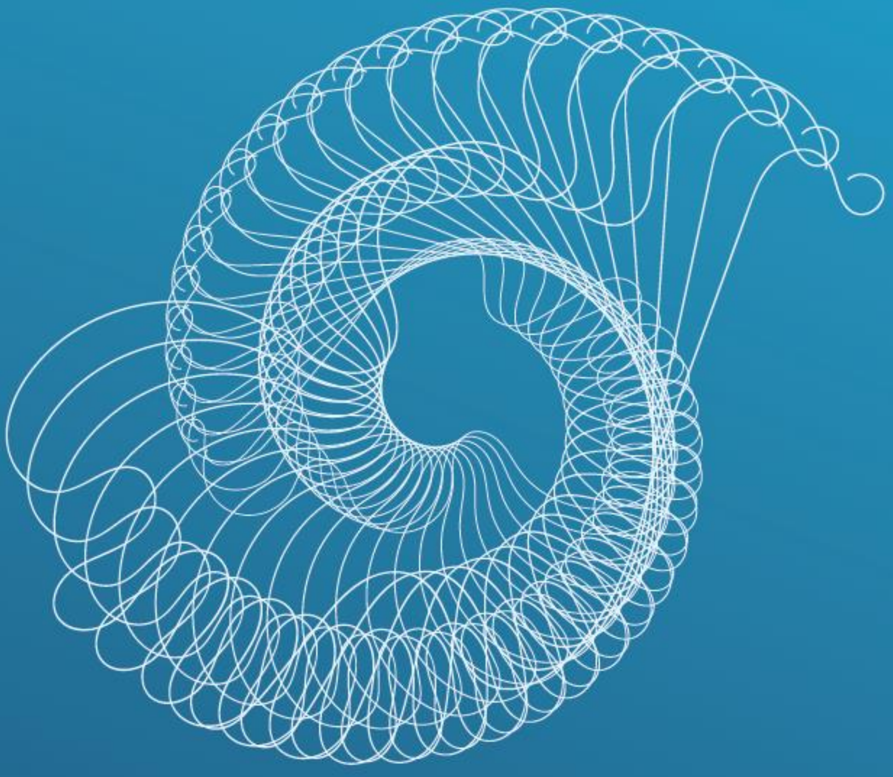
JOSCEX

Journal of Soft Computing Exploration



JOSCEX Volume 1 Number 1 Pages 1-100 Semarang September 2020 e-ISSN xxxx - xxxx p-ISSN xxxx - xxxx

Journal of Soft Computing Exploration

Publisher:
SHM Publihser, Karanglo Street 64, Pedurungan, Semarang City, Central Java, Indonesia

Journal of Soft Computing Exploration

Vol. 1, No. 1, September 2020

Journal of Soft Computing Exploration is a journal that publishes scientific research papers related to soft computing and disseminates it widely to the public, especially soft computing enthusiasts. We hereby invite friends to post articles and citation articels in our journals. We appreciate if you would like to subbmit your paper for publication in JOSCEX.

The process of publishing articles in the Journal of Soft Computing Exploration will be carried out immediately after going through several stages, namely being examined by editors, reviewers, and final checking. Researchers who submit to JOSCEX are expected to be responsive to make revisions when contacted by JOSCEX so that they can produce qualified work, add knowledge, especially in the field of soft computing, and can be published on JOSCEX.

Publisher:
SHM Publihser, Karanglo Street 64, Pedurungan, Semarang City, Central Java, Indonesia

Indexing:
Journal of Soft Computing Exploration indexed is in process

EDITORIAL TEAM

Journal of Soft Computing Exploration

Vol. 1, No. 1, September 2020

Contact:
Secretariat of the Journal of Soft Computing Exploration
SHM Publihser, Karanglo Street 64, Pedurungan, Semarang City, Central Java, Indonesia
Telp. (024) 6735842, email: shmpublisher@gmail.com
Website: http://https://shmpublisher.com/index.php/joscex

Journal of Soft Computing Exploration

Vol. 1, No. 1, September 2020

Table of Contents

JOSCEX
Journal of Soft Computing Exploration

Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review

Ilham Esa Tiffani
Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 20, 2020
Revised Sept 6, 2020
Accepted Sept 19, 2020

Keywords:

Text Mining;
Sentiment Analysis;
Naïve Bayes Classifier;
N-Gram;
Hotel Review

ABSTRACT

The information needed in its development requires that proper analysis can provide support in making decisions. Sentiment analysis is a data processing technique that can be completed properly. To make it easy to classify hotels based on sentiment analysis using the Naïve Bayes Classifier algorithm. As a classification tool, Naïve Bayes Classifier is considered efficient and simple. In this study consists of 3 stages of sentiment analysis process. The first stage is text pre-processing which consists of transform case, stopword removal, and stemming. The second stage is the implementation of N-Gram features, namely Unigram, Bigram, Trigram. The N-Gram feature is a feature that contains a collection of words that will be referred to in the next process. Next, the last click is the hotel review classification process using Na menggunakanve Bayes Classifier. OpinRank Hotels Review dataset on Naïve Bayes Classifier using N-Gram namely Unigram, Bigram, Trigram with research results that show Unigram can provide better test results than Bigram and Trigram with an average accuracy of 81.30%.

Corresponding Author:

Ilham Esa Tiffani
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: tiffaniilham@gmail.com

1. INTRODUCTION

The development of information technology and websites, it allows the managers of the world of tourism to provide more detailed information about the tourism products offered. Currently there are many travel websites that provide facilities for internet users to write their opinions and personal experiences online [1].

A text can consist of only one word or sentence structure [2]. Information in the form of text is important information and is widely obtained from various sources such as books, newspapers, websites, or e-mail messages. Retrieval of information from text (text mining), among others, can include text or document categorization, sentiment analysis, search for more specific topics (search engines), and spam filtering [3]. Text mining is one of the techniques that can be used to do classification where, text mining is a variation of data mining that tries to find interesting patterns from a large collection of textual data [4].

The classification method itself many researchers use the Naïve Bayes Classifier where a text will be classified in machine learning based on probability [5]. Naïve Bayes Classifier is a pre-processing technology in the classification of features, which adds scalability, accuracy and efficiency which is certainly very much in the process of classifying a text. As a classification tool, Naïve Bayes Classifier is considered efficient and simple, and sensitive to feature selection [6].

The data used in this study contains hotel reviews in English so that it can be seen that the grammar used by a person is very diverse in writing the review, diversity makes the features generated through N-Gram will be very much. Therefore, here we will use N-Gram word characters with N = 1, 2, 3 to retrieve features in a review which will then be classified with the Naïve Bayes Classifier Algorithm.

It is expected that the N-Gram Naïve Bayes Classifier Algorithm in this study can be classified correctly and appropriately. So that the main purpose of this study can be fulfilled which is to know the effect of N-Gram features on Naïve Bayes Classifier for sentiment analysis of hotel reviews.

2. METHOD

The step of research include text pre-processing, the application of N-Gram features, and the Naïve Bayes Classifier classification. The research starts by inputting OpinRank Hotels Review datasets. Next, the data will be processed in the text pre-processing stage, namely with a transform case, stopword removal, and stemming. Then the N-Gram feature selection will be carried out, namely unigram, bigram, trigram. Based on the selected features, the classification process will be carried out using the Naïve Bayes Classifier algorithm. Then the classification model is tested using test data and evaluated using a confusion matrix to produce accuracy values. Flowchart of the research method can be seen in Figure 1.

2.1 Dataset

The data used in this study is OpinRank Hotels Review Dataset (in English) obtained from the UCI Machine Learning Repository. The data contains 1000 documents consisting of 500 documents labeled positive and 500 labeled negative. The dataset is obtained in .txt format and then the file is converted into a table with two columns: the first column contains text and the second column contains labels defined by "0" means negative and "1" means positive as shown in Table 1.

Table 1. Data samples in CSV format

Text	Label
Poor location.. This hotel is located in a run down part of the city. The hotel room smelt of ammonia, the toilet would not flush and we could not sleep due to the traffic/street noise. The breakfast was poor and over priced at $12.50. We would not stay there again.	0
I had two nights stay at this hotel, very nice sleep, the bed was fantastic. Staffs' service was good and helpful.	1

2.2 Text Pre-processing

The text pre-processing phase performed in this study is Transform Case, Stopword Removal, Stemming. Text pre-processing is the stage of the initial process of the text to prepare the text into data that will be further processed.

2.2.1 Transform case

The process to change the form of words, in this process the characters are made into lowercase or lower case all. The steps of the transform case process are as follows:

Step 1 : Data input used in the form of hotel reviews.
Step 2 : Hotel review data if there are characters that use capital letters (uppercase), then these characters will be changed to lowercase (lowercase).
Step 3 : Hotel review data becomes lowercase which is then used in the stopword removal process.
Step 4 : The process is complete. The results of the process of the transform case stage can be seen in Table 2

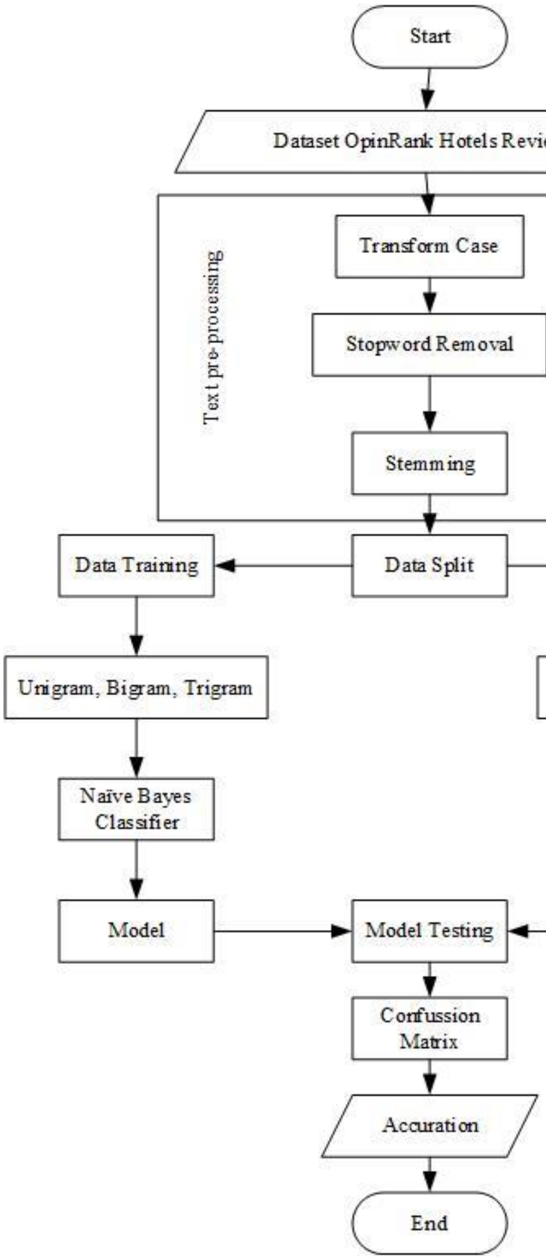


Figure 1. Naive Bayes Classifier algorithm with N-Gram flowchart

Table 2. Results of the Transform Case Process

Review of Data	Transform Case Results
My husband and I stayed at the Chamberlain Hotel for three nights	my husband and i stayed at the chamberlain hotel for three nights

2.2.2 Stopword removal

Stopword Removal is the process of removing words that often appear but do not have any effect in the extraction of text classifications. The steps for the stopword removal process are as follows:

Step 1 : The word from the transform case result will be compared with the word in the stopword list.
Step 2 : Check whether the word is the same as the stopword list or not.
Step 3 : If the word is the same as the stopword list, then the word will be deleted.
Step 4 : The process is complete. The results of the process of the stopword removal stage can be seen in Table 3.

Tabel 3. Results of the Stopword Removal Process

Transform Case Results	Stopword Removal Results
my husband and i stayed at the chamberlain hotel for three nights	husband stayed chamberlain hotel nights

2.2.3 Stemming

The process of mapping and decomposing the shape of a word into basic word forms. The purpose of the stemming process is to eliminate the affixes that exist in each word. The words in the stopword list are pronouns, connectors and pointers. The steps in the stemming process are as follows:

Step 1 : The word from the stopword removal result is checked, whether the word from the stopword removal result is a basic word or not
Step 2 : If the root word then the process has stopped or finished but if it is not a root word then delete the suffix (the affix which is located at the end of the word)
Step 3 : Word resulting from suffix deletion if it is a base word, the process is complete, but if it is not
Step 4 : The process is complete. The results of the process of the stemming stage can be seen in Table 4.

Tabel 4. Results of the Stemming Process

Stopword Removal Results	Stemming Results
husband stayed chamberlain hotel nights	husband stay chamberlain hotel night

2.3 N-Gram

N-Gram is a n-character chunk taken from a string [7]. N-Gram is used in the process of making a model by dividing a sentence into parts of words. In N-Gram, 'N' shows the number of words that will be grouped into one section. On research [8] divide the N-Gram into three types, namely:

a. Unigram: token consisting of only one word.
b. Bigram: a token consisting of two words.
c. Trigram: a token consisting of three words.

The rules used to form the three types of tokens are overlapping tokens. Examples of the process of N-Gram characters generated from the comments results of the Stemming stage can be seen in Table 5.

Table 5. N-Gram Process

N-Gram Results	
Unigram	husband, stay, chamberlain, hotel, night
Bigram	husband stay, stay chamberlain, chamberlain hotel, hotel night
Trigram	husband stay chamberlain, stay chamberlain hotel, chamberlain hotel night

In this study took up to 3 words because in the structure of English phrases with a single meaning have a maximum of 3 words. Phrases are added to a sentence to make the sentence more complex. The advantage of N-Gram is based on the characteristics of N-Gram as part of a string, so errors in some strings will only result in differences in some N-Gram.

2.4 Naïve Bayes Classifier

Naïve Bayes Classifier is a statistical classification that can be used to predict the probability of membership of a class. Naïve Bayes Classifier is based on the Bayes theorem which has the same classification capabilities as the Decision Tree and Neural Network. Naïve Bayes Classifier is proven to have high accuracy and speed when applied to databases with large data [9]. Naïve Bayes Classifier is a popular and good algorithm for high-dimensional data such as text [10].

The flow of the Naïve Bayes Classifier can be seen in Figure 2 as follows:

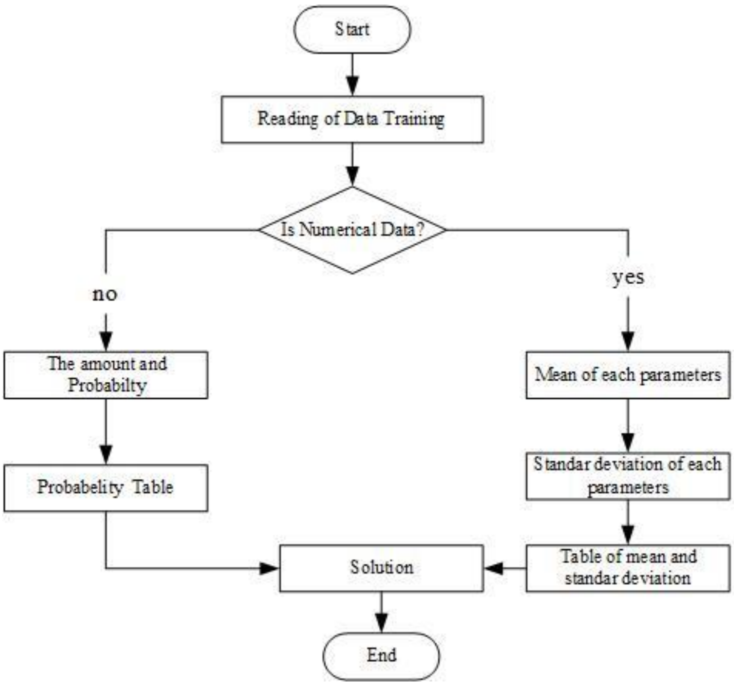


Figure 2. Naïve Bayes Classifier flowchart

Classification is the process of classifying a collection of objects, data or ideas into groups, where each member has one or more of the same characteristics. The classification stage using the Naïve Bayes Classifier is divided into 2 processes, namely training and testing. The training process is carried out to produce a probabilistic model of features that will later be used as a reference calculation for classifying testing data. The stages of sentiment classification use the Naïve Bayes Classifier as follows:

a. Training Process
   1. Count $P(c_i)$
   2. Count $P(w_k \mid c_i)$ for each w_k on the model
b. Testing Process
   1. Count $P(c_i)$ Πk $P(w_k \mid c_i)$ for each category
   2. Decide $c*$, i.e. categories with values $P(c_i)$ Πk $P(w_k \mid c_i)$ the highest

3. RESULT AND DISCUSSION

3.1 Result

In research, the proposed algorithm is tested using the python programming language. The classification process in the dataset uses the Naïve Bayes Classifier algorithm, the classification in the dataset with the Naïve Bayes Classifier algorithm applied to Unigram produces 81.30% accuracy, the dataset classification with the Naïve Bayes Classifier algorithm applied to Bigram produces an accuracy of 71.60%, and the classification of the dataset with the Naïve Bayes Classifier algorithm is applied Trigram produces 71.90% accuracy.

3.2 Discussion

3.2.1 Naïve Bayes Classifier algorithm + *Unigram*

This classification stage applies the Naïve Bayes Classifier algorithm with Unigram on the OpinRank hotels review dataset. The training data will be divided into 10 subset data to conduct the learning process. This process takes 10 iterations to then get the classification model. Then the algorithm will be tested with a confusion matrix. From the classification of the Naïve Bayes Classifier algorithm by applying Unigram produces accuracy as shown in Table 6.

Table 6. Accuracy results of Naïve Bayes Classifier + Unigram algorithm

***k* to**	**Accuracy**
1	68,00%
2	75,00%
3	79,00%
4	85,00%
5	83,00%
6	84,00%
7	86,00%
8	85,00%
9	84,00%
10	84,00%
Average	81,30%

3.2.2 Algoritma *Naïve Bayes Classifier* + *Trigram*

This classification phase applies the Naïve Bayes Classifier algorithm with Trigram on the OpinRank hotels review dataset. The training data will be divided into 10 subset data to conduct the learning process. This process takes 10 iterations to then get the classification model. Then the algorithm will be tested with a confusion matrix. From the classification of the Naïve Bayes Classifier algorithm by applying Trigram produces accuracy as shown in Table 7.

Table 4.8. Accuracy results of Naïve Bayes Classifier + Trigram algorithm

***k* to**	**Accuracy**
1	63,00%
2	74,00%
3	75,00%
4	75,00%
5	73,00%
6	72,00%
7	70,00%
8	71,00%
9	74,00%
10	72,00%
Average	**71,90%**

The accuracy of the Naïve Bayes Classifier algorithm using Unigram, Bigram, Trigram compared to related research results in better accuracy. OpinRank hotels review dataset using Naïve Bayes Classifier in related research has an accuracy of 55.00%, the classification of the Naïve Bayes Classifier algorithm with Unigram is able to produce an average accuracy of 81.30%, the classification of the Naïve Bayes Classifier algorithm with Bigram is able to produce an average accuracy of 71.60%, and the classification of the Naïve Bayes Classifier algorithm with the Trigram is capable of producing an average accuracy of 71.90. Comparison of accuracy results can be seen in Figure 3.

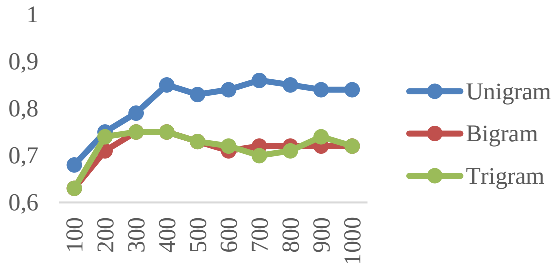


Figure 3. Graph of accuracy results of Naïve Bayes Classifier algorithm with N-Gram

Based on the results of the implementation of Unigram, Bigram, Trigram on the Naïve Bayes Classifier algorithm that has been done, it can be seen that the accuracy for sentiment analysis of hotel reviews using

OpinRank Hotels Review datasets is taken from the UCI Machine Learning Repository after going through text pre-processing and then applying the N-Gram feature and classification using the Naïve Bayes Classifier can improve accuracy so that it can be used by subsequent researchers as a reference in conducting hotel review sentiment analysis research.

4. CONCLUSION

The application of Unigram, Bigram, Trigram on the Naïve Bayes Classifier algorithm for the analysis of hotel review sentiments in this study is the OpinRank hotel reviews dataset that has been done in the text pre-processing stage will be divided into training data and testing data. The training data will be given the features of Unigram, Bigram, Trigram by breaking a sentence into words whose results will be classified with the Naïve Bayes Classifier algorithm so as to produce a more optimal Naïve Bayes Classifier model. The final result of the classification is testing the model of data testing. Based on these tests, an accuracy of the Naïve Bayes Classifier algorithm can be seen using a confusion matrix. The average accuracy of 10 subsets of data obtained using Unigram in the Naïve Bayes Classifier algorithm is 81.30%, the average accuracy of 10 subset data obtained using Bigram in the Naïve Bayes Classifier algorithm is 71.60% and the results of the average accuracy 10 subsets of data obtained using Trigram in the Naïve Bayes Classifier algorithm are 71.90%.

REFERENCES

[1] C. Wang, Y. Zhang, J. Song, Q. Liu and H. Dong, "A novel optimized SVM algorithm based on PSO with saturation and mixed time-delays for classification of oil pipeline leak detection, " *Sys. Sci. & Con. Eng.*, vol. 7, no. 1, pp. 75-88, 2019.

[2] R. Carter and M. McCarthy, Cambridge Grammar of English Paperback with CD ROM: A Comprehensive Guide. Cambridge, UK: Cambridge University Press, 2006, pp.179-185.

[3] N. Buslim, A. E. Putra, and L. K. Wardhani, "Chi-square feature selection effect on naive bayes classifier algorithm performance for sentiment analysis document, " presented at the 7th International Conference on Cyber and IT Service Management, Jakarta, Indonesia, Nov. 6-8. 2019.

[4] R. Feldman and J. Sanger, The Text Mining Handbook . Cambridge, UK: Cambridge University Pres. 2009.

[5] W. Zhang and F. Gao, "An Improvement to Naive Bayes for Text Classification, " *Procedia Engineering*, vol. 15, no. 2, pp. 2160–2164, 2011.

[6] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature Selection for Text Classification with Naïve Bayes, " *Expert System Application*, vol, 36, no. 3, pp. 5432–5435, 2009.

[7] J. Violos, K. Tserpes, I. Varlamis, and T. Varvarigou, "Text classification using the n-gram graph representation model over high frequency data streams, " *Front. Appl. Math. Stat*. vol. 4, no. 41, pp. 1-19, 2018.

[8] Z. Drus and H. Khalid, " Sentiment Analysis in Social Media and Its Application: Systematic Literature Review, " presented at the fifth Information Systems International Conference, Surabaya, Indonesia, July 23-24, 2019.

[9] R. E. Banchs. Text Mining with MATLAB®. New Delhi, India: Springer-Verlag New York, 2013, pp. 49-75.

[10] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of Hoax News Detection using Naïve Bayes Classifier in Indonesian Language," presented at the 11th International Conference Informatics Communication Technology System ICTS, Surabaya, Indonesia, Oct. 30-31, 2017.

Journal of Soft Computing Exploration

Support Vector Machine (SVM) Optimization Using Grid Search and Unigram to Improve E-Commerce Review Accuracy

Sulistiana[1], Much Aziz Muslim[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 18, 2020
Revised Sept 1, 2020
Accepted Sept 21, 2020

Keywords:

Text Mining
Support Vector Machine
Unigram,
Grid Search

ABSTRACT

Electronic Commerce (E-Commerce) is distributing, buying, selling, and marketing goods and services over electronic systems such as the Internet, television, websites, and other computer networks. E-commerce platforms such as amazon.com and Lazada.co.id offer products with various price and quality. Sentiment analysis used to understand the product's popularity based on customers' reviews. There are some approaches in sentiment analysis including machine learning. The part of machine learning that focuses on text processing called text mining. One of the techniques in text mining is classification and Support Vector Machine (SVM) is one of the frequently used algorithms to perform classification. Feature and parameter selection in SVM significantly affecting the classification accuracy. In this study, we chose unigram as the feature extraction and grid search as parameter optimization to improve SVM classification accuracy. Two customer review datasets with different language are used which is Amazon reviews that written in English and Lazada reviews in the Indonesian language. 10-folds cross validation and confusion matrix are used to evaluating the experiment results. The experiment results show that applying unigram and grid search on SVM algorithm can improve Amazon review accuracy by 26,4% and Lazada reviews by 4,26%.

Corresponding Author:

Sulistiana
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: sulistiana@students.unnes.ac.id

1. INTRODUCTION

Commercial sites around the world mostly run on the online platform. People trade their products on different e-commerce sites [1]. Products can be bought from any sites with various prices. Customers usually want a product with the best quality and low price but cannot check it directly. Reviews from other customers can be so helpful to decide whether she will buy the product or not [1]. A review has important information about customers' problems and their experience that can be helpful for creating a conceptual design, personalization, product recommendation, better customer understanding, and customer acquisition [2]. Sentiment analysis used to understand product popularity from customers' reviews worldwide.

Sentiment analysis is proceed using text mining method [3]. Text mining is a part of data mining that used to finding patterns from natural language texts [4]. Classification is one of data mining techniques for predicting a decision.

Classification is a process for grouping some data into classes based on characters and patterns similarity [5]. Classification in machine learning needs to identify data features to determine the class category. This features identification called feature extraction. Based on Laoh et al. in [6], applying n-gram as a feature extraction method can improve accuracy for sentiment analysis tasks [6].

Support Vector Machine (SVM) is one of commonly used algorithms for classifying data [7]. Ravi and Khettry in [8] used Naive Bayes, SVM, Random Forest Classifier, and K-Nearest Neighbor to classify Amazon review dataset. Bigram is used as the feature extraction with tokenization, punctuation removal, stopword removal, and stemming as the preprocessing steps. The results obtained from the experiment is 62,5%, 70%, 77,65%, and 65% for SVM, Naive Bayes, Random Forest Classifier, and K-Nearest Neighbor respectively.

SVM performance depends on the kernel [9]. Linear kernel is the simplest kernel and has only one parameter C [10]. Parameter C has a big impact on SVM classification performance because it determines the trade-off between minimizing errors and maximizing classification margin [11]. Practically, changing C value can control training errors, testing errors, number of support vectors, and SVM margin [9].

Parameter configuration has a significant impact on improving accuracy [12]. Hence, optimal values for the learning parameters are needed to build an accurate model. Grid search is a parameter optimization method. The advantage of using grid search is higher learning accuracy and its capability to parallelize since each process is independent of one another [13].

This study aims to improve SVM accuracy for classifying e-commerce customers review dataset using grid search combined with unigram as the feature extraction then compare it to the previous work.

2. METHOD

In this work, the proposed algorithm was implemented using Python 3.8 with libraries Django 3.0.3, NLTK 3.4.5, Numpy 1.17.4, Openpyxl 3.0.3, Pandas 0.25.3, Sastrawi 1.0.1, Scikit-Learn 0.22, and Xlrd 1.2.0. The first step to conduct this research is collecting datasets. Then, we normalize data by applying some text preprocessing steps: transform cases, punctuation removal, tokenize, stopword removal, and stemming.

After the preprocessing step, we extracted features from data using the word unigram method. This process will chunk text data into one-word-length strings. Next, we split the dataset into training dan testing data with a ratio of 75:25.

We used SVM as the classifier with linear kernel. Training data is used in the learning process to build a classification model. 10-folds cross validation is applied in the training process as well as grid search for finding the optimal parameter C. Next, we tested the model using testing data. The result obtained from the experiment was mapped into a confusion matrix to calculate the accuracy. Figure 1 shows the flowchart of the research methodology used in this work.

2.1 Dataset

In this research, we used two datasets with different language. The first one is Amazon customers review dataset that written in English and the second one is Lazada customer review in the Indonesian language. These language differences aim to prove accuracy improvement. Both are public datasets from Kaggle.com that consist of positive and negative comments. Each dataset contains 1500 reviews. On Amazon dataset, positive comments labeled as 1 and negative comments as 0. For Lazada dataset, we used rating-based labeling to determine the label for each data. Comment with rating 3 to 5 considered as a positive comment, whereas comments with rating 1 and 2 will be labeled as negative ones.

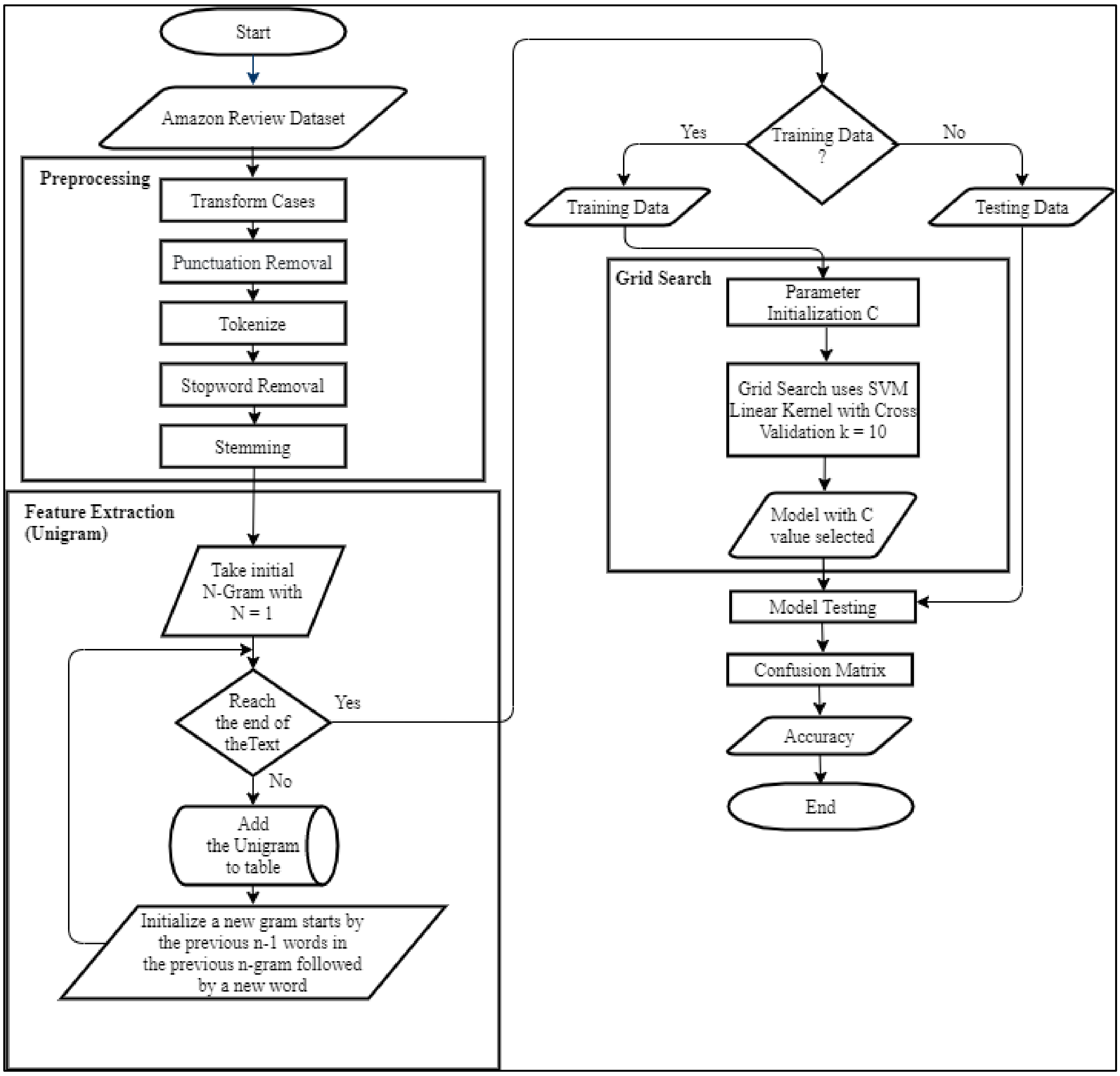

Start
Amazon Review Dataset
Preprocessing
Transform Cases
Punctuation Removal
Tokenize
Stopword Removal
Stemming
Feature Extraction (Unigram)
Take initial N-Gram with N = 1
Reach the end of theText
Yes
No
Add the Unigram to table
Initialize a new gram starts by the previous n-1 words in the previous n-gram followed by a new word
Training Data ?
Yes
No
Training Data
Testing Data
Grid Search
Parameter Initialization C
Grid Search uses SVM Linear Kernel with Cross Validation k = 10
Model with C value selected
Model Testing
Confusion Matrix
Accuracy
End

Figure 1. Flowchart SVM algorithm with Grid Search and Unigram

2.2 Text Preprocessing

Text preprocessing is the first stage to prepare the data so that they can proceed in the next steps. Text preprocessing in this study consists of 5 steps: (1) *transform cases*; (2) *punctuation removal*; (3) *tokenize*; (4) *stopword removal*, and (5) *stemming*. The result of the text preprocessing stage can be seen in Table 1.

2.3 N-Gram

N-gram is defined as a contiguous sequence of *n* items from a given text [15]. The overlapping token method used to divide n-sized token. In this research, we used n = 1 or unigram for feature extraction. Table 2 shows an example of word unigram extraction.

Table 1. Text Preprocessing Result

Text	Token
Stuning even for the non-gamer: This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^_^	"stune"; "even"; "nongam"; "sound"; "track"; "beauti"; "paint"; "seneri"; "mind"; "well"; "would"; "recomend"; "even"; "peopl"; "hate"; "vid"; "game"; "music"; "play"; "game"; "chrono"; "cross"; "game"; "ever"; "play"; "best"; "music"; "back"; "away"; "crude"; "keyboard"; "take"; "fresher"; "step"; "grate"; "guitar"; "soul"; "orchestra"; "would"; "impress"; "anyon"; "care"; "listen"

Table 2. Unigram Result

Stune, even, nongame, sound, track, beauty, paint, seneri, mind, well, would, recommend, even, people, hate, vid, game, music, play, game, chrono, cross, game, ever, play, best, music, back, away, crude, keyboard, take, fresher, step, grate, guitar, soul, orchestra, would, impress, anyon, care, listen

2.4 Support Vector Machine

SVM is a machine learning method that works based on the Structural Risk Minimization principle to find the best hyperplane for separating two classes in input space [14]. Figure 2 illustrates how SVM works by finding hyperplane with the maximum margin that separates two classes.

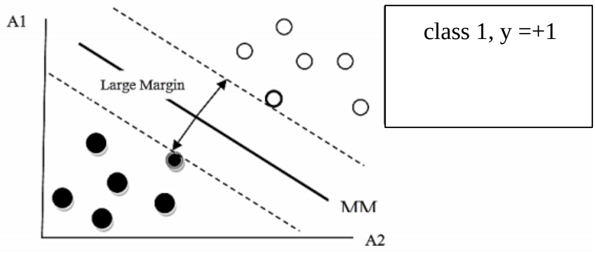


Figure 2. Separating 2 Data Classes with Maximum Margin

Our experiment consists of two phases: training and testing phase. The result from the training phase is a probabilistic model that will be used to classify data in the testing phase.

2.5 Grid Search

Grid search is an exhaustive search based on a defined subset of the hyper-parameter space. The hyper-parameters are specified using minimal value (lower bound), maximal value (upper bound), and a number of steps [16]. Grid Search divides the range of parameters to be optimized into a grid and crosses all points to get the optimal parameters. Grid Search optimizes SVM parameter using cross validation technique as a performance metric. According to Lin et al., applying cross validation technique can prevent overfitting problem [17]. Grid search aims to identify the best hyperparameter combination so that the classifier can predict the unknown data correctly. The pseudocode for grid search algorithm can be seen in Figure 3.

Figure 3 explains grid search pseudocode that start with initializing candidates of parameter C. In this study, we used 11 candidates, 0.10, 0.18, 0.26, 0.34, 0.42, 0.50, 0.58, 0.66, 0.74, 0.82, and 0.90. We also applied 10-folds cross validation during the training phase to find the optimal value for parameter C.

```
ALGORITHM: Grid Search for parameter C on SVM
Initialize list of C candidates
FOR every c in list of C candidates
    Train SVM with c on TrainingSet
    Evaluate SVM classification on ValidationSet
    IF accuracy > MaxAccuracy
        THEN save MaxC = c
    ENDIF
ENDFOR
RETURN MaxC
```

Figure 3. Grid search pseudocode

2.6 Validation and Evaluation

Validation and evaluation are done to measure the performance of our proposed method. In this step, we split the dataset into training dan testing data. Training data used to train the model. Then, this model is tested using testing data. The results obtained from this process used to measure model performance. We use confusion matrix for evaluation. Table 3 shows the confusion matrix for binary classification.

Table 3. Confusion Matrix for Binary Classification

f_{ij}		Predicted Class (j)	
		Class =1	Class =0
Actual Class (i)	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

Table 3. illustrates confusion matrix for binary classification with classes 0 and 1. *fij* denotes the number of data from class i that predicted as class j by the classifier. For instance, cell *f11* is the number of data from class 1 that correctly predicted as class 1, while cell f10 shows the number of data from class 1 that predicted as class 0.

Based on this confusion matrix, we obtain numbers of correctly predicted data (*f*11+*f*00) and numbers of wrongly predicted data (*f*10 + *f*01). Then, we can calculate the error rate and accuracy. The error rate is defined as the ratio between numbers of wrongly predicted data and the total number of data, while accuracy is defined as the ratio between numbers of correctly predicted data and the total number of data. Equation 1 is a formula to calculate data accuracy.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ total\ data} = \frac{f11+f00}{f11+f10+f01+f00} \tag{1}$$

3. RESULT AND DISCUSSION

3.1 Result

This section will discuss the results from our experiments.

3.1.1 SVM classification Results

Our first scenario is to apply SVM only as a classifier. As mentioned before, we split the dataset into training and testing data with a ratio of 75:25 and using confusion matrix to evaluate our model performance. The experiment results show that accuracy for Amazon dataset is 54,40% whereas for Lazada dataset is 85,87%.

3.1.2 SVM + Unigram classification Results

In our second scenario, we modified the configuration of the first scenario by adding unigram as the feature extraction for both datasets. Unigram will extract features from the dataset by dividing text data into one-word-length strings. From the experiment, both Amazon and Lazada dataset shows an accuracy improvement. The accuracy of the Amazon dataset in this scenario is 80,80% which is improved by 26,40%. Meanwhile, for the Lazada dataset, the accuracy is 88,00%, which improved 4,26%.

3.1.3 SVM + Unigram + Grid Search Classification Results

The third scenario is combining SVM with unigram and grid search. This configuration applied for both Amazon and Lazada datasets. Grid search performed in the training phase to find the optimal parameter C from the 11 candidates that already mentioned in section 2.5. We also applied 10-folds cross validation in this phase. Table 4 shows the results obtained from the grid search process for Amazon dataset.

Table 4. The grid search results for Amazon dataset

Experiment	Parameter C	Accuracy
1	0,10	76,81
2	0,18	75,11
3	0,26	75,02
4	0,34	74,76
5	0,42	74,85
6	0,50	74,76
7	0,58	74,49
8	0,66	74,49
9	0,74	74,58
10	0,82	74,50
11	0,90	74,49
Optimal Parameter	**0,1**	**76,81**

From Table 4 we can see that the highest accuracy obtained when the C value is 0,1. This means that the optimal parameter C for Amazon dataset is 0,1. For Lazada dataset, the grid search results described in Table 5. As we can see from Table 5, the highest accuracy obtained when the C value is 0,58. This means that the optimal parameter C for Lazada dataset is 0,58.

The optimal values will be applied to build a model that will be used in the testing phase. The testing results show that the accuracy for Amazon dataset is 80,80% and for Lazada dataset is 90,13%.

Table 5. The grid search results for Lazada dataset

Experiment	Parameter C	Accuracy
1	0,10	84,71
2	0,18	85,15
3	0,26	84,71
4	0,34	84,80
5	0,42	85,15
6	0,50	85,24
7	0,58	85,24
8	0,66	84,80
9	0,74	84,80
10	0,82	84,62
11	0,90	84,36
Optimal Parameter	**0,58**	**85,24**

3.2. Discussion

Our experiment results show that combining SVM with unigram and grid search can improve the accuracy of Amazon and Lazada datasets as can be seen in Table 6 and Table 7 respectively.

Table 6. The accuracy improvement of Amazon Dataset

	Train	***Test***
SVM	52,35%	54,40%
SVM+*Unigram*	76,81%	80,80%
SVM+*Unigram*+*Grid Search*	76,81%	80,80%

Based on Table 6, we can see that there is no improvement between the results from scenario 2 (SVM + Unigram) and scenario 3 (SVM+Unigram+Grid Search) for Amazon dataset. This happens since the optimal parameter C for Amazon dataset is the default value.

Table 7. The accuracy improvement of Lazada Dataset

	Train	***Test***
SVM	84,45%	85,87%
SVM+*Unigram*	84,71%	88,00%
SVM+*Unigram*+*Grid Search*	85,24%	90,13%

Since the optimal C value is 0,58, which is different from the default value, we can see the improvement of accuracy for Lazada dataset in each scenario as depicted in Table 7.

3.2.1 Comparision with Previous Study

Next, we try to compare the results from our proposed method with previous work. In the previous study [8], Ravi dan Khettry also conducted an experiment to calculate the accuracy of e-commerce review dataset. They used Amazon customer reviews as the dataset. Bigram is applied as the feature extraction. To normalize the data, some preprocessing steps were done that consist of tokenization, punctuation removal, stopword removal, and stemming. We build a model using this configuration and tested it with our datasets. We obtained the accuracy for Amazon dataset is 73,60% and for Lazada dataset is 86,93%. The accuracy comparison between previous work and our proposed method described in Table 8 and Figure 4.

Table 8. Accuracy comparison with previous work

Dataset	**SVM accuracy results before applying the method**	**Previous Work**	**Proposed Method (Grid Search + Unigram)**
Amazon	54,40%	73,60%	80,80%
Lazada	85,87%	86,93%	90,13%

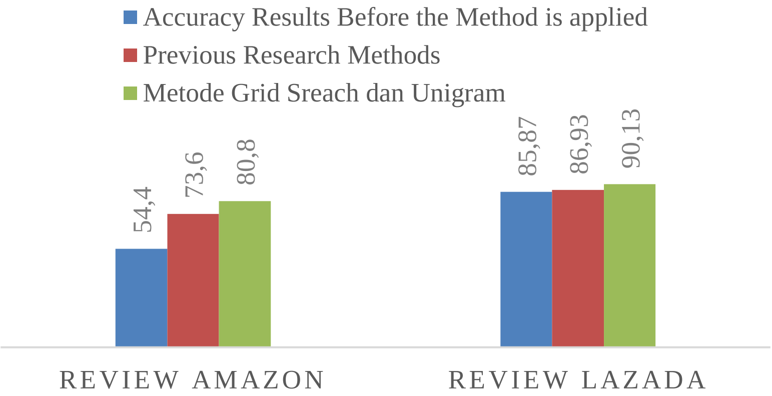


Figure 4. Comparison with previous work

4. CONCLUSION

The experiment results in this study show that combining SVM with grid search and unigram can improve the classification accuracy for both datasets. Amazon review dataset accuracy increased by 26,4%, from 54,40% to 80,80% with optimal parameter C 0,1. While for Lazada review dataset, the accuracy increased by 4,26% from 85,87% to 90,13% with optimal parameter C 0,58. Based on the results, we conclude that applying grid search and unigram can help to find the optimal parameter and improve SVM accuracy. These results are better than the previous study that can only improve accuracy by 19,2% on Amazon dataset and 1,06% on Lazada dataset.

REFERENCES

[1] T.U Haque, N. N, Saber, and F. M. Shah, “Sentiment analysis on large scale Amazon product reviews. ” presented at Int. Conf. on Inno. Res. & Dev., Bangkok, Thailand., May 11-12, 2018.

[2] J. Zhan, H. Tong, and Y. Liu, Y. “Gather customer concerns from online product reviews – A text summarization approach., ” *Expert Systems With Applications*, vol. 36, no. 2, pp. 2107–2115. 2009

[3] U. L. Larasati, M. A. Muslim, R. Arifudin, and Alamsyah, “Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis ” *Journal of Soft Computing Exploration*, vol. 6, no. 1, pp. 138-149. 2019.

[4] H. C. Yang, and C. H. Lee, “A text mining approach for automatic construction of hypertexts, ” *Expert Systems with Applications*, vol. 29, no. 4, pp. 723–734. 2005.

[5] A. F. Indriani, and Muslim, M. A, “SVM Optimization Based on PSO and AdaBoost to Increasing Accuracy of CKD Diagnosis, ” *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, DOI: 10.24843/LKJITI.2019.v10.i02.p06. 2019

[6] E. Laoh, I. Surjandari, and N. I. Prabaningtyas, “Enhancing hospitality sentiment *review*s analysis performance using SVM N-grams method, ” presented at 16th Int. Conf. on Service Systems and Service Management, Depok, Indonesia, July 15-18, 2019.

[7] M. A. Muslim, B. Prasetiyo, B., E. Listiana, E. L. H. Mawarni, A. Juli, Mirqotussa’adah., S. H. Rukmana, and A. Nurzahputra, Data Mining Algoritma C4.5, Semarang: CV. Pilar Nusantara, 2019.

[8] A. Ravi, A. R. Khettry, and S. Y. Sethumadhavachar, “Amazon Reviews as Corpus for Sentiment Analysis Using Machine Learning” presented at Int. Conf. on Adv. in Comp. & Data Sci, Ghazibad, India, April 12-13, 2019

[9] S. Amari, and S. Wu, “Improving support vector machine classifiers by modifying kernel functions, ” *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[10] K. Srivastava, and L. Bhambhu, “Data classification using support vector machine, ” *J of Theo. & Appl. Inf. Tech.*, vol 1, no. 5, pp. 1–7. 2009.

[11] A. Tharwat, “Parameter investigation of support vector machine classifier with kernel functions, ” *Knowledge and Information Systems*. vol. 6, no. 2, pp. 24-31, 2019.

[12] J. Alex, and B. S. Smola, “A tutorial on support vector regression” *Statistics and Computing*, vol. 14, no. 3, pp. 199–222. 2004.

[13] A. Zakrani, A. Najm, and A. Marzak, “Support Vector Regression Based on Grid-Search Method for Agile Software Effort Prediction, ” *Colloquium in Information Science and Technology*, vol. 8, no. 2, pp. 26-32. 2018.

[14] R. Feldman, and J. Sanger. The Text Mining Handbook. New York: Cambridge University Press. 2006

[15] M. Agyemang, K. Barker, and R.S. Alhajj, “Mining web content outliers using structure oriented weighting techniques and N-grams, ” *Proceedings of the ACM Symposium on Applied Computing*, Santa Fe, New Mexico, USA, March 13-17, 2005.

[16] I. Syarif, I., A. Prugel-Bennett, and G. Wills., “SVM parameter optimization using grid search and genetic algorithm to improve classification performance, ”.*Telkomnika*, vol. 14, no. 4, pp.1502–1509, 2016.

[17] S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, Z. J,” Particle swarm optimization for parameter determination and feature selection of support vector machines, ” *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2009.

Journal of Soft Computing Exploration

Electrical Energy Monitoring System and Automatic Transfer Switch (ATS) Controller with the Internet of Things for Solar Power Plants

Novi Kurniawan
Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 10, 2020
Revised AUg 25, 2020
Accepted Sept 9, 2020

Keywords:

Internet of Things;
Android;
NodeMCU;
Automatic Transfer Switch;
Monitoring;

ABSTRACT

Internet of Things is a technology that connects communication devices with electronic devices that are used everyday using the internet as a medium to communicate between devices and users. The use of IoT technology can be implemented in solar power generation systems. The IoT technology implemented in this study is to monitor and control the use of batteries in solar power plants. Current technology, battery usage can only be monitored closely to get information about battery capacity and battery usage. When the battery is empty or cannot be used to meet electricity needs, it is not equipped with a diversion of existing electricity sources such as PLN electricity. So, we need a renewable technology to get information about batteries and transfer of electricity sources that can be accessed remotely and can be accessed via the internet. The design of this smart monitoring system has stages, namely planning,design , coding , and test . The results of this study are able to see data in the form of battery capacity, electric current and electric power used in Android applications. The data is obtained from sensors that are on smart monitoring connected to the internet network and stored on a database server. Then the data residing on the database server will be retrieved by the application to be displayed to users in the form of graphics and usage lists. Furthermore, the Automatic Transfer Switch system works if the battery capacity sensor has read less than 11.4V then the relay will automatically transfer electricity to PLN. The Android smartphone application is used as a monitoring tool to view data in realtime.

Corresponding Author:

Novi Kurniawan
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: novikurniawan1998@students.unnes.ac.id

1. INTRODUCTION

At this time is an era of technological development that can improve the quality of human life [1]. With internet technologies such as wifi and wireless Internet access, growth continues to facilitate communication [2]. Survey conducted by the Association of Indonesian Internet Network Providers (APJII) in 2018, 171.17 million people in Indonesia have been connected to the internet, an increase of 10.12% from the previous year. 93.9% of smartphone users claim that they access the internet every day [3]. 87.6% of smartphone sales in circulation in 2016 were Android smartphones [4]. The field of research that enables users to manage and optimize electronic and electrical equipment using the internet is often called the Internet of Thing (IoT) [5].

Internet of Things is a technology that connects all devices and the Internet using sensor devices and to identify and manage information [6]. IoT is defined as the interconnection of embedded computing devices (embedded computing) that are uniquely identified in the presence of internet infrastructure [7]. rapid development in the concept of IoT to advance the scheme, namely Internet-of-Everything (IoE) [8]. The concept of IoE is to link all objects with the Internet where all references are for ease of use [9]. IoT can be described as a connecting device commonly used in everyday life such as smartphones, televisions, sensors and actuators to the internet where all devices are connected to each other to enable new types of communication between humans (users) and objects [10]. The progress of IoT which includes anywhere, anytime, and anyone to connect with these objects with the hope that is to expand the network and get ahead of the IoT. The use of IoT technology can be implemented in solar power generation systems.

Electrical energy needs become the main needs in the industrial sectors, housing, education, hospitals, and other places, but sometimes the electricity that is distributed by PLN does not always exist continuously, sometimes there are blackouts and when electricity is needed in important places that should be fulfilled its electricity needs on an ongoing basis would certainly be m ROBLEM great. The control system or control the current began to shift in the automation of control systems or commonly called Autom a tic Transfer Switch (ATS), so that human intervention in controlling very small. When compared with manual workmanship, an equipment system that is controlled by automation will provide benefits in terms of efficiency, safety, and accuracy [11] .

Based on research [12] that applies IoT technology using the ESP8266 Wifi module, so that it can send data to a computer device, but the distance that can be reached is only limited to the ability of the wifi module used (the computer must be connected to the wifi contained in the monitoring tool). Backup system / backup is absolutely necessary in electronic devices that require uninterrupted electrical energy. Reserves are used to replace PLN's main source. Research [13] says that inadequate electricity supply has led to the proliferation of standby generators, especially in developing countries. However, the methods and equipment used to change power supply remain challenging from inefficiency to cost. Most industries still use manual power supply replacement methods, which are covered by various setbacks including: time wasting, heavy operations, vulnerability to fire and high frequency of maintenance. In the study, Aumuzuyi presents a microcontroller-based Automatic Transfer Switch system, which eliminates the challenges of a manual changeover system. Simulation results prove the duration analysis yields very good results.

Based on the description of the problems above, the research focuses on designing an IoT-based monitoring system that can send data to a database server, so that it can be monitored in real time by an Android smartphone anywhere and anytime without being constrained by distance. The Automatic Transfer Switch (ATS) system will also be implemented using the NodeMCU ESP32 Microcontroller which hopes to be more effective because this Microcontroller has been designed as an IoT-based Microcontroller with a Wifi module that is already attached to the NodeMCU. The ATS system will be optimized by applying restrictions on the use of batteries in PLTS at a minimum voltage of 11.4V, so that the tool can make the switch from PLTS to PLN or vice versa.

2. METHOD

This research uses Agile Methods application development method with Extreme Programming (XP) model . According to Pressman, the agile model combines philosophy and steps of development. The philosophy in question covers the demand and development in stages in the scale of its development. The Extreme Programming (XP) approach is a software development model that starts from the planning, design, coding, and test stages [14].

2.1 Planning

In this stage the initial needs of the user are needed or in The XP Process are called user stories. This stage explains the task content, system output requirements and the main features of the application being developed.

2.1.1 Problem identification

This stage identifies problems that may arise during application development. Such as adjusting the voltage level to the condition> 11.4V relay turns on and the voltage condition <11.4V relay does not connect the voltage. As well as features sending data from the tool to the web-server. Smartphones display data flow according to web- server. Problem identification is done so that the application is successfully built properly.

2.1.2 Create user stories

a. User stories on android

Story 1 : users can access the application with a login facility.
Story 2 : users can access the button on the main menu of the application.
Story 3 : users can see the data flow in the real-time menu.
Story 4 : users can zoom in on the graph in realtime data flow.
Story 5 : Users can adjust the data flowing on the graph.
Story 6 : users can see the data flow in decimal form.
Story 7: users can download data from the web-server

b. User stories on the tool

Story 1 : users can see the voltage and current variables according to multitester.
Story 2 : the user can see the relay active at a voltage> 11.4V and the relay turns off at a voltage <11.4V.
Story 3 : Users can see the voltage and current data recorded on the device monitoring to web-server.
Story 4 : users can see the process of recording realtime data.
Story 5 : users can see the recorded data in accordance with the function on monitoring tool.

2.2 Design

The Design Stage describes the system design that will be created according to user requests. Focusing on hardware and software architecture design plans and user interface design.

2.2.1 System design

This section explains the design of the IoT monitoring tool. System design on an IoT - based monitoring tool . There are three sections that include devices Android, web- server as well as a monitoring tool can be seen in Figure 1 .

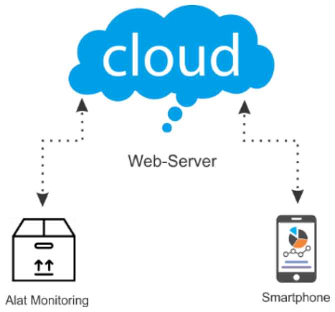


Figure 1. System Design

Android devices function as data visualization into a graph called real- time monitoring. Web-server functions to store and process the data sent. Monitoring tools have a voltage and current sensors that record and send it to the web- server. Recording occurs every 5 seconds and sends it to the web-server continuously.

2.2.2 Design tool

This section explains the design of IoT-based monitoring tools. The monitoring tool uses the NodeMCU ESP32 microcontroller as the main component. NodeMCU ESP32 is connected to voltage, current and relay sensors . The NodeMCU ESP32 wifi module connects the NodeMCU ESP32 with the web-server, when the voltage and current sensors are recording data. Data that is successfully recorded will be sent to the web- server with a span of 12 times in one minute. The web-server converts the data that has been sent into a variable and stores it as a datasheet. The data storage results are visualized on an Android device into a real-time graph. The design of the monitoring tool can be seen in Figure 2.

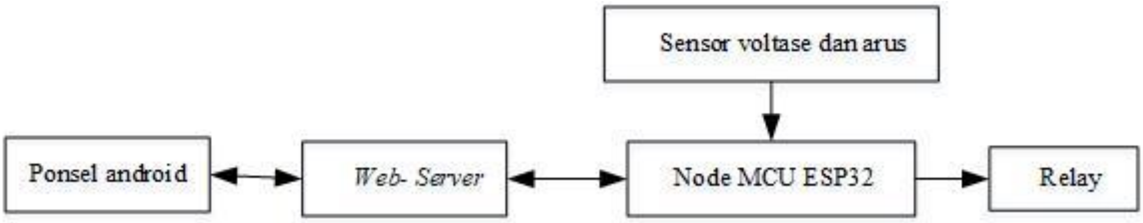


Figure 2. Design of monitoring tools

Circuit design monitoring tools can be seen in Figure 3 .

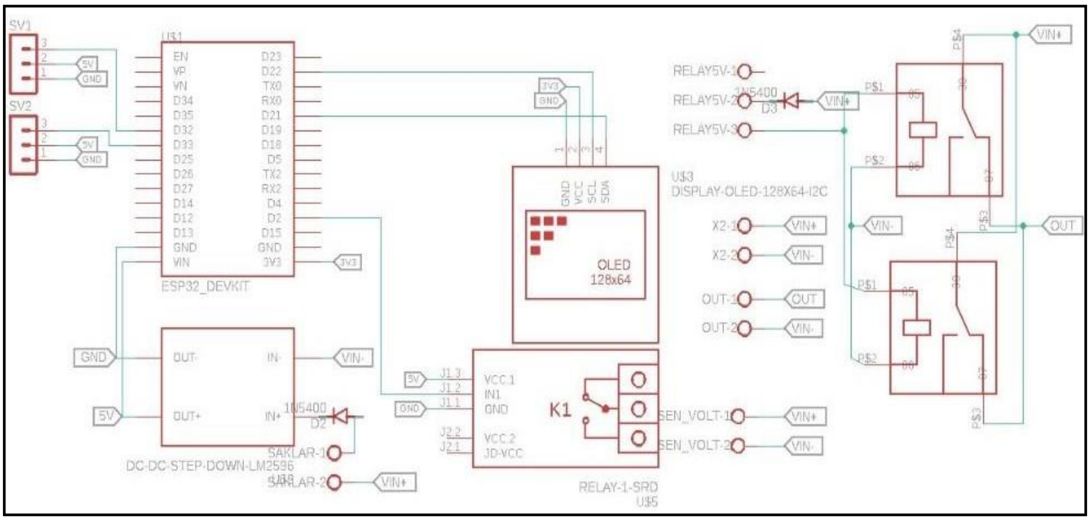


Figure 3. Circuit design

2.2.3 Interface design

The interface design of the IoT monitoring tool consists of 6 pages. The application interface page is as follows:

a. Main Page
b. Menu page
c. Realtime Graph Pages
d. Database page

2.2.4 Database design

IoT-based monitoring tools have a database that is stored on the web-server. The database contains 5 lines which include id, voltage, current, estimated time and date. The id line identifies the data input sequence in the database variable. Voltage and current are inputs provided by the NodeMCU ESP32 microcontroller to the monitoring tool. Date functions to display the time data is recorded as well as a pause for each incoming data per minute

2.2.5 A diagram of how the tool works.

The working principle of this tool is the NodeMCU ESP32 module programmed to read the voltage and electric current sensor . Data received from the sensor is processed by Arduino and sent to the web server. So that it can be monitored and can be controlled directly by an Android-based smartphone or laptop through the website. Block diagram of how the tools used in this study can be seen in Figure 4 .

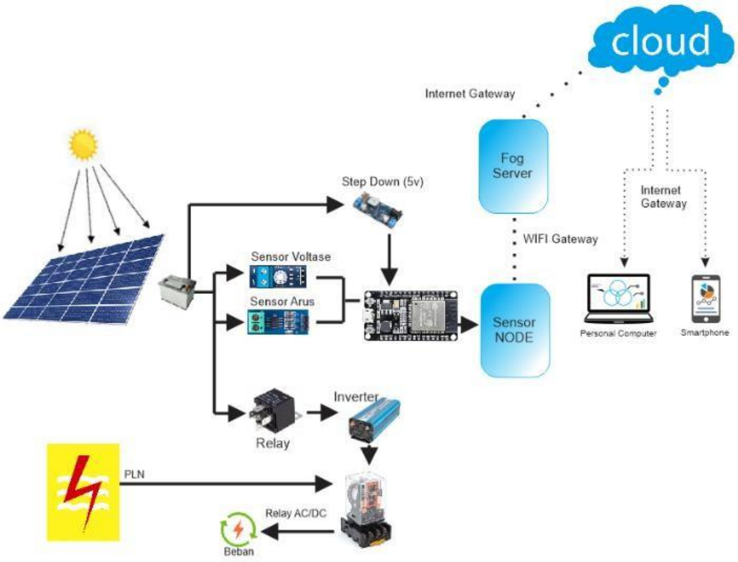


Figure 4. Diagram of how the tool works

2.3 Coding

Retrieval of data from the database using the PHP programming language. Data is changed in the form of JSON first so that it can be displayed on the battery monitoring application. Making an application android smartphone using thinkable. NodeMCU coding uses Arduino IDE which is given a code to connect to a wifi network using the programming language C. Relay Automatic Transfer Switch is controlled if the sensor reads the voltage on the battery less than 11.4V.

2.4 Testing

At this stage the test is done after the tool can be used. The test used is the black box testing method. Black box testing focuses on the functional specifications of the software.

Black box testing tends to find the following things [15]:

a. Incorrect or non-existent function.
b. Interface errors.
c. Errors in data structures and database access.
d. Performance errors.
e. Error initialization and termination.

In black box testing, special knowledge of application code or internal structure and programming knowledge are generally not needed. This test allows developers to obtain a series of input conditions that meet the functional requirements of a program [16] . The advantage of black box testing is that testing can be done from the user's point of view and helps to expose ambiguity or inconsistency in due diligence [17].

3. RESULT AND DISCUSSION

3.1 Result

The results of this study are the monitoring of PV-VP monitoring tool with an android application connected to the Internet of Things (IoT). Error testing and stability of data flow testing are done using Exponential Smoothing. The monitoring tool functions as a research object where voltage and current variables can be monitored through the Android application .

3.1.1 Application design

Visible page real time which contains the value ofthe data voltage, current, and estimated battery usage time da graphs can be seen in Figure 5 .

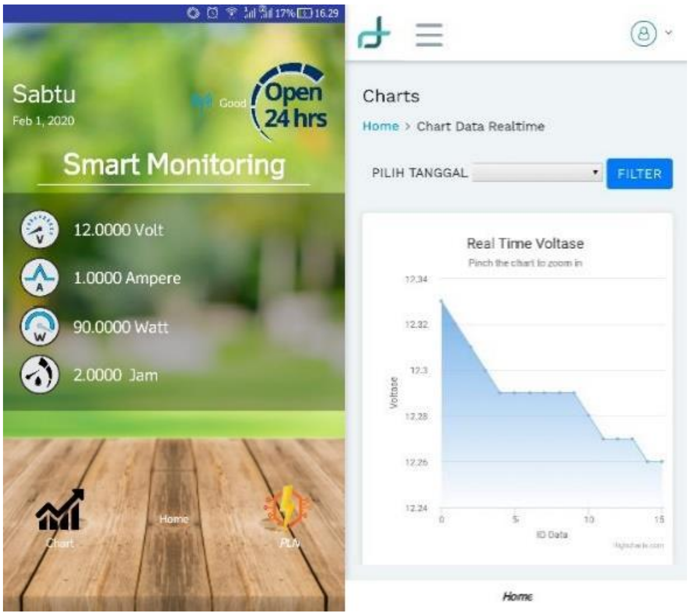


Figure 5. Realtime pages and graph

3.1.2 Testing

Testing is done to test the tool that has been made in accordance with the function and without constraints. The testing to be carried out can be seen in Table 1

Table 1. Testing Tools

No	Requirements tested	Test Item	Conclusion
1	Relay	Able to switch to DC12V input when the battery is more than 11.4V	Valid
2	Sensor	Read battery capacity, current and power used	Valid
3	LCD display	Displays readings of wifi connections, battery status, relay status, data upload status to the web server	Valid
4	ESP32	Sending sensor data to the database	Valid
5	Send Data	The database can store data that has been sent by ESP32	Valid
6	Battery	The battery can power the inverter by supplying a monitoring device with a DC12V voltage	Valid
7	Inverter output	The inverter can turn on the AC220V load by supplying a monitoring device	Valid
9	AC220V output	Can turn on the load	Valid

Tool testing is done using a 1200W inverter and a 40Ah 12V battery and a load of 6 9W AC lamps. testing is done by turning on the lights continuously until the battery can not supply electricity and replaced with PLN.

Sending data in the form of battery capacity, electric current and electric power used for 5 seconds. in the graph above can be seen the voltage based on the data received last to receive information that the battery capacity is 11.4 V. There is a possibility a few moments later before sending data back the battery capacity is 11.4V and the relay automatically cuts off the electricity coming from the battery.

Furthermore testing is carried out on the application to ensure that the application has been made in accordance with the previous plan. In testing this application uses black box testing. The following results of testing with black box testing can be seen in Table 2.

Table 2. Testing of application

No	Testing	Expected results	Test result	Conclusion
1	Open application	Preliminary results	Shows the start page	Valid
2	Showing data record	Displays all data records that have been stored in the database	Data can be displayed in accordance with the database	Valid
3	Showing graph	Displays usage graph	Displays usage graph	Valid
4	Change the date	Change the date and display the usage graph for that date	Date changed and successfully displayed the usage graph according to the desired date	Valid
5	Showing realtime data	Display voltage data in realtime	Display voltage data in realtime	Valid

3.1.3 Evaluating

The evaluation phase is the stage for evaluating the whole starting from the experimental stage, designing tools, making applications, coding and testing. Errors found in the testing phase both in design and coding are evaluated in order to function properly. Evaluation of tools and applications is done when the device is connected to a battery and given an electric load. The following list of evaluations is in Table 3 .

Table 3. Research evaluation

No	Error	Evaluation
1	Usage data cannot be displayed in the application	Check internet connection
2	On the LCD the device only brings up Search for Wifi	Check wifi network
3	The inverter makes a noise	Turn off and turn on the Inverter again

3.2 DISCUSSION

Designing monitoring tools on solar power plants with the internet of things. There are two main components, namely smart monitoring and the Automatic Transfer Switch system. In smart monitoring using NodeMCU as the control brain and coding using Arduino IDE. Data received by the sensor on the monitoring tool will be sent to the database via the internet. Then the data will be retrieved for display on the monitoring application. The application will display data in the form of realtime data and graphics on a predetermined date. The Automatic Transfer Switch (ATS) system works if the battery capacity sensor has read less than 11.4V .

4. CONCLUSION

Based on the discussion of this research, the design of a monitoring tool on solar power plants with the concept of the internet of things is making Android applications using thunkable and using NodeMCU as the control center, and Automatic Transfer Switch (ATS) as a battery controller to move the power source. The results of this study are:

a. Users can view data in realtime in the form of battery capacity, electric current and electric power used through an application on an Android smartphone. The data is obtained from sensors located on monitoring devices that are connected to the internet network so that the data is sent and stored on a database server. Based on black-box testing and testing tools produce a percentage of 100%. Android application displays the results of recorded data on the webserver to the smartphone screen. With an average data change rate of 0.005% for each data recording.

b. The Automatic Transfer Switch (ATS) system works if the battery capacity sensor has read less than 11.4V then the automatic relay ATS will move the electricity needs to the PLN so that the user does not need to move the electric current manually to the PLN and does not worry the battery will overload due to usage The battery is controlled by the ATS system.

REFERENCES

[1] F. Xia, L.T. Yang, L.Wang, and A. Vinel, "Editorial Internet of Things, " *International Journal of Communication Systems*, vol. 25, no.1, pp. 1101-1102, 2012..

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A Vision, architerctural elements, and future directions, " *Future generation computer systems*. vol. 29, no. 7, pp. 1645-1660, 2016.

[3] APJII. 2016. *Survey Internet APJII 2018*. [Online], Available: http://www.apjii.or.id/survey2018.

[4] R. Roviaji, and M. A, Muslim. "Pembuatan Sistem Informasi Gardu Induk PT.PLN (Persero) App Semarang Se-Kota Semarang Dengan Java Android, " in Proc. *Seminar Ilmu Komputer Dan Teknologi Informasi*, Samarinda, Indonesia, 2017.

[5] A. Junaidi. "Internet of Things, Sejarah, Teknologi dan Penerapannya: Review, " *Jurnal ilmiah Teknologi Informasi Terapan*, vol. 1, no. 3, pp. 62-66, 2015.

[6] M. Malenko, and M. Baunach "Real-time and Security Requirements for Internet-of-Things Operating Systems. In: Halang W., Unger H. (eds) Internet der Dinge, " Informatik aktuell DOI: 10.1007/978-3-662-53443-4_4.

[7] M. H. Miraz, M. Ali, P. S. Excell, and R. Picking, "A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of nano things (IoNT)" in Proc. international IEEE conference on Internet Technologies and Applications, Wrexham, North East Wales, UK, 2015, pp 219-224.

[8] S. Abdelwahab, B. Hamdaoui, M. Guizani and A. Rayes, "Enabling smart cloud services through remote sensing: An internet of everything enabler, " IEEE Internet of Things Journal, vol. 1, no. 3, pp. 276-288, 2014.

[9] S. Kumar, "Ubiquitous Smart Home System Using Android Application, " *Int. J. of Comp. Net. & Comm.* vol. 6, no. 1, pp. 33-43, 2014.

[10] A. Soetedjo, Y. I. Nakhoda, and C. Saleh, Choirul, "Embedded Fuzzy Logic Controller and Wireless Communication for Home Energy Management Systems, " *Electronics*, vol. 7. no. 9. pp 1-21, 2018.

[11] T. Kaur, J. Gambhir, and S. Kumar, "Arduino based solar powered battery charging system for rural SHS" presented at 7th India Int. Conf. on Pow. Elec., Patiala, India, Nov. 17-19, 2016.

[12] C.K. Amuzuvi, and E. Addo, "A microcontroller-based *Automatic Transfer Switch*ing system for a standby electric generator, " *Ghana Mining J*, vol. 15., no. 1, pp.85-92. 2015.

[13] R.S. Pressman and B. Maxxim. Software engineering: a practitioner's approach. Palgrave macmillan. 2015

[14] T.Murnane , K. Reed and R. Hall, "On the learnability of two representations of equivalence partitioning and boundary value analysis, " presented at Australian Software Engineering Conference , Melbourne, Vic., Australia, April 10-13, 2007.

[15] E. A. Wibowo, and R. Arifudin, "Aplikasi Mobile Learning Berbasis Android, " *Unnes Jurnal of Mathematics*. vol.5, no. 2, pp. 108-117, 2016.

[16] S. Nidhra, and J. Dodenti, "Blackbox and Whitebox Testing Techniques – A Literature Review, " *Int. J. of Embedded Systems and Applications*, vol. 2. no. 2, pp. 29-50, 2012.

JOSCEX Journal of Soft Computing Exploration

SVM Optimization with Correlation Feature Selection Based Binary Particle Swarm Optimization for Diagnosis of Chronic Kidney Disease

Doni Aprilianto
Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jul 24, 2020
Revised Aug 6, 2020
Accepted Sept 2, 2020

Keywords:

Support Vector Machine;
Correlation-Based Feature Selection;
Chronic Kidney Disease;
Binary Particle Swarm Optimization.;

ABSTRACT

Data mining has been widely used to diagnose diseases from medical data. In this study using chronic kidney disease dataset taken from UCI Machine Learning. The dataset has 25 attributes with 400 samples. With 25 attributes that allow redundant data. Redundant data in datasets can reduce computational efficiency and classification accuracy. To increase accuracy of classification algorithm can be done by reducing dimensions of dataset. Correlation-based Feature Selection (CFS) can quickly identify and filter redundant attributes. However, CFS has disadvantage that selected attribute is not necessarily the best attribute. These weaknesses can be overcome by Binary Particle Swarm Optimization (BPSO). BPSO chooses attributes based on the best fitness value. The purpose of this study is to improve accuracy of Support Vector Machine (SVM) by implementing combination of CFS and BPSO as feature selection. Accuracy of SVM in predicting CKD is 63.75%. Whereas, accuracy of SVM by applying CFS as feature selection is 88.75% and average accuracy of ten execution SVM algorithms by applying a combination of CFS and BPSO as feature selection is 95%. Thus, combination of CFS and BPSO as feature selection on the SVM algorithm can improve results of accuracy in diagnosing CKD by 31.25%.

Corresponding Author:

Don Aprilianto
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: doniapr14@students.unnes.ac.id

1. INTRODUCTION

In recent years, data mining has been widely used in the health sector, bioinformatics, banking, document classification marketing, etc. [1]. in the health sector, data mining is used to diagnose diseases such as breast cancer, diabetes, heart disease and others [2]. To predict a decision in data mining can use classification techniques. Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Artificial Neural Network are examples of classification algorithms [3].

Popular algorithm for data classification is support vector machine (SVM). However, SVM requires a large amount of memory if the processed data has high dimensions or has a lot of features [4]. Accuracy and computational efficiency of a classification algorithm are strongly influenced by datasets that process, datasets with high dimensions can present data that is redundant and irrelevant so that it will affect the performance of the classification algorithm [5].

To improve accuracy and performance of classification algorithm can be done by reducing dimensions and eliminating redundant data. This process is called feature selection. The feature selection algorithm is divided into 3 main categories namely, filter, wrapper and hybrid. The filter method has a fast computation time, but not necessarily finding the best combination of subset attributes. The wrapper method will produce the best combination of subset attributes, but requires large computer memory. The hybrid method will combine the filter method and the wrapper method to get the advantages of each method. In hybrid method, filter method is used to reduce dataset dimensions and wrapper method will be used to find the best combination of attributes [6].

One of the algorithms in filter method that effective for handling redundant and irrelevant data is correlation-based feature selection (CFS). CFS is an algorithm that ranks subset attributes and finds relevant attributes based on correlation-based heuristic evaluation functions [7]. CFS can quickly identify and filter out irrelevant and redundant attributes [8]. CFS will choose attributes that have a strong correlation with the target class but do not correlate with other attributes. However, CFS does not necessarily choose attributes that provide the best accuracy results if the data sample is limited [5]. To get the best combination of attributes and good correlation, CFS can be combined with the wrapper method.

Over the past few years, many wrapper methods have been developed for attribute selection. An example of a wrapper algorithm developed is evolutionary programming (EP), ant colony optimization (ACO), differential evolution (DE), genetic algorithms, particle swarm optimization (PSO) [9]. For the optimization, PSO gives more competitive results than Genetic Algorithms [2]. At PSO, each particle is flown in search space to find the best solution (fitness) called pbest. Then, the overall best value is called gbest. To overcome the feature selection problem, the particles in PSO will be represented in binary form which is then called Binary Particle Swarm Optimization [10].

In this study propose a combination of CFS algorithm and BPSO algorithm as feature selection to improve the accuracy of the SVM algorithm for diagnosing chronic kidney disease (CKD). CFS was chosen because it can choose attributes that have good correlation results and delete redundant data quickly, while BPSO was chosen because it was able to provide a combination of attributes that produced the best accuracy.

2. METHOD

In this study, the combination of CFS and BPSO was carried out as a feature selection. CFS is used to reduce the dimensions of the dataset based on the correlation between features and target class but does not correlate with other features. BPSO is used to find the best combination of features. The classification method used is the Support Vector Machine algorithm. From the classification results, we will get an increase in accuracy from Support Vector Machine before and after the combination of CFS and BPSO is applied. The flowchart of the method used in this study is shown in Figure 1.

2.1 Data preprocessing

The data used in this study is the Indian Chronic Kidney Disease taken from the UCI Machine Learning Repository. This dataset has 25 attributes, of which 11 attributes are numeric and 14 are nominal.

The CKD dataset has 400 data samples and there are more than 15% missing values. With a missing value of more than 15% it will greatly affect the performance of the classification model that has been formed [11]. In this study, the handling of the missing value is done by using the most frequent.

2.1.1 Correlation-based feature selection

Correlation-based Feature Selection (CFS) is a filter algorithm that ranks subset attributes according to heuristic evaluation functions based on correlation [12]. CFS will evaluate features by considering the predictive capabilities of each feature and the level of redundancy between them. If the correlation between attributes and class is known, and the correlation between each attribute is given, then the correlation can be predicted by using Eq. 1.

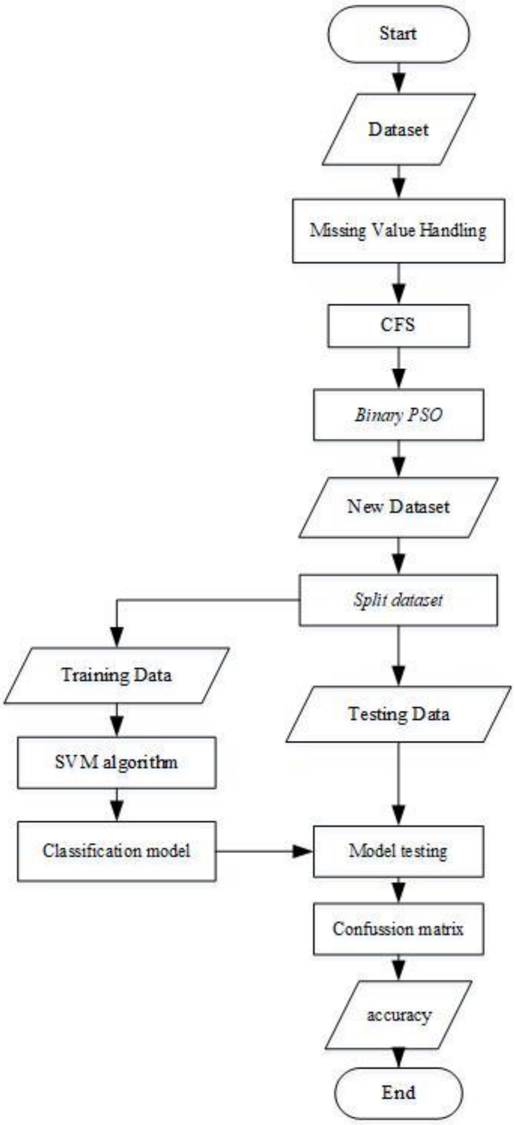


Figure 1. Flowchart SVM with CFS and BPSO as feature selection.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \tag{1}$$

where Ms is correlation between summed components and external variables. K is number of components. rcf is average correlation between components and outside variables.rff is correlation between average components.

CFS is an automatic algorithm that does not require users to specify the number of features to be selected. CFS uses the best first search method to get the best features from a feature subset. Then the correlation between features can be calculated using symmetrical uncertainty (SU) as Eq. 2.

$$SU = 2.0 \times \left[\frac{H(S_j) + H(S_i) - H(S_i,S_j)}{H(S_j) + H(S_i)}\right] \tag{2}$$

Where H(Sj) dan H(Si) is entropy of attributes. H(S_i, S_j) is conditional entropy.

2.1.2 Binary particle swarm optimization

Binary Particle Swarm Optimization (BPSO) was introduced by Kennedy and Eberhart in 1997. BPSO is used to solve discrete optimization problems. The difference between PSO and BPSO lies in the representation of the particles [10]. In BPSO each particle is shown in discrete values. While in standard PSO, particles are represented in continuous values. The concept of PSO is that each particle is flown in search space to find the best solution (fitness) called pbest. Then, the best overall value (global value) called gbest. Each particle has two vectors namely position vectors and velocity vectors to move around in search space. Each particle has memory and each particle will track the best position beforehand [13].

The stages of the binary particle swarm optimization algorithm in the feature selection are as follows:

Step 1 : Initialize particles with random velocity and positions.
Step 2 : Calculate the fitness value of each particle in the population.
Step 3 : If fitness value of particle i is smaller than the fitness value of pbest, then set pbest from particle i to particle position i.
Step 4 : If pbest value is less than the current gbest value, then set gbest to the current pbest.
Step 5 : Update position and velocity of the particles by using Eq. 3 and 4.

$$Vid_{k+1} = w \times \ Vid_k + c_1 \ \times \ rand1 \times (Pid - Xid) + c_2 \times rand2 \times (Gid - Xid) \tag{3}$$

$$Xid_{k+1} = Xid_k + Vid_{k+1} \tag{4}$$

Step 6 : Where Vid is the individual velocity. Xid is position of individual. w is inertia weight parameter. c_1 and c_2 is learning rate constant, the value is between 0 and 1. rand1 and rand2 is random parameter between 0 and 1. Pid is Pbest (personal best) individual i in d dimension. Gid is Gbest (global best) in d dimensions.
Step 7: Iteration will stop if maximum generation is fulfilled; if not return to step 2.

2.2 Support vector machine

Support Vector Machine (SVM) was first proposed by Vladimir Vapnik. Proposed in the field of statistical learning theory and structural risk minimization [14]. SVM has been used in a variety of problems such as data classification, image classification, text categorization, tone recognition, digit recognition of handwriting [15].

The stages of the Support Vector Machine algorithm in classifying datasets are as follows:
Step 1: Prepare training data. The training data consists of 80% of the entire dataset.
Step 2: Finding boundaries between classes. When each point in a class is connected to another point, a line that separates between the classes will appear. This limit is known as the convex hull. Each class has its own convex hull and because the class (assumed) is linearly separated, this hull does not intersect.
Step 3: Determine a hyperplane that maximizes the margin between classes. Can be done in the following ways:

a. First, any hyperplane stated in two attributes, $x_1$and x_2, can be written Eq. 5.

$$w \ \cdot \ x + b = 0 \tag{5}$$

where w is weight ($w = w_1, w_2, \dots, w_n$), x= number of attributes ($x = x_1, x_2, \dots, x_n$), b = bias.

b. An optimal hyperplane, defined uniquely by $b_0 + w_0 \cdot x = 0$. After defining a hyperplane in this mode, it can determine the margin. Margin can be written Eq. 6.

$$margin = \frac{2}{\sqrt{w_0}} \,.\, w_0 \tag{6}$$

c. Maximizing this quantity requires quadratic programming, which is a process that has a strong position in the theory of mathematical optimization. Furthermore, w can be easily stated in some examples of training data, known as support vectors, can be written Eq. 7.

$$|w_0 \ = \ \sum y_i x_i| \tag{7}$$

where y_i is the class label and x_i is called support vector. i is a zero coefficient only for support vector.

Step 4: After setting boundaries and hyperplane, each new test can be classified by calculating on which side of the data results in the hyperplane. This can be found by replacing the test x example into the hyperplane equation. If you count +1, then it includes a positive class and if it is calculated as -1, then it belongs to the negative class.

2.3 Evaluation

The proposed method begins by dividing the dataset into training data and testing data. In this study the data distribution was done using the splitter method. This method divides the data into two subsets with a proportion of 80% for training data and 20% for testing data.

Measurement of classification performance is done by confusion matrix, confusion matrix is a useful tool to analyze how well the classifier recognizes tuples from different classes. Confusion matrix is done by calculating the number of predicted classes against the actual class. These results are expressed in True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). FP and FN state that the classifier is wrong in recognizing tuples, meaning positive tuples are recognized as negative and negative tuples are recognized as positive. While TP and TN state that the classifier recognizes tuples correctly, meaning positive tuples are recognized as positive and negative tuples are recognized as negative. The matrix confusion table can be shown in Table 2.

Table 1. Confusion matrix

Classification		**Predicted class results**		
		Yes	**No**	Total
Actual Class	**Yes**	TP	FN	P
	No	FP	TN	N
	Total	*P'*	*N'*	P+N

Accuracy is the percentage of the total data classified correctly. Accuracy measurements can be written by using Eq. 8.

$$Accuracy = \frac{TP+TN}{P+N} \text{ x } 100\% \tag{8}$$

3. RESULT AND DISCUSSION

In this study, the proposed algorithm testing uses the Python programming language by utilizing a scikit-learn library, sk-feature and pyswarms library. The data used is the CKD dataset taken from the UCI Machine Learning. This dataset has 25 attributes that have 1 class and 24 attributes.

3.1 Correlation Based Feature Selection Process

CFS will choose attributes that have the highest correlation weighting value. From the CFS process, 13 selected attributes were obtained. The list of attributes and weights of the CFS process is shown in Table 3.

Tabel 2. List of attributes and results of CFS weight

No	Attributes	CFS Weight
1	*Pc*	0,54636158
2	*Pe*	0,54568856
3	*Appet*	0,54506278
4	*Bp*	0,54405311
5	*Rc*	0,54397886
6	*Ane*	0,54383461
7	*rbc*	0,54211877
8	*Cad*	0,53743288

9	*Al*	0,53624068
10	*Pcv*	0,53390589
11	*Dm*	0,50241938
12	*Sg*	0,4529584
13	*Htn*	0,35495722

3.2 Binary Particle Swarm Optimization Process

Attributes chosen by the CFS algorithm, do not necessarily produce the best combination of attributes. Therefore, the BPSO algorithm is used to determine the best feature combination of the attributes chosen by CFS. At this stage, 10 tests are executed to determine the best combination of features. The BPSO parameters used in this study are shown in Table 4.

Table 3. BPSO parameters

Parameters	Value
Number of particles	60
Iteration	100
Inertia weight	0,9
C1	2
C2	2

3.3 Application of Algorithms

At this stage, 3 tests were carried out, namely stand alone SVM algorithm, SVM algorithm by implementing CFS and SVM algorithm by implementing a combination of CFS and BPSO. In the first application, the SVM algorithm will process the CKD dataset with 25 attributes. The application of SVM algorithms gets an accuracy of 63.75%. The results of this accuracy state that the SVM algorithm can classify CKD datasets well because the accuracy results are greater than the error rate. However, the results of this accuracy can be improved by applying several preprocessing methods.

The second application, the SVM algorithm will be combined with the CFS algorithm. So SVM will process the CKD dataset with 13 attributes and 1 class. The accuracy of this classification model is 88.75%. The accuracy of the application of this model can increase the accuracy of the SVM algorithm by 25%. However, these results can still be improved by selecting the best feature combination using the BPSO algorithm.

The third application, the SVM algorithm will be combined with the CFS and BPSO algorithms. In this implementation, 10 tests were executed to determine the best combination of features. The accuracy of this classification model can be seen in Table 5

Table 4. Average SVM Accuracy Results with a combination of CFS and BPSO as feature selection

Execution	Number of attributes	Accuracy (%)
1	12	96,25 %
2	10	95 %
3	11	95 %
4	10	96,25 %
5	10	93,75 %
6	11	95 %
7	11	96,25 %
8	12	96,25 %
9	9	91,25 %
10	10	95 %
mean	10,6	95

The accuracy of the application of this model can increase the accuracy of the SVM + CFS algorithm by 6.25% and can increase the SVM algorithm by 31.25%. comparison of accuracy of each application of the algorithm can be seen in Figure 2.

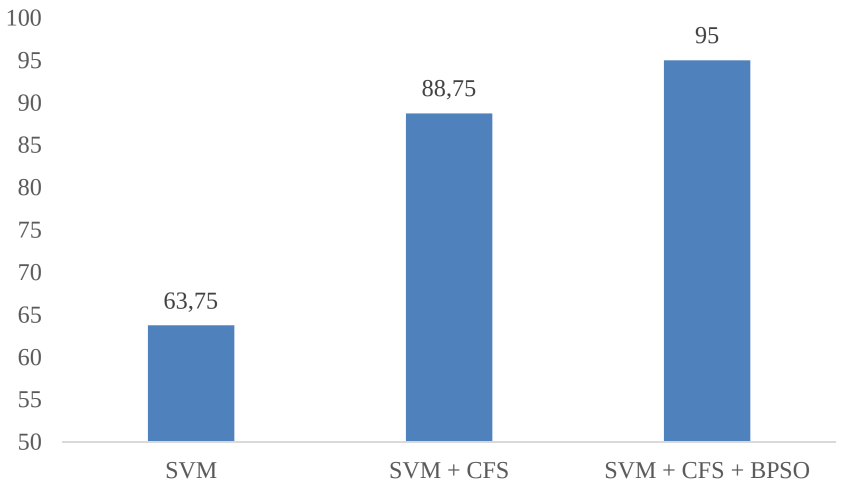


Figure 2. The comparison result of accuracy

4. CONCLUSION

In this study, the testing of SVM algorithm by applying the CFS algorithm and the BPSO algorithm was carried out using the CKD dataset taken from the UCI Machine Learning Repository. CFS algorithm is used to get attributes with good correlation while BPSO is used to get the best combination of attributes. The results of this study showed that the accuracy of the application of the SVM algorithm was 63.75%, while after the CFS algorithm was used, the accuracy of 88.75% and the accuracy of ten SVM algorithms was obtained by applying a combination of feature selection CFS and BPSO of 95 %. Thus, it can be concluded that the application of a combination of CFS and BPSO as feature selection on SVM algorithm can improve the results of accuracy in diagnosing CKD by 31.25%.

REFERENCES

[1] C. Sreedhar, N. Kasiviswanath, and P. C. Reddy, “Clustering large datasets using K-means modifed inter and intra clustering (KM-I2C) in Hadoop, ” *J. of Big Data*, vol. 27, no. 4, pp. 1-19, 2017.

[2] M.A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, “Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis, ” presented at the 5th Int. Conf on Mathematics, Science and Education, Bali, Indonesia, Oct. 8–9, 2018.

[3] D. O. Sahin, and E. Kılıc, “Two new feature selection metrics for text classification, ” *Automatika*, vol.60, no. 2, pp. 162-171, 2019

[4] V. Kotu, and B. Deshpande, Predictive Analytics and Data Mining. Massachusetts, USA: Morgan Kaufmann, 2015, pp. 63-163.

[5] I. Jain, V. K. Jain, and R. Jain, “Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification, ” *Appl. Soft Comp.*, vol. 62, no.-, pp. 203-215. 2018.

[6] K. Sutha, and J. J. Tamilselvi, “A review of feature selection algorithms for data mining techniques, ” *Inte. J. on Comp. Sci. & Eng.*, vol. 7, no. 6, pp. 63-67, 2015.

[7] P. Yildirim, “Filter based feature selection methods for prediction of risks in hepatitis disease, ” *Int. J. of Mach. Lear. & Comp.*, vol. 5, no. 4, pp. 258-263. 2016.

[8] S. Sasikala, S. Appavu, and S. Geetha, “Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set, ” *Appl. Comp. & Info.*, vol. 12. no.-, pp. 117-127. 2017.

[9] I.A. Ashari, M.A. Muslim, and Alamsyah. “Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing, ” *Sci. J. of Info.* vol. 3, no. 2, pp. 149-158, 2016.

[10] M.S. Muhammad, K.V. Selvan, S.M.W. Masra, Z. Ibrahim, and A.F.Z. Abidin, “An Improved Binary Particle Swarm Optimization Algorithm for DNA Encoding Enhancement, ” presented at the IEEE Symposium on Swarm Intelligence, Paris, France, April 11-14, 2011.

[11] W. Abedalkhader, and N. Abdulrahman, “Missing Data Classification of Chronic Kidney Disease, ”, *Int. J. of Data Mining & Knowledge Manage. Process*, vol. 7, no. 5, pp. 55-61.2017

[12] N. Gopika, and A. M. E. M. Kowshalaya, “Correlation based feature selection algorithm for machine learning, ” presented at the 3rd Int. Conf. on Commu. & Elec. Sys., Coimbatore, India, Oct. 15-18, 2018.

[13] N. D. Jana, and J. Sil, “Interleaving of Particle Swarm Optimization And Differential Evolution Algorithm For Global Optimization, ” *Int. J. of Comp. & Appl.,* vol. 38. no. -, pp. 116-133, 2016.

[14] J. Nayak, B. Naik, and H. S. Behera, “A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges, ” *Int. J. of Data. Theory & Appl.*, vol. 8, no.-, pp. 169-186. 2016.

[15] D. K. Srivastava, and L. Bhambhu, “Data classification using support vector machine, ” *J. of Theoretical & Appl. Inf. Tech*., vol. 12, no.-, pp. 1-7. Feb 2010.

Journal of
Soft Computing
Exploration

Increasing Accuracy of C4.5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease

Aprilia Lestari[1], Alamsyah[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 2, 2020
Revised Aug 20 , 2020
Accepted Sept 12, 2020

Keywords:

Data Mining;
C4.5;
Gain Ratio Boosting;
CKD;

ABSTRACT

Data information that has been available is very much and will require a very long time to process large amounts of information data. Therefore, data mining is used to process large amounts of data. Data mining methods can be used to classify patient diseases, one of them is chronic kidney disease. This research used the classification tree method classification with the C4.5 algorithm. In the pre-processing process, a feature selection was applied to reduce attributes that did not increase the results of classification accuracy. The feature selection used the gain ratio. The Ensemble method used adaboost, which well known as boosting. The datasets used by Chronic Kidney Dataset (CKD) were obtained from the UCI repository of learning machine. The purpose of this research was applying the information gain ratio and adaboost ensemble to the chronic kidney disease dataset using the C4.5 algorithm and finding out the results of the accuracy of the C4.5 algorithm based on information gain ratio and adaboost ensemble. The results obtained for the default iteration in adaboost which was 50 iterations. The accuracy of C4.5 stand-alone was obtained 96.66%. The accuracy for C4.5 using information gain ratio was obtained 97.5%, while C4.5 method using information gain ratio and adaboost was obtained 98.33%.

Corresponding Author:

Aprilia Lestari
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: aprilialestari11@gmail.com

1. INTRODUCTION

Along with the rapid development of technology development, people can easily access information anytime and anywhere. Data information that has been available is very much and will require a very long time to process large amounts of information data. Therefore, to process large amounts of data, data mining techniques are used. Data mining can be applied in various fields. One of them in the health sector is to predict and classify a disease from a patient's medical record data. Data mining methods can be used to classify patient diseases based on the severity of the disease, one of which is to classify patients with kidney disease and not. Chronic kidney disease or kidney failure is a very serious problem in all corners of the world, where the kidneys are damaged and become a cause of non-maximal kidney function [1].

Data mining is used to determine patterns in data mining knowledge and is useful in solving data problems in large data warehouses [2]. The term Data mining is also referred to as knowledge discovery. Data mining has a variety of types of methods, for that the selection of the right method will depend on the purpose and process. One of the methods in data mining is classification. The classification method has input in the form of a collection of records, where each record is marked with tuple (x.y). X is an attribute and Y is a specific attribute / target showing the class label. Classification has several algorithms including Naïve Bayes and C4.5, each of which has different accuracy [3]. Some techniques that exist in classification, decision trees are classification techniques that are very popular and widely used.

Decision tree is the most powerful approach in scientific discovery and data mining, and a very effective tool in various fields such as data and text extraction, information extraction, machine learning, and pattern recognition [4]. One of the most popular decision tree techniques is C4.5. C4.5 algorithm is one of the algorithms developed by J. Ross Quinlan which is the development of algorithm ID3 (Iterative Dichotomiser 3) [5].

This research was conducted using Chronic Kidney Disease Dataset obtained from the UCI repository of learning machine. The following are some of the researches that are relevant to CKD. In the research [6] compared two algorithms namely C4.5 standalone and C4.5 with Pessimistic prunning applied to the Chronic Kidney Disease dataset. Standalone C4.5 has an accuracy of 95% and C4.5 with pressimistic prunning resulting in an accuracy of 96.5%. Based on research [7] discusses the prediction of chronic kidney sufferers using decision tree and naïve bayes algorithms. The dataset used for this research is the chronic kidney disease dataset. The results of this research are decision trees resulting in an accuracy of 91% and Naïve Bayes resulting in an accuracy of 86%. In the research by [8] which states that the C4.5 algorithm has the highest accuracy when applied to the Chronic Kidney Disease dataset compared to the Expectation Maximization (EM) and Artificial Neural Network (ANN) algorithms. The C4.5 algorithm produces an accuracy of 96.75%, EM 70% and ANN 75%.

For improving the accuracy of the C4.5 algorithm, this research used methods in pre- processing and ensemble methods. In the pre-processing process, a feature selection is applied to reduce attributes that do not increase the results of classification accuracy. The feature selection used the gain ratio. The Ensemble method used adaboost which well know as boosting.

The purpose of this research was applying the Information Gain Ratio and Adabost ensemble to the Chronic Kidney Disease dataset using C4.5 algorithm and finding out the results of the accuracy of the C4.5 algorithm based on Information Gain Ratio and Adaboost ensemble.

2. METHOD

2.1 Feature Selection

Feature selection is the process for selecting a subset of original attributes, so that the feature space optimally decreases according to certain criteria. Feature selection which aims to reduce the number of certain features focusing on relevant data and improve quality therefore Feature selection is able to work better than processes that are driven by the selected features [11].

2.2.1 Information Gain Ratio

Information gain ratio is the ratio of obtaining information gain with intrinsic information. To reduce the bias towards multi value attributes by taking the number and size of branches in a calculation when selecting attributes. This is useful as a consideration for logarithmic probabilities to measure the impact of this type of calculation in a dataset.

2.2 Algoritma C4.5

The C4.5 algorithm was introduced by Quinlan as an improved version of ID3. In ID3, induction of decision trees can only be done on categorical type features (nominal / ordinal), while numerical types (internal / ratio) cannot be used. The improvement that distinguishes C4.5 algorithm from ID3 is that it can handle features with numeric types, pruning decision trees, and deriving rule sets. The C4.5 algorithm also uses gain criteria in determining features that are node breakers in the induced tree [12].

In the C4.5 algorithm, building a decision tree the first thing to do is to choose attributes as roots. Then a branch is created for each value in the root. The next step is to divide the case in branches. Then repeat the process for each branch until all the cases in the branch have the same class [13].

$$GainRatio(A) = \frac{Gain(A)}{SplitEntropy(A)} \tag{1}$$

For calculating the gain, Equation 2 is used as follows [15].

$$\mathrm{Gain}(S, A) = \mathrm{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \mathrm{Entropy}(S_i) \tag{2}$$

Description:
S = Set of Case
A = Attributs
n = The number of A Attribute Partition
$|S_i|$= The Case Number in i Partition
|S| = The Case Number

Meanwhile, calculating the entropy value can be seen in Equation 3.

$$\mathrm{Entropy}(S) = \sum\nolimits_{i=1}^{n} - p_i * \log_2 p_i \tag{3}$$

Description:
S = Set of case
n = The partition number of S
p_i = The proportion S_i to S, which $log_2 pi$ can be calculated using Equation 4.

$$\log(X) = \frac{\ln(X)}{\ln(2)} \tag{4}$$

Entropy is used to determine which node will be the next training data solver. A higher entropy value will increase the potential for classification. What needs to be considered is that if entropy for nodes is 0 means that all vector data are on the same class label and that node becomes a leaf containing a decision (class label). What also needs to be considered in the calculation of entropy is if one of the elements w_i is 0 then the entropy is confirmed to be 0 too. If the proportion of all w_i elements is equal, it is certain that entropy is worth [16].

For calculate Split Entropy Equation 5 is used as follows.

$$\mathrm{SplitEntropy}_A(S) = -\sum_{i+1}^{n} \frac{|Si|}{|S|} * \log_2 \frac{|Si|}{|S|} \tag{5}$$

Description:
S = Set of Case
A = Attributs
n = The number of A Attribute Partition
$|S_i|$= The Case Number in i Partition
|S| = The Case Number

2.3 Adaptive Boosting (Adaboost)

Adaboost is a boosting algorithm part of ensemble learning that is used to improve classification performance [17]. According to research conducted by Nurzahputra & Muslim [18] states that adaboost is a part of machine learning introduced by Freud and Schapire (1995) which is used to improve accurate prediction rules by uniting many inaccurate and weak regulations.

Adaboost and its variants have been successfully applied in several fields because of their strong theoretical basis, accurate predictions and great simplicity. The steps in the adaboost algorithm are as follows.

a. Input: A collection of research samples with labels {(xi,yi), …, (xn,yn)}, a component learn algorithm, the amount of rotation T.
b. Initialize: Weight of a training sample $w_i^1 = \frac{1}{N}$, for all i=1, …, N
c. Do for t= 1, …, T
   1) Use component learning algorithms to train a classification component, h_t, to weight of a training sample.
   2) Calculate the training error on $h_t = \varepsilon_t \Sigma_i^N = 1 w_i^t , y_i \neq h_t(x_i)$
      1) Determine the weight for component classifier $h_t = \propto_t = \frac{1}{2} \ln\left(\frac{1- \varepsilon_t}{\varepsilon_t}\right)$

3) Update weight of a training sample $w_i^{t+1} = \frac{w_i^t \exp\{-a_t y_i h_t(x_i)\}}{c_t}, i = 1, \dots, N$ C_t is a normalization constant.

4) Output: $f(x) = sign(\sum_t^T = 1 \propto_t h_t(x))$

3. RESULT AND DISCUSSION

In this research a web-based system was made using the Python programming language to find out the results of applying information gain ratio and Adaboost Ensemble to the C4.5 algorithm in the diagnosis of chronic kidney disease. To make it necessary data related to the diagnosis of chronic kidney disease that will be used as a testing system. This research used a chronic kidney disease dataset obtained from the UCI machine learning repository. This data consists of 24 attributes and 1 class.

At the data processing stage data processing was carried out before the algorithm is applied or commonly called pre-processing. The Chronic kidney disease dataset obtained is in the form of a file with an extension of .arff, changes to the file extension to .xlsx for data processing.

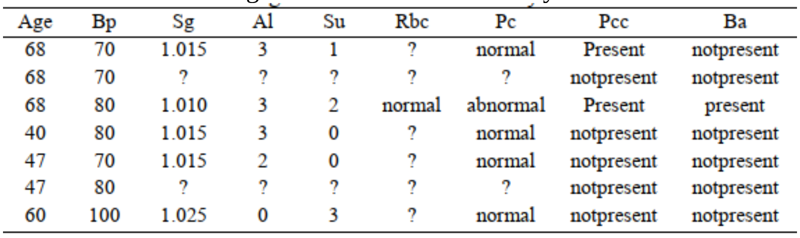
3.1 Formatiing Stage

The following formatting stage is the formatting of standards in the dataset used in the study. For example, in the Rbc (Red Blood Cells) attribute by changing the label on the Rbc attribute to 0 for negative (abnormal) and 1 for positive (normal).

3.2 Handling Missing Value Stage

Handling of missing value is part of pre-processing which aims to optimize mining results. Missing values in datasets are usually marked with the symbol "?" As in Table 1, which is an example of a Chronic Kidney Disease dataset that has a missing value.

Tabel 1. Missing value in the chronic kidney disease dataset



Age	Bp	Sg	Al	Su	Rbc	Pc	Pcc	Ba
68	70	1.015	3	1	?	normal	Present	notpresent
68	70	?	?	?	?	?	notpresent	notpresent
68	80	1.010	3	2	normal	abnormal	Present	present
40	80	1.015	3	0	?	normal	notpresent	notpresent
47	70	1.015	2	0	?	normal	notpresent	notpresent
47	80	?	?	?	?	?	notpresent	notpresent
60	100	1.025	0	3	?	normal	notpresent	notpresent

For replacing the missing value in the dataset using the calculation model mean (average) with the Equation 6.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6}$$

a. The value to replace the missing value in attribute Sg:

$$\bar{x}(Sg) = \frac{\sum_1^{359.145} x_{(Sg)}}{353} = \frac{359.145}{353} = 1{,}017$$

b. The value to replace the missing value in attribute Al:

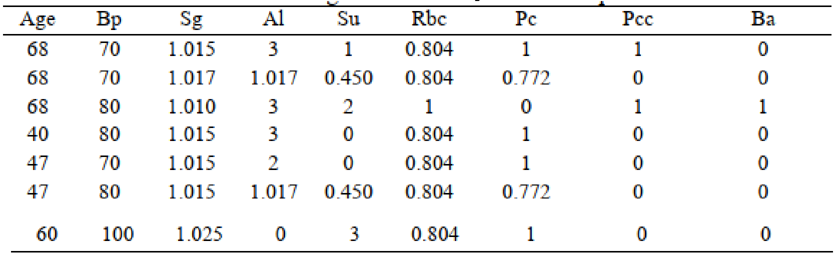
$$\bar{x}(Al) = \frac{\sum_1^{360} x_{(Al)}}{354} = \frac{360}{354} = 1{,}016949$$

c. The value to replace the missing value in attribute Su:

$$\bar{x}(Su) = \frac{\sum_1^{158} x_{(Su)}}{351} = \frac{158}{351} = 0{,}450142 = 0$$

The chronic kidney disease dataset that has been replaced is presented in Table 2.

Tabel 2. The chronic kidney disease dataset



Age	Bp	Sg	Al	Su	Rbc	Pc	Pcc	Ba
68	70	1.015	3	1	0.804	1	1	0
68	70	1.017	1.017	0.450	0.804	0.772	0	0
68	80	1.010	3	2	1	0	1	1
40	80	1.015	3	0	0.804	1	0	0
47	70	1.015	2	0	0.804	1	0	0
47	80	1.015	1.017	0.450	0.804	0.772	0	0
60	100	1.025	0	3	0.804	1	0	0

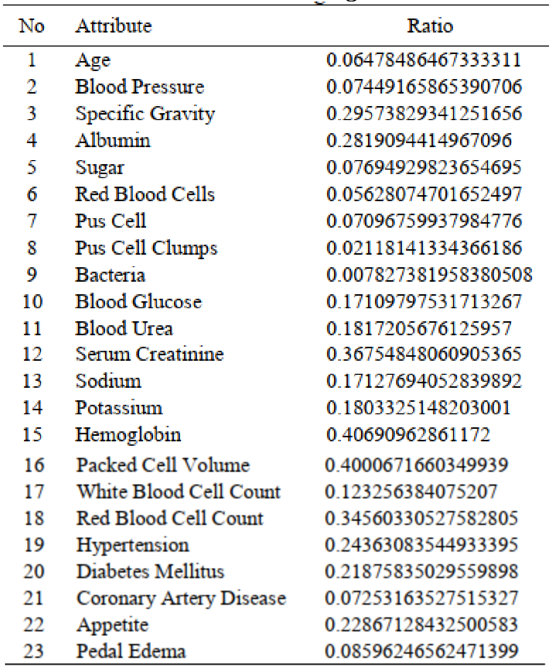
At the stage of class balancing is done by applying the SMOTE algorithm. The SMOTE algorithm is applied to make new data more balanced. German Credit's initial dataset has 1000 samples with 700 loyal (good) classes and 300 churn (bad) classes. Therefore it is necessary to balance the class by creating new data in the churn class. The new dataset of the SMOTE algorithm results in 300 churn class data, so there are 1300 new sample data. This is done so that data can be classified optimally. The attribute selection stage is done by selecting attributes in the data used. In this attribute selection stage there is a dimension reduction in the data in order to optimize attributes that will affect the accuracy of the Naive Bayes algorithm. Dimension reduction in attributes is done by using Genetic Algorithms. Removal of attributes is done one by one from attributes that have the smallest fitness value and will be mining. The process of selecting attributes and mining will stop when the results of the accuracy have exceeded the specified minimum limit.

After going through the pre-processing stage, new data will go through the classification process using the Naive Bayes algorithm. From the results obtained, there is an increase in the accuracy of the Naive Bayes algorithm and the Naive Bayes algorithm by applying the SMOTE algorithm and attibutes selection of Genetic Algorithms.

3.3 Feature Selection Implementation Stage

The stages of applying the feature selection are pre-processing steps in data mining to select features from the original attributes. In this study the application of a feature selection in the chronic kidney disease dataset aims to select attributes that fit certain criteria to improve quality so that optimal results are obtained. The results of information gain ratio for each CKD attribute are shown in Table 3.

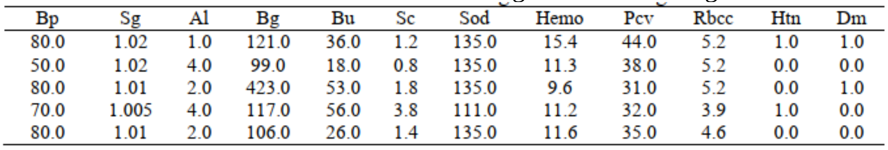
Tabel 3. Result of information gain ratio in CKD



No	Attribute	Ratio
1	Age	0.06478486467333311
2	Blood Pressure	0.07449165865390706
3	Specific Gravity	0.2957382934125l656
4	Albumin	0.2819094414967096
5	Sugar	0.07694929823654695
6	Red Blood Cells	0.05628074701652497
7	Pus Cell	0.07096759937984776
8	Pus Cell Clumps	0.02118141334366186
9	Bacteria	0.007827381958380508
10	Blood Glucose	0.17109797531713267
11	Blood Urea	0.1817205676125957
12	Serum Creatinine	0.36754848060905365
13	Sodium	0.17127694052839892
14	Potassium	0.1803325148203001
15	Hemoglobin	0.40690962861172
16	Packed Cell Volume	0.4000671660349939
17	White Blood Cell Count	0.123256384075207
18	Red Blood Cell Count	0.34560330527582805
19	Hypertension	0.24363083544933395
20	Diabetes Mellitus	0.21875835029559898
21	Coronary Artery Disease	0.07253163527515327
22	Appetite	0.22867128432500583
23	Pedal Edema	0.08596246562471399

After the feature selection calculation stage is carried out using the information gain ratio method, the results of the selected attributes are obtained. The results of the feature selection using the information gain ratio method are shown in Table 4.

Table 4. The results of the feature selection using the information gain ratio method



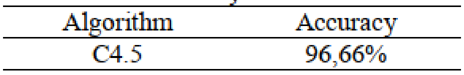
Bp	Sg	Al	Bg	Bu	Sc	Sod	Hemo	Pcv	Rbcc	Htn	Dm
80.0	1.02	1.0	121.0	36.0	1.2	135.0	15.4	44.0	5.2	1.0	1.0
50.0	1.02	4.0	99.0	18.0	0.8	135.0	11.3	38.0	5.2	0.0	0.0
80.0	1.01	2.0	423.0	53.0	1.8	135.0	9.6	31.0	5.2	0.0	1.0
70.0	1.005	4.0	117.0	56.0	3.8	111.0	11.2	32.0	3.9	1.0	0.0
80.0	1.01	2.0	106.0	26.0	1.4	135.0	11.6	35.0	4.6	0.0	0.0

3.3 Data Mining Stage

3.3.1 Implementation of C4.5 Algorithm

In this stage the model used is to apply the C4.5 algorithm to the CKD. New data that is ready to be processed is carried out by sharing training data as a model and testing data to measure the ability of the model formed. In this study the data distribution used a random sub sampling method. Where training data: data testing = 70%: 30% divided randomly. The application of the stand-alone C4.5 algorithm obtained an accuracy of 96.66% is presented in Table 5.

Table 5. The accuracy of C4.5 stand-alone

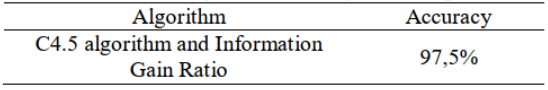


Algorithm	Accuracy
C4.5	96,66%

3.3.1 Implementation of C4.5 Algorithm and Information Gain Ratio

In this stage, the original attribute of the chronic kidney disease dataset consisted of 24 attributes and 1 class, after the information gain ratio method was applied as selecting attributes 12 attributes were selected. In this study the data distribution using the splitter method contained in the sklearn library, the method is random sub sampling. The data sharing system is by sub-sampling random method, where the data is divided into 70% and 30% and the data is taken randomly at each execution. The application of Information Gain Ratio to preprocessing data resulted in an accuracy of C4.5 is 97.5%, the results are presented in Table 6.

Table 6. The accuracy of C4.5 and information gain ratio

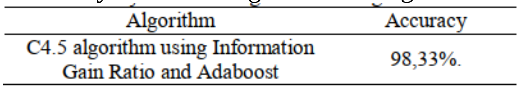


Algorithm	Accuracy
C4.5 algorithm and Information Gain Ratio	97,5%

3.3.1 Implementation of C4.5 Algorithm and Information Gain Ratio and Adaboost

The results of the decision tree will be known the gain value of the attributes that make up the dataset. From the gain value, each attribute is initialized as the initial weight in the calculation of adaboost. After the initialization weight is known, then iterations are determined in adaboost. The default iteration in adaboost is 50 iterations. The accuracy of the C4.5 algorithm based on information gain ratio and adaboost by using sub-random sampling as the splitter is 98.33%. The results of the accuracy obtained are presented in Table 7.

Table 7. The accuracy of C4.5 using information gain ratio and adaboost



Algorithm	Accuracy
C4.5 algorithm using Information Gain Ratio and Adaboost	98,33%.

The results of this accuracy are far better than just using the C4.5 or C4.5 algorithm based on information gain ratio only.

4. CONCLUSION

The application of information gain ratio and adaboost ensemble is a combination of two methods that are useful for increasing accuracy in the C4.5 algorithm. The original attribute of the chronic kidney disease dataset consisted of 24 attributes and 1 class, after the information gain ratio method was applied as selecting attributes 12 attributes were selected. The default iteration in adaboost is 50 iterations. The accuracy of stand-alone C4.5 was 96.66%, for C4.5 with information gain ratio of 97.5%, while C4.5 method was based on information gain ratio and adaboost was 98.33%. So, it can be concluded that combining information gain ratio and adaboost methods can improve classification accuracy.

REFERENCES

[1] Boukenze, B., Mousannif, H., & Haqiq, A. (2016). Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease. International Journal of Database Management Systems (IJDMS), 8(3).

[2] Shajahaan, S. S., Shanthi, S., & ManoChitra, V. (2013). Application Data mining Techniques to Model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering, 3(11): 362-369.

[3] Pranatha, A. A. (2012). Analisis Perbandingan Lima Metode Klasifikasi pada Dataset Sensus Penduduk. Jurnal Sistem Informasi, 4(2): 127-134.

[4] Neeraj, B., Girja, S., Ritu, D. B., & Manisha, M. (2013). Decision Tree Analysis on J48 Algorithm for Data mining. International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE), 3(6): 1114-1119.

[5] Muzakir, A., & Wulandari, R. A. (2016). Model Data mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. Scientific Journal of Informatics, 3(1): 19-26.

[6] Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018). Optimization of C4.5 Algorithm-Based Particle Swarm Optimization for Breast Cancer Diagnosis. International Conference on Mathematics, Science and Education, 983(1): 012-063.

[7] Padmanaban, K. A & Parthiban, G. (2016). Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease. Indian Journal of Science and Technology, 4(2): 1-5.

[8] S, T., Bai, M., & Majumdar, J. (2017). Analysis and Prediction of Chronic Kidney Disease Using Data Mining Techniques. International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 4(9): 25-32.

[9] Gola, J., Britz, D., Staudt, T., Winter, M., Schneider, A. S., Ludovici, M., & Mucklich, F. (2018). Advanced microstructure classification by Data mining methods. Computational Materials Science, 148: 324-335.

[10] Nurzahputra, A., Safitri, A. R., & Muslim, M. A. (2017). Klasifikasi Pelanggan pada Customer Churn Prediction Menggunakan Decision Tree. Prosiding Seminar Nasional Matematika. Semarang: Universitas Negeri Semarang: 717- 722.

[11] Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M., & Mendes, M. P. (2018). Feature Selection Approaches for Predictive Modelling of Groundwater Nitrate Pollution: An Evaluation of Filters, Embedded and Wrapper Methods. Science of the Total Environment, 624(2018): 661-672.

[12] Prasetyo, E. (2014). Data mining: Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: Andi Offset.

[13] Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining. Yogyakarta: CV Andi Offset.

[14] Neeraj, B., Girja, S., Ritu, D. B., & Manisha, M. (2013). Decision Tree Analysis on J48 Algorithm for Data mining. International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE), 3(6): 1114-1119.

[15] Quinland, J. Ross. (1986). Introduction of Decision Tree. Machine Learning. 1(1): 81-106

[16] Han, J. (2012). Data mining Concepts and Techniques. San francisco: Morgan Kauffman.

[17] Listiana, E., & Muslim, M. A. (2017). Penerapan Adaboost Untuk Klasifikasi Support Vector Machine Guna Mengingkatkan Akurasi Pada Diagnosa Chronic Kidney Disease. Prosiding Seminar Nasional Teknologi dan Informatika, 875- 881.

[18] Nurzahputra, A., & Muslim, M. A. (2017). Peningkatan Akurasi pada Algoritma C4.5 Menggunakan Adaboost untuk Meminimalkan Resiko Kredit. Prosiding Seminar Nasional Teknologi dan Informatika. Kudus: Universitas Muria Kudus: 243-247

JOSCEX
Journal of Soft Computing Exploration

Data Security System of Text Messaging Based on Android Mobile Devices Using Advanced Encrytion Standard Dynamic S-BOX

Akhmad Sahal Mabruri [1], Alamsyah[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jul 29, 2020
Revised Aug 12, 2020
Accepted Sept 3, 2020

Keywords:

Android
Chat Messaging
Cryptography
Advanced Encrytion Standard
Dynamic S-BOX

ABSTRACT

Most of the recent technologies are turning to mobile platforms, Android becames one of the most widely used OS. Eventhough it has complete features, even it's not safe enough such like Chat Messenger. The security of messages distribution is a challenge to increase of vulnerable distribution of information through the network today. Therefore, a data security or cryptographic algorithm is needed to secure the messages so that it cannot be read by irresponsible people. National Institute of Standard and Technology (NIST) established the Advanced Encrytion Standard (AES**)** cryptographic algorithm as a standard encryption algorithm that is safe and can be used globally. AES algorithm is included in block cipher cryptography that uses substitution boxes (S-BOX) in its operations, so that algorithmically can make input and output unrelated. So, it can provide more varied output in the process, we need a dynamic S-BOX. In this research, dynamic S-BOX generalized using XOR operations from affine transformations with 8-bit binary element matrices arranged and randomly to produce as many as 256 S-Boxes. The application of dynamic AES with S-BOX algorithm on Android-based messenger chat application is built using the Java programming language and database hierarchy for data storage. The implementation results showed that the algorithm was running well and could encrypt the text of the message to ciphertext and decrypt the ciphertext to the original message. This research can be used as a reference so that further researchers can merge the AES algorithm with other algorithms to improve the security of encryption in text files, documents, images, videos or other types of files.

Corresponding Author:

Akhmad Sahal Mabruri
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: mabroery7@gmail.com

1. INTRODUCTION

The security of information distribution is an important aspect. The delivery of information requires a process that can ensure the security and data integrity [1]. One of study that can be implemented is cryptography. Cryptography came from the Greek language, which is from the terms of crypto which means secret and graphia which means writting. Cryptography is a science or art used to maintain the security of messages when messages were sent from one place to another. The term well known as encryption and decryption [2]. In cryptography, there is an algorithm that became encryption standards set by the National Institute of Standard and Technology (NIST) in October 2000, namely the Advanced Encryption Standard (AES). AES is included in the type of cryptographic algorithms which symmetrical and cipher block characteristics [3]. This algorithm uses the same key when encryption and decryption and the input and output in the form of blocks with a certain number of bits. The bit size in AES can be 128-bit, 192-bit or 256-bit, named AES-128, AES-192, and AES-256 [4]. The selection of the size of the data block and key will determine the number of processes that must be passed for the encryption and decryption process. However, de-facto there are only two variants of AES, namely AES-128 and AES-256, because users will rarely use 192-bit long keys [5]. Because AES has at least 128-bit key lengths, AES is resistant to exhaustive key search attacks with current technology. Using a 128- bit key length, there are $2^{128} = 3{,}4 \times 10^{38}$ possible keys.

AES is included in block cipher encryption, like other block cipher encryption, AES also uses substitution boxes to replace inputs and outputs to be unrelated. AES has four transformation data for Substitution bytes (SubBytes) / Inverse SubBytes (Inv SubBytes), ShiftRow / Inverse ShiftRow (Inv ShiftRow), MixColumn / Inverse MixColumn (Inv MixColumn) and AddRoundkey. [6]. But unlike other block cipher algorithms that use dynamic S-BOX, AES specifies a static S-BOX that is used in the encryption and decryption process. S-BOX has an important role in the implementation of AES. S-BOX is used to randomize input bits that will produce output bits [7].

This study discussed about how to design and implement the dynamic S-BOX Advance Encryption Standard (AES) method for the security of text message data that will be sent and received by the user. So, it will be designed a chat or messaging application that implements AES cryptographic algorithm using dynamic S-BOX. The purpose of this study is securing data and generating text message encryption applications on chat messengers by using dynamic S-Box AES algorithm method.

2. METHOD

This research was built by developing AES method and Dynamic S-Box into the systems. The system was developed by waterwall model which consist five stages. The waterfall model can be seen in Figure 1. Communication stage in the waterfall model of a software or system to be built requires analysis of software requirements in the form of collecting additional data in journals, articles, books, etc. Furthermore, in the planning stage, namely the plan for implementing a generated dynamic S-Box on the AES-128 bit cryptographic algorithm. So, it produces a user requirement document or in other words data that relates to the user's desire in developing the software. The next stage that must be passed is modeling, this process will translate the requirements to a software design. This process focuses on data structure design, software architecture and algorithmic / procedural details. After modeling, the next stage is construction or developing the system using the coding process. After the coding is complete, testing of the system will be done. The goal is that when testing can be found errors in the system can be corrected. The last stage is deployment, in this process making the system is in the final stage. Then the system that has been made must be regularly maintained.

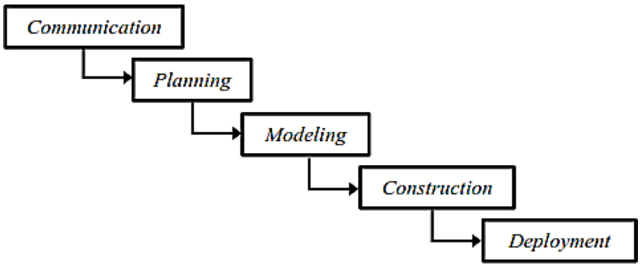


Figure 1. The waterfall model

2.1 Advanced Encrytion Standard

Advanced Encrytion Standard (AES) is one of cryptography algorithm which symmetry and block cipher. This algorithm uses the same key to encrypt and decrypt as well as input and output in the form of blocks with a certain number of bits [7]. AES supports a variety of block sizes and keys that will be used. However, Rijndael has a fixed block and key size of 128, 192, 256 bits. The selection of block size and key data will determine the number of processes that must be passed for the encryption and decryption process. This encryption technique includes the type of block cipher as well as DES. The main difference between the AES encryption technique and the DES encryption technique is that AES uses substitutions or often called S-boxes

2.2 Dynamic S-BOX

S-Box is a matrix that contains simple substitutions which mapped one or more bits with one or more other bits. It’s also well known as most cipher block algorithms. S- Box mappped m input bits into n output bits, so that the S-Box is called m × n S-Box [8]. Substitution process that mappped input based on look-up tables. Usually the input from the operation on the S-Box is used as the index and the output is the entry.

2.3 Android

Android is an operating system for mobile phones based on Linux. Android is a computer code-based software that can be distributed openly or well know as open source so that programmers can create new applications in it [9]. There is an Android Market that provides thousands of applications which free and paid, and has a native Google application that is integrated, such as push email GMail, Google Maps, and Google Calendar. The android application was developed with the Java programming language, using the Android software development kit (SDK). This SDK consists of a set of tools for development, including a debugger and software library, a QEMU-based handset emulator, documentation, sample code and tutorials.

2.4 Firebase Real Time Database

Firebase Realtime Database is a NoSQL cloud-hosted database service which owned by Firebase SDK. This service offers data storage that can be synchronized in real time to all connected clients and can support flick mode for situations when an internet connection is not available [10]. Firebase in the backend has several components, namely Cloud messaging, Authentication, Realtime, Storage, and Hosting. Cloud messaging can be used to send messages between users. First, authentication simplifies application integration by limiting user access. Second, firebase can be used in realtime or offline. Third, the storage also allows to keep image, audio and video content. The last component is that hosting which is possible to be developed globally [11].

2.5 Chat Messenger

Chat messenger is one of technology that provides a feature of communication between two or more people using a network that allows users to send and receive messages in real time [12]. In addition, chat messenger is also called an instant messaging application or instant message on a computer network technology that allows users to send messages to other users who are connected on a computer or internet network. Communication in chat is generally in the form of text (text chat).

3. RESULT AND DISCUSSION

In this research the implementation of AES algorithm in chat applications developed using application namely Android Studio. The development used the waterfall method which consists of Analysis of Requirement, Design, Implementation and Testing. Analysis of Requirement was conducting by preparing all the needs in the developing of application. The design stage was conducted by making the application interface and databases. The implementation stage was conducted by coding and applying AES using dynamic S-BOX. The testing stage was carried out using the Black Box method. The stages of 128-bit AES algorithm encryption in securing text messages are as follows:

The encryption stages with AES algorithm using Dynamic S-BOX:

Plaintext:	selamat pagi pak
In Hexa:	73 65 6C 61 6D 61 74 20 70 61 67 69 20 70 61 6B
Key:	kunciku
In Hexa:	6B 75 6E 63 69 6B 75 00 00 00 00 00 00 00 00 00

Step 1: Continue the calculation, prepare two 4x4 matrices from plaintext and key

Plaintext (Hex) : $\begin{bmatrix} 73 & 6D & 70 & 20 \\ 65 & 61 & 61 & 70 \\ 6C & 74 & 67 & 61 \\ 61 & 20 & 69 & 6B \end{bmatrix}$

Key (Hex) : $\begin{bmatrix} 6B & 69 & 00 & 00 \\ 75 & 6B & 00 & 00 \\ 6E & 75 & 00 & 00 \\ 63 & 00 & 00 & 00 \end{bmatrix}$

Step 2: Conduct XOR on plaintext with roundkey. The XOR process between the corresponding columns of the two matrices begins by switching the hexadecimal data of each column into binary form. This step is called addroundkey, which will generate a new matrix.

$$\begin{bmatrix} 73 & 6D & 70 & 20 \\ 65 & 61 & 61 & 70 \\ 6C & 74 & 67 & 61 \\ 61 & 20 & 69 & 6B \end{bmatrix} \text{XOR} \begin{bmatrix} 6B & 69 & 00 & 00 \\ 75 & 6B & 00 & 00 \\ 6E & 75 & 00 & 00 \\ 63 & 00 & 00 & 00 \end{bmatrix}$$

Binner Hex 73 XOR 6B= 0111 0011 XOR 0110 1011= 0001 1000 = 18
Binner Hex 65 XOR 75= 0110 0101 XOR 0111 0101 = 0001 0000 = 10
Binner Hex 6C XOR 6E= 0110 1100 XOR 0110 1110 = 0000 0010 = 02
Binner Hex 61 XOR 63= 0110 0001 XOR 0110 0011 = 0000 0010 = 02
Binner Hex 6D XOR 69= 0110 1101 XOR 0110 1001= 0000 0100 = 04
Binner Hex 61 XOR 6B= 0110 0001 XOR 0110 1011= 0000 1010 = 0A
Binner Hex 74 XOR 75= 0111 0100 XOR 0111 0011 = 0000 0111 = 07
Binner Hex 20 XOR 00= 0010 0000 XOR 0000 0000 = 0010 0000 = 20
Binner Hex 70 XOR 00= 0111 0000 XOR 0000 0000 = 0111 0000 = 70
Binner Hex 61 XOR 00= 0110 0001 XOR 0000 0000 = 0110 0001 = 61
Binner Hex 67 XOR 00= 0110 0111 XOR 0000 0000 = 0110 0111 = 67
Binner Hex 69 XOR 00= 0110 1001 XOR 0000 0000 = 0110 1001 = 69
Binner Hex 20 XOR 00= 0010 0000 XOR 0000 0000 = 0010 0000 = 20
Binner Hex 70 XOR 00= 0111 0000 XOR 0000 0000 = 0111 0000 = 70
Binner Hex 61 XOR 00= 0110 0001 XOR 0000 0000 = 0110 0001 = 61
Binner Hex 6B XOR 00= 0110 1011 XOR 0000 0000 = 0110 1011 = 6B

So as producing a matrix, as follows:

$$\begin{bmatrix} 18 & 04 & 70 & 20 \\ 10 & 0A & 61 & 70 \\ 02 & 01 & 67 & 61 \\ 02 & 20 & 69 & 6B \end{bmatrix}$$

Step 3: After getting the XOR result matrix between plaintext and roundkey, the substitution process is conducted with dynamic S-BOX, which the dynamic S- BOX table on the keyword "kunciku" is shown in Table 1.

Tabel 1. Dynamic S-BOX Table on the Keyword “kunciku”

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**a**	**b**	**c**	**d**	**e**	**f**
0	8	23	28	16	153	0	4	174	91	106	12	64	149	188	192	29
1	161	233	162	22	145	50	44	155	198	191	201	196	247	207	25	171
2	220	150	248	77	93	84	156	167	95	206	142	154	26	179	90	126
3	111	172	72	168	115	253	110	241	108	121	235	137	128	76	217	30
4	98	232	71	113	112	5	49	203	57	80	189	216	66	136	68	239
5	56	186	107	134	75	151	218	48	1	160	213	82	33	39	51	164
6	187	132	193	144	40	38	88	238	46	146	105	20	59	87	244	195
7	58	200	43	228	249	246	83	158	215	221	177	74	123	148	152	185
8	166	103	120	135	52	252	47	124	175	204	21	86	15	54	114	24
9	11	234	36	183	73	65	251	227	45	133	211	127	181	53	96	176
a	139	89	81	97	34	109	79	55	169	184	199	9	250	254	143	18
b	140	163	92	6	230	190	37	194	7	61	159	129	14	17	197	99
c	209	19	78	69	119	205	223	173	131	182	31	116	32	214	224	225
d	27	85	222	13	35	104	157	101	10	94	60	210	237	170	118	245

e	138	147	243	122	2	178	229	255	240	117	236	130	165	62	67	180
f	231	202	226	102	212	141	41	3	42	242	70	100	219	63	208	125

The dynamic S-Box in this research was reshaped based on the modification of the affine transformation. The dynamic S-Box table has the same size as the original AES S-box table, but in that dynamic S-box there are 256 S-Boxes and the work process is constantly changing and randomly. In this research, we would create a generalized S-Box table that uses generalized XOR operations from affine transformations with 8-bit matrix binary elements arranged randomly to produce 256 different matrix shapes.

$$\begin{bmatrix} b'_0 \\ b'_1 \\ b'_2 \\ b'_3 \\ b'_4 \\ b'_5 \\ b'_6 \\ b'_7 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Substitution Results from Dynamic S-BOX, as follows:

$$\begin{bmatrix} C6 & 99 & 3A & DC \\ A1 & 0C & 84 & 3A \\ 1C & 17 & EE & 84 \\ 1C & DC & 92 & 14 \end{bmatrix}$$

Step 4: In the substitution results with the new S-BOX, the shiftrows have been carried out. The shiftrows process is very simple by doing wrapping (cyclic).

$$\begin{bmatrix} C6 & 99 & 3A & DC \\ 0C & 84 & 3A & A1 \\ EE & 84 & 1C & 17 \\ 14 & 1C & DC & 92 \end{bmatrix}$$

Step 5: After the results of the shiftrows was obtained, then mixcoloumn is conducted by multiplying the result matrix shiftrows with the rijndael matrix. This transformation is expressed as matrix multiplication. All plaintexts are carried out the same process so that the new mixcoloumns matrix will be obtained.

$$\begin{bmatrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{bmatrix} \times \begin{bmatrix} C6 \\ 0C \\ EE \\ 14 \end{bmatrix} = \begin{bmatrix} 79 \\ E3 \\ 31 \\ 9B \end{bmatrix}$$

$$\begin{bmatrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{bmatrix} \times \begin{bmatrix} 99 \\ 84 \\ 84 \\ 1C \end{bmatrix} = \begin{bmatrix} 26 \\ 01 \\ 2A \\ 88 \end{bmatrix}$$

$$\begin{bmatrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{bmatrix} \times \begin{bmatrix} 3A \\ 3A \\ 1C \\ DC \end{bmatrix} = \begin{bmatrix} FA \\ B6 \\ 47 \\ CB \end{bmatrix}$$

$$\begin{bmatrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{bmatrix} \times \begin{bmatrix} DC \\ A1 \\ 17 \\ 92 \end{bmatrix} = \begin{bmatrix} DE \\ 2E \\ FE \\ F6 \end{bmatrix}$$

So as to produce the mixcoloumns matrix as follows:

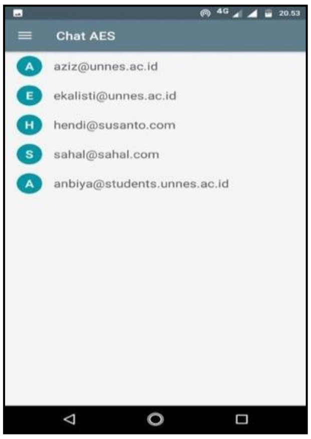
$$\begin{bmatrix} 79 & 26 & FA & DE \\ E3 & 01 & B6 & 2E \\ 31 & 2A & 47 & FE \\ 9B & 88 & CB & F6 \end{bmatrix}$$

Step 6: Conduct XOR between the matrix of mixcoloumns with the key generated from the key expansion process. Repeat the 3rd to 7th steps for the 1st to 9th iterations because this study uses AES-128. The results of the AES algorithm encryption calculation use dynamic S-BOX, which is as follows:

```
addRoundKey(1) 18100202040A0120706167692070616B
 subBytes(1)       C6A11C1C990C17DC3A84EE92DC3A8414
shiftRows(1)       C60CEE149984841C3A3A1CDCDCA11792
mixColumns(1)  79E3319B26012A88FAB647CBDE2EFEF6
                                  -
                                  -
                                  -
addRoundKey(9) A720A27BF7DC51C4C25F76F874C18D1D
 subBytes(9)       37DC514A03EDBA774EA4532AF91336CF
shiftRows(9)       37ED53CF03A4364A4E135177F9DCBA2A
mixColumns(9) DECC36628D4015038FEC667E06A53422
addRoundKey(10)51B95801C15E0E600FF27D1D0EBB2F41
subBytes(10)       BA3D01171333C0BB1DE294CFC0817EE8
 shiftRows(10)     BA3394E813E27E171D8101BBC03DC0CF
RoundKeyAkhir 0346FA8BE6890B1768F46FD8BD56B5CF
Output              0346FA8BE6890B1768F46FD8BD56B5CF
```

The algorithm is run every time you send or open a conversation to be able to read the message. Then, the figure of application itself can be seen in Figure 2 which is the list of conversations or chat rooms on the application.

When choosing a chat room, the user is required to fill the encryption key on the application as a key to read the existing message and as a key to encrypt the message to be sent. For interfaces when key input can be seen in Figure 3.



Chat AES

A aziz@unnes.ac.id

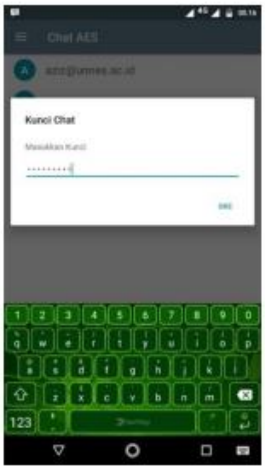
E ekalisti@unnes.ac.id

H hendi@susanto.com

S sahal@sahal.com

A anbiya@students.unnes.ac.id

Figure 2. *Chat Room Interface*



Chat AES

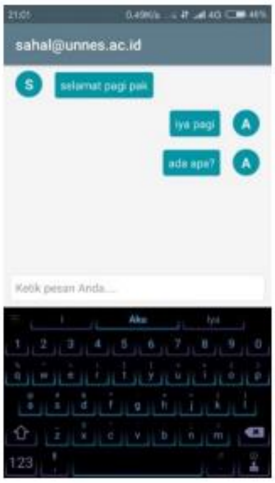
aziz@unnes.ac.id

Kunci Chat

Masukkan Kunci

Figure 3. *Input Encrypt Key*

In the chat interface, the message text can be seen if the key entered was the same as the key used when sending the message, and if the key entered was different then the message can not be read. The interface of the chat room interface that was successfully can read seen in Figure 4, and the interface of the chat room opened with a different key can be seen in Figure 5.


sahal@unnes.ac.id

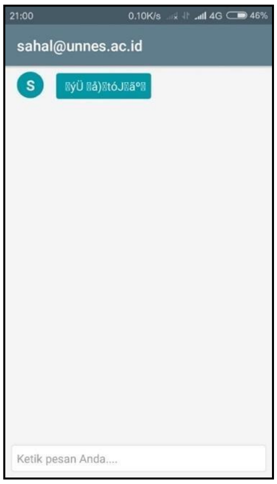
S: selamat pagi pak

iya pagi :A

ada apa? :A

Ketik pesan Anda....

Figure 4. Chat Room Interface with Correctly Key


sahal@unnes.ac.id

S: ▯ýÜ ▯å)▯tóJ▯ã°▯

Ketik pesan Anda....

Figure 5. Chat Room Interface with Incorrectly Key

4. CONCLUSION

The developing of this application was made using the Android Studio framework with the java programming language and using hierarchical databases. There were 4 stages that must be conducted, namely, analysis of requirement to determine the needs of users and systems to the application, the design stage to design the appearance and process of the application, system design that has been done before, the implementation stage to build an application based on the results of the analysis and testing conducted black box method. The dynamic S-BOX of AES chat application could run properly. In the form of text messages that encrypt the plain text into ciphertext. AES dynamic S-BOX chat application had also managed to decrypt the changed text (ciphertext) as before by using the same key when opening the received message.

REFERENCES

[1] M.A. Muslim, B. Prasetiyo, and Alamsyah, "Implementation Twofish Algorithm for Data Security in A Communication Network Using Library Chilkat Encryption Activex, " *Journal of Theoretical & Applied Information Technology*, vol. 84, no. 3, pp. 370-375, 2016.

[2] R. Venkateswaran, and V. Sundaram, "Information Security: Text Encryption and Decryption with Poly Substitution Method and Combining the Features of Cryptography, " *International Journal of Computer Applications*, vol. 3, no. (7), pp. 28-31, 2010

[3] C. Sanchez-Avila, and R. Sanchez-Reillol, The Rijndael Block Cipher (AES Proposal): A Comparison With DES. Proceedings of 35th *International Carnahan Conference IEEE.* London, England, Okotber 16, 2001.

[4] M. Rao, T. Newe, and I. Grout, AES Implementation on Xilinx FPGAs Suitable for FPGA Based WBSNs. Proceedings of 9th International Conference Sensing Technology (ICST) IEEE. Massey, New Zealand, December 8 2015.

[5] O. Dunkelman, N. Keller, and A. Shamir, "Improved Single-Key Attacks on 8-Round AES-192 and AES-256., " *Journal of Cryptology*, vol. 28, no. 3, pp. 397-422, 2015

[6] H. Hamzah, N. Ahmad, M.H. Jabbar and C.F. Soon, "AES S-Box/Inv S- Box Optimization Using FPGA Implementation, " *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* vol. 9, no. 3, pp. 133-136. 2017

[7] Alamysah, A. Bejo, A. and Adji, T.B. (2017). *AES S-Box Construction Using Different Irreducible Polynomial and Constant 8-bit Vector*. Proceedings of *IEEE Conference on Dependable and Secure Computing*. Taipe, Taiwan, August, 7, 2017.

[8] U. Çavuşoğlu, S. Kaçar , I. Pehlivan, and A. Zengin, "Secure image encryption algorithm design using a novel chaos based S-Box."*Chaos, Solitons & Fractals*, vol. 95, pp. 92-101. 2017.

[9] B. Shebaro, O. Oluwatimi, and E. Bertino, "Context-based access control systems for mobile devices, " *IEEE Transactions on Dependable and Secure Computing.* vol. 12, no. 2, pp.150-163, 2015.

[10] Alepis, E. and Nita, S. (2017). Mobile Application Providing Accessible Routes for People with Mobility Impairments. Proceeding of *8th International Conference Information, Intelligence, Systems & Applications (IISA) IEEE*. Lanarca, Cyprus, August, 27, 2017.

[11] M.A. Mohammed, A.S. Bright, C. Apostolic, F.D. Ashigbe, and C. Somuah, “Mobile-Based Medical Health Application-Medi-Chat App, ” *International Journal of Scientific & Technology Research*, vol. 4, no. 8, pp. 70-76, 2015.

[12] R. Sanjaya, and A. S. Girsang, Implementation Application Internal Chat Messenger Using Android System. Proceeding of International Conference in Applied Computer and Communication Technologies (ComCom*)*, Jakarta, Indonesia, May, 17, 2017.

JOSCEX
Journal of Soft Computing Exploration

Improve the Accuracy of C4.5 Algorithm Using Particle Swarm Optimization (PSO) Feature Selection and Bagging Technique in Breast Cancer Diagnosis

Raka Hendra Saputra[1], Budi Prasetyo[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jul 25, 2020
Revised Aug 10, 2020
Accepted Sept 3, 2020

Keywords:

Data mining
Decision Tree
Classification
C4.5
PSO
Bagging
Breast Cancer

ABSTRACT

Breast cancer is the second leading cause of death due to cancer in women currently. It has become the most common cancer in recent years. In early detection of cancer, data mining can be used to diagnose breast cancer. Data mining consists of several research models, one of which is classification. The most commonly used method in classification is the decision tree. C4.5 is an algorithm in the decision tree that is often used in the classification process. In this study, the data used was the Breast Cancer Wisconsin (Original) Data Set (1992) obtained from the UCI Machine Learning Repository. The purpose of this study was to select features that will be used and overcome class imbalances that occur, so that the performance of the C4.5 algorithm worked more optimal in the classification process. The methods used as feature selection are PSO and bagging to overcome class imbalances. Classification was tested using the confusion matrix to determine the accuracy that was generated. From the results of this study, the application of PSO as a feature selection and bagging to overcome class imbalances with the C4.5 algorithm succeeded in increasing accuracy by 5.11% with an initial accuracy of 93.43% to 98.54%.

Corresponding Author:

Raka Hendra Saputra
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: rakahendrasaputra@students.unnes.ac.id

1. INTRODUCTION

Breast cancer is the second leading cause of death due to cancer in women currently. It has become the most common cancer among women in developed and developing countries in recent years [1]. Identification of breast cancer can be done manually, but this process is difficult because we must remember all the information needed for each particular situation that cause in low accuracy. Mortality from breast cancer can be reduced if it can be detected early. There are conventional methods for breast cancer detection but machine learning classifiers need to be done because they can get higher accuracy [2].

Data mining is a pattern recognition technology as well as statistical and mathematical techniques to find meaningful correlations, patterns and new trends by sorting out the data storage stacks that store large data [3].

In the medical field, data mining can be used to diagnose some diseases such as breast cancer, heart disease, diabetes, etc. [4].

Classifications in data mining are two forms of data analysis process used to extract models that describe data classes or predict future data trends. In the classification process, there are 2 phases; the first phase is training data, wherein this phase the data are studied and analyzed using classification algorithms. The model or classifier studied is presented in the form of a pattern or classification rule; the second phase is the use of models for classification, and testing data is used to estimate the accuracy generated based on classification rules [5].

The problem that often occurs is the classification has a large number of features in the dataset, but not all of them will be used. Irrelevant and redundant features can reduce performance [6]. Unnecessary features can make generalizations more difficult and increase the size of the search space which makes a major obstacle in machine learning and data mining. To maximize accuracy in classification, we can use feature selection in selecting features that will be used [7].

Feature selection is widely used to overcome irrelevant exaggerated features. Feature selection simplifies a collection of data by reducing dimensions and identifies the relevant features without reducing the prediction of accuracy [8]. Particle Swarm Optimization (PSO) is a metaheuristic optimization for feature selection because it has been proven to be competitive compared to genetic algorithms in some cases, especially in the field of optimization [4]. Metaheuristic optimization has proven to be a superior methodology for getting a good solution in a reasonable time [9]. In addition, too many available features, the dataset also often occurs data imbalances.

Data imbalance is one of the classic problems in classification in machine learning. Data imbalance has been proven to reduce the performance of machine learning algorithms [10]. Imbalance can be interpreted, for example one class (majority class) is more than the other class (minority class) [11].

Breast Cancer Wisconsin (Original) Data Set has 2 classes, namely benign written 2 in class as much as 458 (65.5%) and malignant written 4 in class as many as 241 (34.5%).

Two popular methods used in the ensemble method are Bagging and Boosting [13]. Bagging technique is superior compared to boosting when dealing with data that contains noise [14]. In addition, bagging technique is not only easy to be developed, but also strong when dealing with class imbalances if implemented correctly [15]. Bagging technique can be applied to tree-based methods to increase the value of accuracy that will be generated later [16].

A text can consist of only one word or sentence structure [2]. Information in the form of text is important information and is widely obtained from various sources such as books, newspapers, websites, or e-mail messages. Retrieval of information from text (text mining), among others, can include text or document categorization, sentiment analysis, search for more specific topics (search engines), and spam filtering [3]. Text mining is one of the techniques that can be used to do classification where, text mining is a variation of data mining that tries to find interesting patterns from a large collection of textual data [4].

The classification method itself many researchers use the Naïve Bayes Classifier where a text will be classified in machine learning based on probability [5]. Naïve Bayes Classifier is a pre-processing technology in the classification of features, which adds scalability, accuracy and efficiency which is certainly very much in the process of classifying a text. As a classification tool, Naïve Bayes Classifier is considered efficient and simple, and sensitive to feature selection [6].

The data used in this study contains hotel reviews in English so that it can be seen that the grammar used by a person is very diverse in writing the review, diversity makes the features generated through N-Gram will be very much. Therefore, here we will use N-Gram word characters with N = 1, 2, 3 to retrieve features in a review which will then be classified with the Naïve Bayes Classifier Algorithm.

It is expected that the N-Gram Naïve Bayes Classifier Algorithm in this study can be classified correctly and appropriately. So that the main purpose of this study can be fulfilled which is to know the effect of N-Gram features on Naïve Bayes Classifier for sentiment analysis of hotel reviews.

2. METHOD

Stages of data processing consist of several stages, starting from converting the dataset format which was originally .data to .csv, overcoming missing values contained in the dataset, selecting features with PSO, overcoming class imbalance by bagging, and evaluating using confusion matrix. For more details about the methods used in this study can be seen in Figure 1.

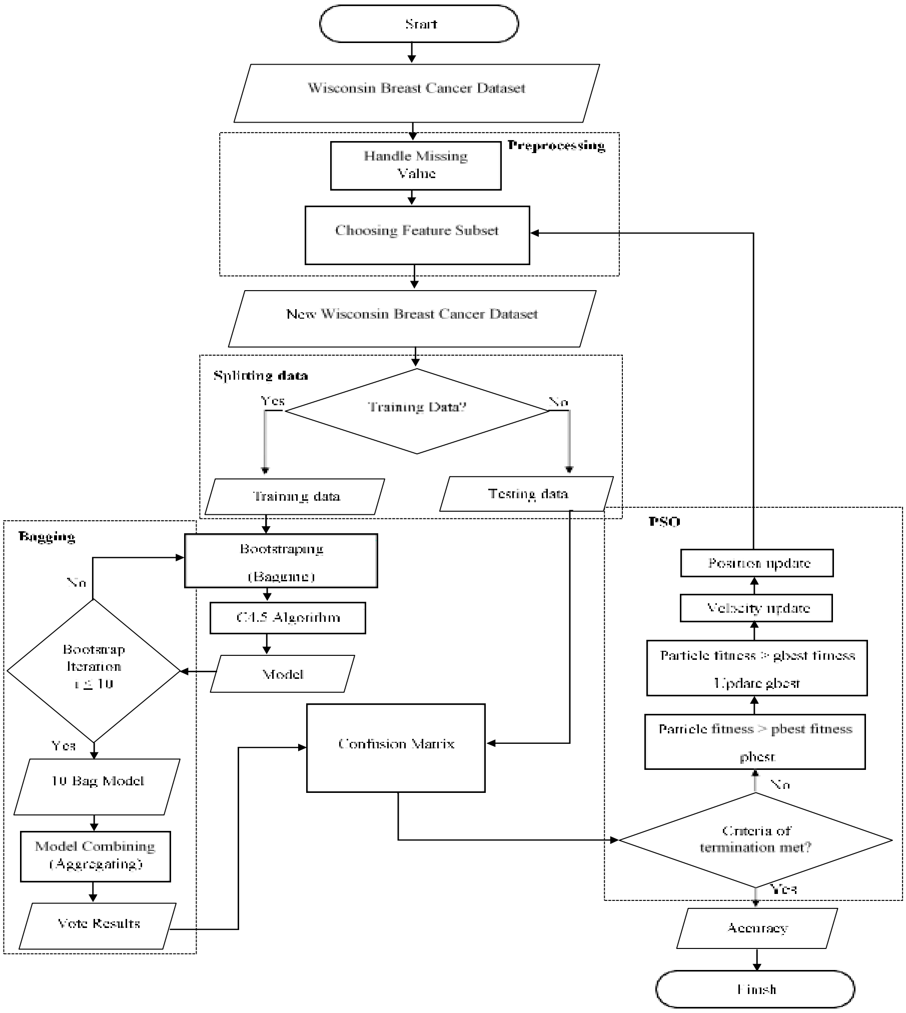
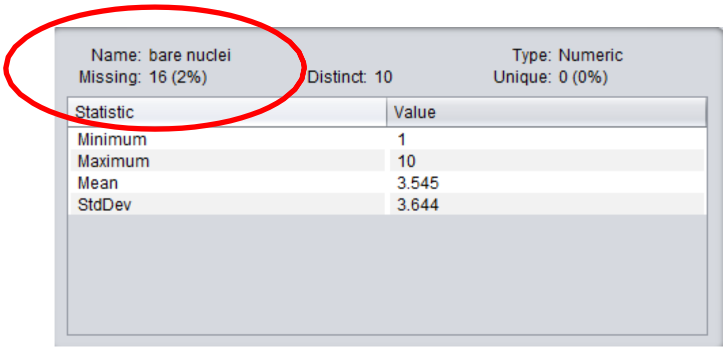


Figure 1. C4.5 algorithm using pso and bagging technique

2.1 Handling Missing Values

The dataset used in this study experienced a missing value of 16 data. This was known with the help of WEKA tools as shown in Figure 2.



Name: bare nuclei
Missing: 16 (2%) Distinct: 10
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	10
Mean	3.545
StdDev	3.644

Figure 2 Missing Value on the Dataset

The missing value was in the bare nuclei attribute which meant as much as 16 data in the bare nuclei attribute was not filled. In this study, 16 data that experienced missing values were overcome by imputation methods, so as not to interfere with the classification process. So that, the amount of data in the dataset was reduced, which was originally 699 data to 683 data.

2.2 Particle Swarm Optimization (PSO)

PSO was selected as the best features available in the Breast Cancer Wisconsin (Original) Data Set. The best features were the features selected to be used in the next process. The PSO steps in selecting the best features are shown in Figure 3.

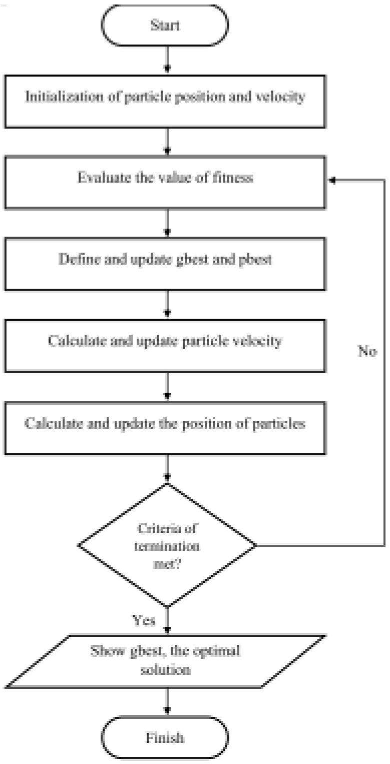

Start
Initialization of particle position and velocity
Evaluate the value of fitness
Define and update gbest and pbest
Calculate and update particle velocity
Calculate and update the position of particles
Criteria of termination met?
No
Yes
Show gbest, the optimal solution
Finish

Figure 3. Flowchart PSO

Based on Figure 3, the PSO steps can be seen more clearly as follows.

Step 1: Initialization of particle position (*xt*), weight of inertia (*w*) = 0.72, and acceleration coefficients (c_1 and c_2) = 0.7. Initialization of particle velocity (*vt*) = 0. Number of particles = 50 and iterations performed = 100.

Step 2: Calculate and evaluate the fitness value of each particle using the C4.5 algorithm.

Step 3: Determine the *pbest* value of each particle based on the accuracy value produced by C4.5. Determine *gbest* value based on the highest *pbest* value.

Step 4: Calculate particle of velocity and position using Equations 1 and 2.

$$v_{id}^{t+1} = w \times v_{id}^{r} + r_{1i} \times (p_{id} - x_{id}^{t}) + c_2 \times r_{2i} \times (p_{gd} - x_{id}^{t}) \quad (1)$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (2)$$

Step 5: Determine the optimal criteria. Determination of particle probability 0 or 1 based on the speed value using the sigmoid function in Equation 3.

$$x(t+1) = \begin{cases} 1 & if \quad rand < s(v(t+1)) \\ 0 & if \qquad\quad not \end{cases} \quad (3)$$

The value rand () is a random number that is uniformly distributed between 0 and 1. The *S*( ) function is a sigmoid function calculated using Equation 4.

$$s(v_{ij}(t+1)) = \frac{1}{1+e^{-v_{id}^{t+1}}} \tag{4}$$

Step 6: Displays *gbest* and optimal solution in the form of selected features that will be used.

2.3 Bagging technique

Bagging is a method that combines bootstrapping and aggregating. Bootstrap samples are obtained by changing the number of elements or resampling the same number of elements as the original dataset [21]. The bagging process was done in training data, while the steps are as in Figure 4.

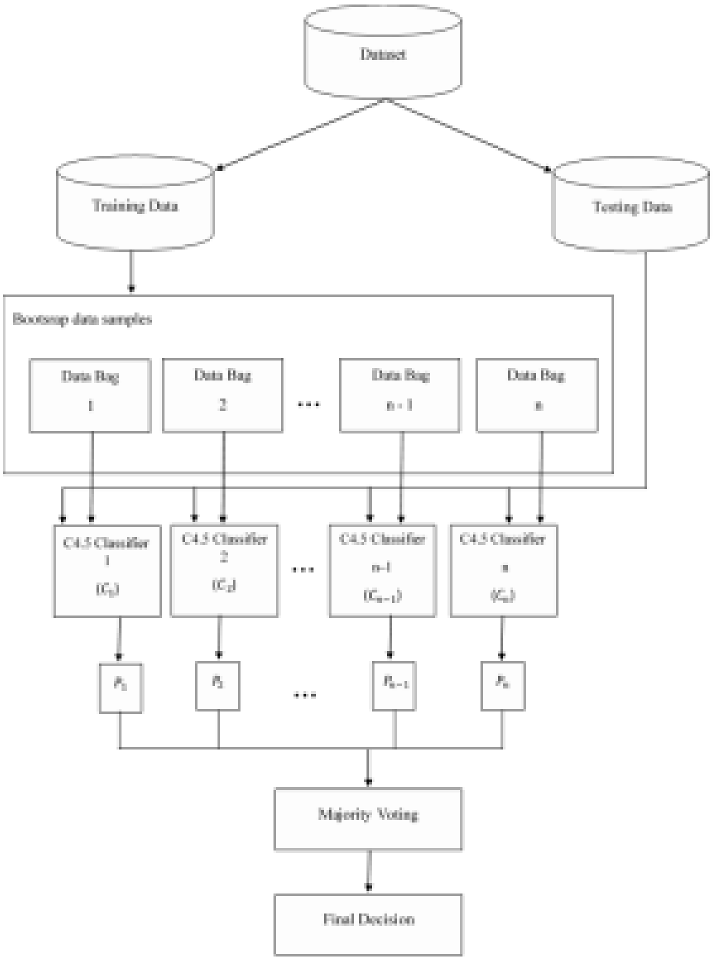


Figure 4. The concept of bagging process

Based on Figure 4, it can be seen more clearly bagging steps are as follows.

Step 1: Perform the bootstrap process on the training data, dividing the data according to the specified number of bags. In this study 100 bags were used.
Step 2: Classify each bag using the C4.5 classifier to get the model.
Step 3: Next, the model obtained was tested using testing data.
Step 4: Each bag produces accuracy, then vote on all results.
Step 5: The results of the final decision were the results based on majority voting.

2.4 C4.5 algorithm

C4.5 algorithm is an algorithm that is widely used in classifications to make decisions because it can produce decision trees that are easy to interpret and understand, it has an acceptable level of accuracy, and are efficient for dealing with discrete and numerical attributes [22]. Stage C4.5 was conducted on the training data in conducting the classification process can be seen in Figure 5.

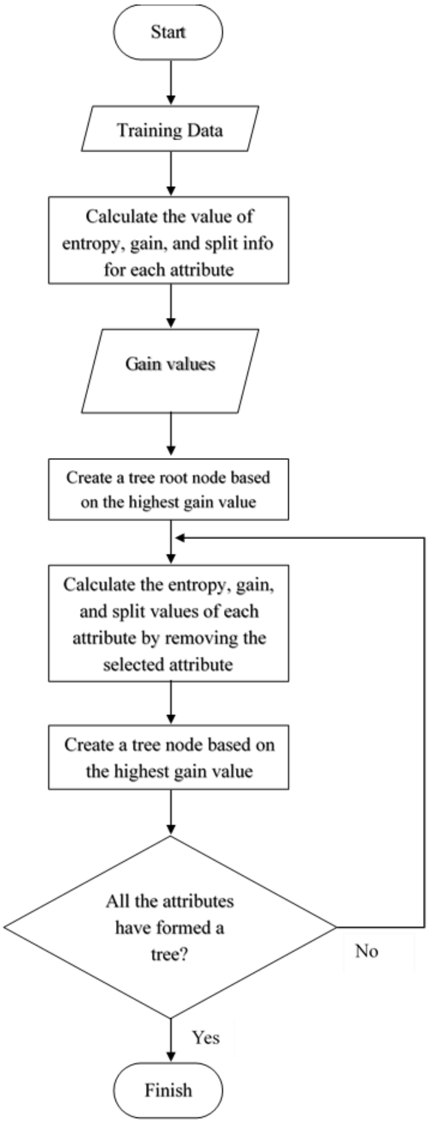

Start
Training Data
Calculate the value of entropy, gain, and split info for each attribute
Gain values
Create a tree root node based on the highest gain value
Calculate the entropy, gain, and split values of each attribute by removing the selected attribute
Create a tree node based on the highest gain value
All the attributes have formed a tree?
No
Yes
Finish

Figure 5. C4.5 Algorithm

Based on Figure 5, it can be seen more clearly the steps of the C4.5 algorithm are as follows.

Step 1: Calculate the entropy of each attribute with Equation 5.

$$Entropy\ (S) = \sum_{i=1}^{n} p_i \times \log_2 p_i \tag{5}$$

Step 2: Calculate the gain info for each attribute by Equation 6.

$$Info\ Gain\ (S.A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \tag{6}$$

Step 3: Calculate the split info for each attribute using Equation 7.

$$Split\ Info\ (S.A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \tag{7}$$

Step 4: Calculate the gain of each attribute by Equation 8.

$$Gain\ (A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \tag{8}$$

Step 5: Calculate the gain of each attribute by Equation 8.
Step 6: Repeat steps 1 to 4 to determine the branch by removing the selected attributes.
Step 7: Create a branch based on the highest gain value.
Step 8: Continue to repeat the process of determining branches until all attributes form a tree.

2.5 Evaluate with the Confusion Matrix

The evaluation stage was carried out at the end of the research process. This stage is useful for testing the model and calculating the resulting accuracy. In this study the evaluation was carried out with a confusion matrix. The steps are as follows.

Step 1: Enter the test results in the confusion matrix table as seen in Table 1.

Table 1. Testing the confusion matrix

Actual	***Predicted***	
	Positive	***Negative***
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

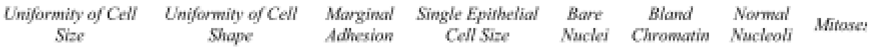
Step 2: Calulate the accuracy value, determine the highest accuracy with Equation 9 .

$$Accuracy = \frac{TP+TN}{P+n} \times 100\% \tag{9}$$

Step 3: State the conclusions from the accuracy results obtained.

3. RESULT AND DISCUSSION

This research was conducted using tools, namely the Python 3 programming language, scikit-learn library, and Pyswarms documentation. While the material used was the Breast Cancer Wisconsin (Original) Data Set obtained from the UCI Machine Learning Repository. The tools and materials in this study were public so they can be accessed and used by anyone who will conduct or prove the validity of the research conducted by previous researchers. PSO feature selection was done to get selected features that will be used for the classification process. The dataset which previously had 9 attributes after being processed by PSO left 8 selected features that will be used in the next process. This can optimize the performance of the C4.5 algorithm in classifying the dataset. The results of the selected features can be seen in Figure 6. A total of 8 selected features when used in the classification process with the C4.5 algorithm produce an accuracy of 95.62%.



Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses

Figure 6. Selected Features by PSO

Bagging was done to overcome the class imbalance that occurs in the dataset used. Bagging was done in training data by dividing the data into 100 bags randomly, the total data of the whole bag was the same as the total training data. Bagging will produce the best bag of 100 bags then the results will be processed by C4.5 algorithm to do the classification. 1 bag with the highest accuracy will be used as the final decision in the classification process. In this study, data from 1 bag selected when processed by the C4.5 algorithm produced an accuracy of 97.81%.

This study recorded every accuracy that results from the classification process that has been done. The results can be seen in Table 2

Table 2. Results of each method used

Algorithm	Accuracy
C4.5	93,43%
C4.5 + PSO	95,62%
C4.5 + Bagging	97,81%
C4.5 + PSO + Bagging	98,54%

Based on Table 2, it was known that there was an increase in each method used. C4.5 algorithm without using PSO and bagging produced an accuracy of 93.43%. C4.5 algorithm with PSO without using bagging produced an accuracy of 95.62%. C4.5 algorithm with bagging without using PSO produced an accuracy of 97.81%. And the purpose method which in this case was an algorithm with PSO and bagging produces an accuracy of 98.54%. So it can be concluded that there was an increase in accuracy of 5.11% when comparing the C4.5 algorithm without PSO and bagging with the purpose method in this study.

When the method used in this study was compared with previous studies, it can be seen that the accuracy produced in this study was 98.54 which shows better than some previous studies using the Breast Cancer Wisconsin (Original) Data Set as in Table 3. Akay [23] in his research showed that the distribution of training and testing data respectively 80% and 20% is the most optimal when used for classification of breast cancer. Lavanya & Rani [24], in her study showed the application of bagging to decision trees which in this case was CART produces an accuracy of 97.85% . Muslim MA et al., [4] in his research succeeded in increasing accuracy by 0.88% by using PSO as a feature selection on the C4.5 algorithm [4]. Shrivas & Singh [25] in his research showed C4.5 using the distribution of training and testing data respectively 80% and 20% for the classification of breast cancer resulting in an accuracy of 92.857% [25].

Table 3. Comparison of research accuracy

Method	Accuracy
Akay	97,91%
Lavanya & Rani	97,85%
Muslim *et al*	96,49%
Shrivas & Singh	92,857%
The Purpose Method	98,54%

4. CONCLUSION

Based on the results of research and discussion related to C4.5 algorithm using Particle Swarm Optimization (PSO) feature selection and bagging technique in breast cancer diagnosis, it can be concluded that PSO was used from a number of features in the dataset. In this case, the feature can be referred to an attribute. The dataset originally had 9 attributes and 1 class became 8 attributes and 1 class after PSO was applied. Bagging was used to overcome class imbalances that occur in the dataset. Bagging produced the best bag to be used in the classification process of the C4.5 algorithm in order to make its performance more optimal. Accuracy results obtained when applied PSO and bagging on the C4.5 algorithm were 98.54%. While, C4.5 without PSO and bagging produced an accuracy of 93.43%. So, it can be seen an increase of 5.11% based on the comparison of the resulting accuracy. This showed that PSO and bagging had an important role in optimizing

REFERENCES

[1] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, “Diagnosis of breast cancer using decision tree data mining technique, ” *International Journal of Computer Applications*, vol. 98, no. 10. 2014.

[2] A. Gupta, and B. N. Kaushik, “Feature selection from biological database for breast cancer prediction and detection using machine learning classifier, ” *J. Artif. Intell*, vo. 11, pp. 55-64, 2018.

[3] D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc. 2004.

[4] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, “Optimization of C4. 5 algorithm-based particle swarm optimization for breast cancer diagnosis, ” *Journal of Physics: Conference Series*, vol. 983, no. 1, 2018.

[5] D. Singh, N. Choudhary, and J. Samota, “Analysis of data mining classification with decision tree technique, ” *Global Journal of Computer Science and Technology*, vol. 13, pp. 1-5, 2013.

[6] B. Xue, M. Zhang, and W. N Browne, “Particle swarm optimization for feature selection in classification: A multi-objective approach, ” *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1656-1671. 2012.

[7] I. A. Gheyas, and L. S. Smith, “Feature subset selection in large dimensionality domains, ” *Pattern recognition*, vol. 43, no. 1, pp. 5-13. 2010.

[8] M. H. Aghdam, S. Heidari, “Feature selection using particle swarm optimization in text categorization,

" *Journal of Artificial Intelligence and Soft Computing Research*, vol. 5, no. 4, pp. 231-238. 2015

[9] S.C. Yusta, "Different metaheuristic strategies to solve the feature selection problem, " *Pattern Recognition*, vol. 30, no. 5, pp. 525-534. 2009.

[10] T. W. Cenggoro, "Deep learning for imbalance data classification using class expert generative adversarial network, " *Procedia Computer Science*, vol. 135, pp. 60- 67. 2018

[11] N. Rout, D. Mishra, and M.K. Mallick, "Handling imbalanced data: a survey, " *International Proceedings on Advances in Soft Computing, Intelligent Systems and Application*s. Singapura, 2018, pp. 431-443.

[12] B. W. Yap, K . A . Rani, H.A.A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah,. *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013).* Singapura: Springer.

[13] D. Opitz and R. Maclin, " Popular ensemble methods: an empirical study, " *Journal of Artificial Intelligence*, vol. 11, pp. 169-198. 1999.

[14] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, A. "Comparing boosting and bagging techniques with noisy and imbalanced data, " *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol .41, no. 3, pp. 552-568, 2010.

[15] W. Feng, W. Huang, W, and J. Ren, "Class imbalance ensemble learning based on the margin theory, " *Applied Sciences*, vol. 8, no. 5, pp. 815. 2018.

[16] C. D. Sutton, "Classification and regression trees, bagging, and boosting." *Handbook of statistics,* vol. 24, pp. 303-329. 2005

[17] M. Bramer, Principles of data mining, London: Springer. 2007

[18] Y. Yang, and W. Chen, "Taiga: performance optimization of the C4. 5 decision tree construction algorithm". *Tsinghua Science and Technology,* vol. 21, no. 4, pp. 415-425, 2016.

[19] B. Boukenze, H. Mousannif, and A. Haqiq, "Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease, " *Int. Journal of Database Managment systems,* vol. 8, no. 30, pp. 1-9, 2016.

[20] K. R. Pradeep, and N. C. Naveen, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics, " *Procedia computer science,* vol. 132, pp. 412-420, 2018.

[21] E. Alfaro, M. Gámez, and N. Garcia, "Adabag: and package for classification with boosting and bagging, ". *Journal of Statistical Software,* vol. 54, no. 2, pp. 1- 35, 2013.

[22] S.J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making, " *Journal of Biomedical Informatics,* vol. 78, pp. 144-155, 2018.

[23] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis, " *Expert systems with applications*, vol. 36, no. 2, pp. 3240-3247. 2009.

[24] D. Lavanya, and K. U. Rani, "Ensemble decision tree classifier for breast cancer data, " *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, pp. 17, 2012.

[25] A.K. Shrivas, and A. Singh, "Classification of breast cancer diseases using data mining techniques, ". *International Journal of Engineering Science Invention*, vol. 5, no. 12, pp. 62-65, 2016.

Journal of Soft Computing Exploration

Improving Algorithm Accuracy K-Nearest Neighbor Using Z-Score Normalization and Particle Swarm Optimization to Predict Customer Churn

Muhammad Ali Imron [1], Budi Prasetiyo [2]

[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 6, 2020
Revised Aug 20, 2020
Accepted Sept 7, 2020

Keywords:

Data Mining
German Credit
K-Nearest Neighbor
Particle Swarm Optimization
Z-Score

ABSTRACT

Due to increased competition in the business world, many companies use data mining techniques to determine the loyalty level of customers. In this business, data mining can be used to determine the loyalty level of customers. Data mining consists of several research models, one of which is classification. One of the most commonly used methods in classification is the K-Nearest Neighbor algorithm. In this study, the data which used are from German Credit Datasets obtained from UCI machine learning repository. The purpose of this study is to find out how Z-Score works to normalize the data and Particle Swarm Optimization to find the most optimal K value parameters, so the performance of the K-Nearest Neighbor algorithm is more optimal during the classification. The methods which were used to normalize the data are Z-score and Particle Swarm Optimization to determine the most optimal K value. The classification was tested using confusion matrix to determine the generated accuracy. From the finding of this study, the application of Z-score normalization and Particle Swarm Optimization with the K Nearest Neighbor algorithm succeed in increasing the accuracy up to 14%. The initial accuracy was 68.5%, and after applying the normalization of Z-Score and Particle Swarm Optimization, the accuracy became 82.5%.

Corresponding Author:

Muhammad Ali Imron
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: aliimron@students.unnes.ac.id

1. INTRODUCTION

Data era is an era where data which rapidly developed, widely distributed and have large capacities require an appropriate and organized processing method so that it can be maximally utilized [1]. From the available data, its' information will be extracted and was expected to be applied into larger data that has never been known before.

In the business world, customers are the main asset. Therefore, various ways have been taken by companies, so customers do not stop to subscribe [2]. To get a new customer costs up to 10 times more than the cost of retaining the existing customers. The high cost to get new customers, of course, companies prefer to retain the existing customers. Based on that fact, many companies turn to retain the existing customers and avoid customer churn [3]. Based on increasing competition and supply in the industrial market, many companies are utilizing data mining to predict [4]. Data mining is an activity used to find interesting patterns for large amounts of data [5].

There are several stages in data mining, those are pre-processing, processing, and post-processing stage. The pre-processing stage consists some stages, such as data cleaning, data integration, data reduction, and data transformation [6]. In the data transformation consists of some methods to process the transformation, such as smoothing, generalization, normalization, aggregation and attribute construction [7]. According to Junaidi, et al. normalization is the process by which a numerical attribute is mapped or scaled within a certain range. Data normalization is useful to minimalize data refraction in data mining because the values of attribute in data usually have different ranges [8].

There are several normalization techniques which often be used, those are min-max normalization, Z-Score normalization and decimal scaling normalization, all of which have the goal of mapping data to a certain scale [9]. Z-Score normalization is data normalization used to provide data ranges using mean and standard deviation [10].

To accomplish optimization problems, Particle Swarm Optimization (PSO) algorithm is one of the meta-heuristic algorithms that commonly used [11]. In some cases, it has been proven that PSO is more competitive [12]. This optimization method is proven effective and succeed to be used to solve multidimensional and multi-parameter optimization problems in machine learning such as neural networks and classification techniques algorithms [13].

K-Nearest Neighbor (KNN) algorithm is a method to classify objects based on learning data which has the closest distance to the object. This technique is very simple and easy to implement. It is similar to clustering technique, which is grouping to the new data based on the distance of the new data to several data / nearest neighbors. Before searching for the distance of the data to the neighbors, we have to determine the value of K neighbor.

2. METHOD

Data processing stage in this study was carried out in several stages, starting from converting the data from .data to .csv extension, the data obtained from UCI machine learning repository had its available numerical data already, thus the transformation of the data from nominal to numeric was not needed. The next stage was normalization of Z-Score and data mining stage. For more details about the methods used in this study, it can be seen in Figure 1.

2.1 Z-score

Certain in the data mining process. Whereas, according to [7]. Normalization is one of the data transformation processes in the data mining process where numerical attributes are scaled in a smaller range. The Z-Score value ranges between infinite negative and positive numbers. Different from the normalized values, the Z-Score does not have a minimum and maximum value set [14]. There are several methods that are usually applied in data normalization, those are: min-max normalization, Z-Score normalization and normalization by decimal scaling. Z-Score normalization is a method of normalizing data when the range of data is not known with any certainties, thus it is necessary to calculate the range using the mean and standard deviation of the data [10].

The application of Z-Score normalization stage in data mining pre-processing process is as follows:
Step 1: Find the average value of each numeric attribute.
Step 2: After found the value of mean or average of each attribute, the next step is look for data variance from the numeric attributes. Variances are used to find out how far the data spread from the mean. Low variance shows that data was clustered very close around the mean and vice versa. In calculating the data variance, each initial value of the data in numeric attribute reduced by the mean value of each attribute. After found the value of the mentioned reduction, the next step is to square the reduction results then add the squared results for each attribute. After found the sum result of each attribute as above, the next step is to divide the aforementioned sum result by (n), for n is the number of the data records and will produce a data variance value for each numeric attribute.

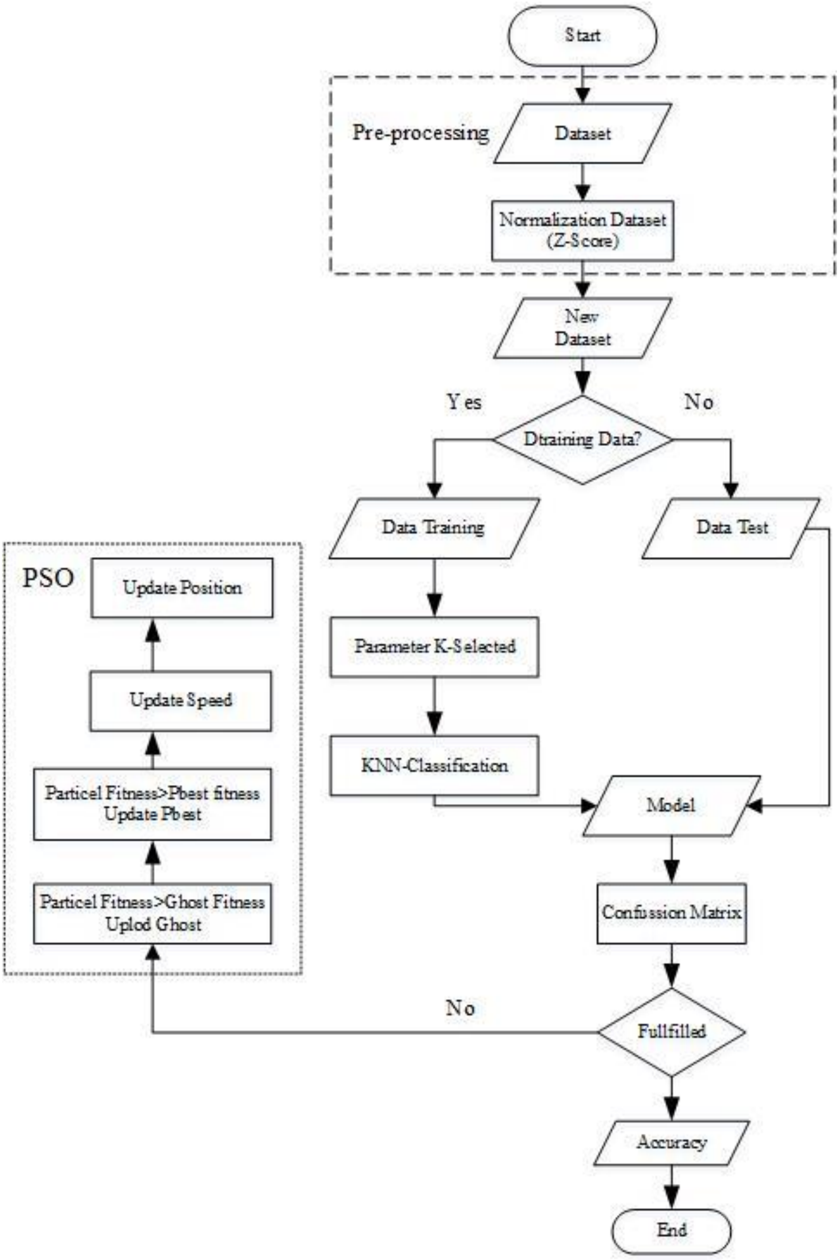


Figure 1. Flowchart of method

Step 3: After found the value of the data variance, the next step is look for the standard deviation value. The value of the standard deviation can be found by calculating the square root of the data variance value of each numeric attribute. To find the standard deviation value, we can use the Equation 1.

$$SD_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{1}$$

Information:

SD_x	= Standard Deviation
n	= Number of Samples
x	= Average
x_i	= Value of x to i

Step 4: The next step is calculate the Z-Score normalization value with Equation 2.

$$Z = \frac{X-\bar{X}}{SD_x} \tag{2}$$

Information:

Z	= Normalization results
x	= Value to be normalized
$\bar{X}$	= Average Value
SD_X	= Standard Deviation

2.2 Partical Swam Optimization

Particle Swarm Optimization (PSO) is one of the basic techniques of the swarm intelligence system to solve optimization problems in the search for space as a solution. This optimization method has been proven effective and has been successfully used to solve multidimensional and multi-parameter optimization problems [15]. The stages of PSO in optimizing can be shown as in Figure 2.

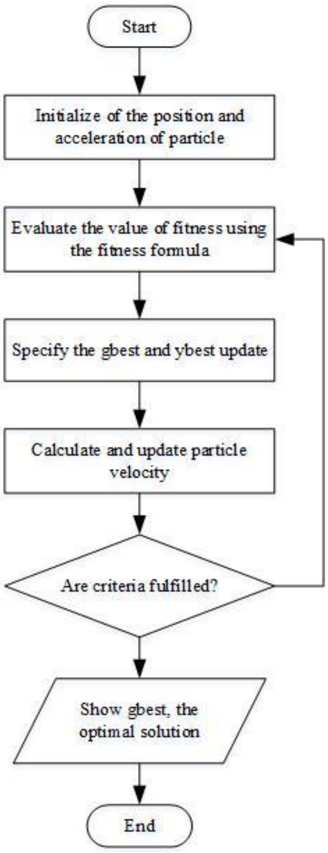

Start
Initialize of the position and acceleration of particle
Evaluate the value of fitness using the fitness formula
Specify the gbest and ybest update
Calculate and update particle velocity
Are criteria fulfilled?
Show gbest, the optimal solution
End

Figure 2 Particle Swarm Optimization

Based on the Figure 2, it can be seen more clearly PSO steps as follows.

Step 1: Initialize particle position (*xt*), weight of inertia (*w*) = 0.72, and acceleration coefficients (*c1* and *c2*) = 0.7. Initialize particle velocity (*vt*) = 0. Number of particles = 50 and iterations performed = 100.

Step 2: Calculate and evaluate the fitness value of each particle by using the KNN algorithm.

Step 3: Determine the *p*best value of each particle based on the accuracy value produced by KNN. Determine *g*best based on the highest *pbest* value.

Step 4: Calculate the velocity and position of the particle by using Eqs. 3 and 4.

$$v_{id}^{t+1} = w \times v_{id}^{t} + c_1 \times r_{1i} \times (p_{id} - x_{id}^{t}) + c_2 \times r_{2i} \times \left(p_{gd} - x_{id}^{t}\right) \tag{3}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{4}$$

Step 5: Show *gbest* and optimal solutions with the best K value which will be used.

2.3 K-Nearest Neighbor Algorithm

K-Nearest Neighbor is a classification algorithm which remains consistent in a large amount of data and classifies based on the closest distance between the data evaluated by the closest point in the training data. The KNN algorithm is more flexible because it is based on the proximity of existing training data [16]. The KNN stage is carried out in the training data during the classification process which can be seen in Figure 3.

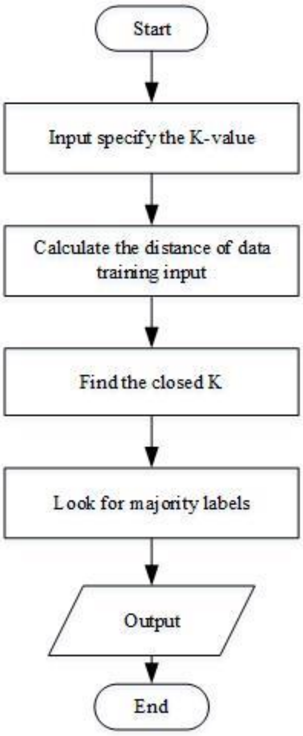

Start
Input specify the K-value
Calculate the distance of data training input
Find the closed K
Look for majority labels
Output
End

Figure 3 . K-Nearest Neighbor

According to the Figure 3, the steps of the KNN algorithm can be seen more clearly.

Step 1: Defining the value of K.
Step 2: Calculating the distance or Euclidean values between the testing data and the training data.
Step 3: Grouping the data based on the distance calculation or Euclidean.
Step 4: Grouping the data based on the distance calculation or Euclidean.
Step 5: Selecting the class that most emerges from the number of the selected K to be used as a prediction result.
The training data on attribute 1 can be seen in Equation 5.

$$X_1 = (X_{11} , X_{12} , . . . ,X_{1\,n}) \qquad (5)$$

The training data on attribute 2 can be seen in Equation 6.

$$X_2 = (X_{21} ,_{22} , . . . ,X_{2\,n}) \qquad (6)$$

Whereas, to find the Euclidean distance expressed by Equation 7.

$$d(X_{1,}X_2) = \sqrt{\textstyle\sum_r^n}(a_r(X_1) - a_r(x_{12}))^2 \qquad (7)$$

2.4 Evaluate using Confusion Matrix

The evaluation stage was carried out at the end of the study process. This stage was useful to test the model and to calculate the accuracy result. In this study, the evaluation was carried out using confusion matrix. The steps are as follows.

Step 1: Enter the test results to the confusion matrix table as in Table 1.

Table 1. Confusion matrix Testing

Actual	**Predicted**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Step 2: Calculate the accuracy value, determine the highest accuracy with Equation 9.

$$Accuracy = \frac{TP+TN}{P+N} \text{ x } 100\% \quad (9)$$

Step 3: State the conclusions from the accuracy results obtaine

3. RESULT AND DISCUSSION

This research was conducted using tools, namely: Sublime Text 3 text editor, Python 3 programming language, skicit-learn library, as well as the documentation pyswarms. While, the material which was used is German Credit Data obtained from the UCI Machine Learning Repository.

Z-Score normalization method can increase accuracy by representing the original data into new data with almost similiar range of values and a narrow range of values. This simplify the data, so that the data mining process can be more optimal and increase accuracy for the KNN algorithm by 80,5%.

The PSO stage was used to optimize the K parameter in KNN algorithm. Each iteration will obtain its' best position with the lowest cost value that called as the best cost. After doing all the iterations, the cost value of each iteration was compared to obtain the final best cost. This final best cost will be used as a recommendation of the optimization process. This recommendation is in the form of the lowest cost value and the chosen K value which can produce the an accuracy of 73,5%.

This study recorded every accuracy results from the classification process that had been done. The results can be seen in Table 2.

Table 2 Results of Each Method Used

Algorithm	Accuracy
KNN	68,5%
KNN+*Z-Score*	80,5%
KNN+PSO	73,5
KNN + *Z-Score* + PSO	82,5%

According to the Table 2, it is seen the improvement of each method used. KNN algorithm without using the normalization of Z-Score and PSO was 68.5%. KNN algorithm with Z-Score normalization without using PSO produces an accuracy of 80.5%. KNN algorithm with PSO without using Z-Score normalization produces an accuracy of 73.5%. Purposed method which in this case is the KNN algorithm using Z-Score and PSO normalization produces accuraty of 82.5%. Thus, it can be concluded that there is an increase in the accuracy of 14% by comparing the KNN algorithm without using the normalization of Z-Score and PSO with the Purposed method in this study.

Table 3 Research Accuracy Results

Method	Accuracy
Sobran *et al*	65,3%
Safitri & Muslim	74,9%
Jeatraku *et al*	77,9%
Purposed Method	82,5%

The method used in this study was compared to the previous studies, it can be seen that the accuracy generated in this study is better than some previous studies using German Credit Data as in the Table 3.

4. CONCLUSION

From the finding and discussion of this study related to the implementation of the normalization of Z-Score and PSO in order to improve the accuracy of KNN algorithm using German Credit Datasets obtained from the UCI Machine Learning Repository, it can be concluded that the application of normalization Z-Score can provide a range of each value in the attribute on German Credit Datasets, thus it is increasing the accuracy of the KNN algorithm. Then, the application of PSO on German Credit Datasets was used to find the best K parameter value, after the optimization results obtained, the best K parameter value was classified using the KNN algorithm. The accuracy results obtained on the application of the KNN algorithm using normalization Z-Score and PSO is 82.5%. The increase of the accuracy is 14% from the application of the KNN algorithm which only has an accuracy of 68.5% for German Credit Datasets objects originating from the UCI repository of machine learning.

REFERENCES

[1] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition. USA: Elsivier. 2012

[2] Y. Liu, and Y. Zhuang, "Research model of churn prediction based on customer segmentation and misclassification cost in the context of big data, " *J. of Comp. & Comm.* vol. 03, pp. 87-93, 2015.

[3] Y. Huang and T. Kechadi, " An effective hybrid learning system for telecommunication churn prediction, " Exp. Sys. with Appl. Vol. 40, pp. 5635-5647, 2013

[4] I. Brandusoiu, and G. Toderean, "Churn Prediction in the Telecommunications Sector using Support Vector Machines, "*Ann. of the Oradea University*, vol. 22, no. 1, pp. 19–22, 2013.

[5] E. Sugiharti, S. Firmansyah, and F. R. Devi, "Predictive Evaluation of Performance of Computer Science Students of Unnes Using Data Mining Based on Naïve Bayes Classifier (NBC) Algorithm, " *J. of Theoretical & Appl. Info. Tech..* , vol. 95, no. 4, pp. 902–911, 2017.

[6] P. Plawiak, M. Abdar, and U. R. Acharya, "Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring, " *App. Soft. Compt*., vol. 84, pp. 105740, 2019.

[7] C. Ordonez, S. Maabout, D. S. Matusevich, and W. Cabrera, "Extending ER models to capture database transformations to build data sets for data mining, " *Data & Know. Eng*. vol. 89, pp. 38-54, 2014.

[8] X. Zhong, and D. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting, " *Neurocomputing***,** vol. 267, pp. 152-168, 2017.

[9] C. Saranya, and Munikandan, "A Study on Normalization Techniques for Privacy Preserving Data Mining, " *Int. J. of Eng. & Tech.*, vol. 5, no. 3, pp. 2701, 2013

[10] H. Goyal, Sandeep, Venu, R. Pokuri, S. Kathula S and N. Battula, " Normalization of Data in Data Mining, " *Int. J. of Soft. & Web Science,* vol. 10, no. 1, pp. 32-33. 2014.

[11] L. A. Ashari, M. A. Muslim, and Alamsyah, "Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing, " *Sci. J. of Info.* vol. 3, no. 2, pp. 149-158, 2016.

[12] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, " Optimization of C4.5 Algorithm-based Particle Swarm Optimization for Breast Cancer Diagnosis " *J. of Physic*, vol. 983, no. 1, pp. 1-5, 2017

[13] S. W. Fei, M. J. Wang, Y. B. Miao, J. Tu, and C. L. Liu, "Particle Swarm Optimization-based Support Vector Machine for Forecasting Dissolved Gases Content in Power Transformer Oil, " *Energy Conversion and Manag.*, vol. 50, no. 6. pp. 1604-1609, 2009.

[14] A. Pandey, and A. Jain, "Comparative Analysis of KNN Algorithm using Various Normalization Techniques " *Int. J. of Comp. Net. & Infor Sec.* vol. 11, no. 04, pp. 36-42, 2017.

[15] Sumathi, S., & Surekha, P. Computational Intelligence Paradigms: Theory and Applications Using Matlab (1st ed.). Boca Raton: CRC Press. 2009

[16] M. R. Hidayah, I. Akhlis, and E. Sugiharti, "Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification , " *Journal of Soft Computing Exploration*, vol. 4, no. 1, pp. 66-74. 2017.

JOSCEX
Journal of Soft Computing Exploration

The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease

Hestu Aji Prihanditya[1], Alamsyah[2]
[1,2]Computer Science Departement, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:

Received Jul 28, 2020
Revised Aug 1, 2020
Accepted Aug 26, 2020

Keywords:

C4.5 Algorithm;
Zscore;
Boosting;
Data Mining
Chronic Kidney Disease

ABSTRACT

In the health sector, data mining can be used as a recommendation to predict a disease from the collection of patient medical record data or health data. One of the techniques can be applied is classification with the C4.5 algorithm. The increasing accuracy can be conducted in data transformation using zscore normalization method. In addition, the implementation of the ensemble method can also improve accuracy of C4.5 algorithm, namely boosting or adaboost. The purpose of this study was determinin the implementation of zscore normalization in the pre-processing and adaboost stages of the C4.5 algorithm and determing the accuracy of the C4.5 algorithm after applying zscore and adaboost normalization in diagnosing chronic kidney disease. In this study, the mining process used k-fold cross validation with the default value k = 10. The implementation of the C4.5 algorithm obtained an accuracy of 96% while the accuracy of the C4.5 algorithm with the zscore normalization method obtained an accuracy of 96.75%. The highest accuracy was obtained from the addition of the boosting method to the C4.5 algorithm and zscore normalization obtained the accuracy of 97.25%. The increasing accuracy was obtained of 1.25% which compared to the accuracy C4.5 algorithm.

Corresponding Author:

N Hestu Aji Prihanditya,
Computer Science Departement, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Semarang, Indonesia
Email: hestuuaji@gmail.com

1. INTRODUCTION

In this technological era, the development of data is growing very rapidly and large. In health sector, it saves a lot of data that can be processed and produce to information or new knowledge. The processing data that can be extracted as information and knowledge from datasets is called data mining. By the existence of data mining, it is expected that it can be able to provide the knowledge as a recommendation as decision making for experts in health sector. Data mining is the process of utilizing several mathematical methods and machines that useful for identifying information from large data [1].

In data mining, there is a pre-processing stage, one of techniques is transformation. Transformation is the process of transforming data so that it can be suitable for data mining [2]. In data transformation stage, there are several methods for conducting the data transformation process such as smoothing, generalization, normalization, aggregation and attribute construction. Normalization is a process where a numeric attribute is mapped or scaled within a certain range in the data mining process [3]. Data normalization is useful to minimize data refraction in data mining because attribute values in data usually have different ranges. There are several normalization Journal of Soft Computing Exploration , Vol. 5, No. 1, May 2018 2 techniques that are often used including normalization of zscore. Zscore normalization is the normalization of data when the range of data is not known with certainty by calculating using the average value and standard data deviation [4].

Data mining has several techniques such as estimation, prediction, classification, clustering, and association. One of the data mining techniques used to predict a decision is classification. Classification is one technique that aims to extract the model into categorical classes [5]. One type of algorithm in data mining classification is the C4.5 algorithm. C4.5 algorithm is an algorithm developed by J. Ross Quinlan from algorithm ID3 that uses the gain ratio as the separation of criteria [6].

In the health sector, data mining can be used as a recommendation as to predict disease from a collection of patient medical record data or health data. Using the classification method, the data such as age, blood pressure, urine concentration and other attributes can be used as supporting factors to make recommendations for predicting the possibility of patients who suffering from chronic kidney disease. Kidney disease is a disease which has not normal kidney function almost as much as 90% and not characterized by certain symptoms [7]. The diagnostic study mostly used a chronic kidney disease dataset obtained from the UCI repository of machine learning datasets. In the classification algorithm, accuracy explained how precisely the algorithm can classify data. Accuracy is very discussedable because if an accuracy has little value or result then it will cause a misinterpretation of classification.

The development of machine learning using the ensemble method can improve accuracy in the way of combining several classifying components. The ensemble method that can be used to improve accuracy on a classifier is bagging and boosting. Boosting is preferred because it has a tendency to increase accuracy higher than bagging. Adaboost is a very popular boosting algorithm to improve classification accuracy. The algorithm can be used in diagnosing a disease, one of which is chronic kidney disease.

So many researchers have conducted research on the C4.5 algorithm specifically using chronic kidney disease datasets from the UCI repository of machine learning datasets and research gaps have been found from these studies. In a research conducted by Sujatha & Ezhilmaran [8], the accuracy of the C4.5 algorithm was 97% for the chronic kidney disease dataset. The preprocessing method was used to replace missing value, then for the data separation using the k-fold cross validation method with a value of k = 2,3,4,5,6. Another research was conducted by Celik et al., [9], in this study obtained in the accuracy result of 96.7% for the C4.5 algorithm.

The purpose of this research was determining the implementation of zscore normalization in the pre-processing stages and adaboost to the C4.5 algorithm and determining the accuracy of the C4.5 algorithm after implementing zscore and adaboost normalization in diagnosing chronic kidney disease.

2. METHOD

2.1 Data Mining

Data mining is a process of exploration of data that has a large number of records and has been taken a certain pattern [10]. The systematically the data mining process has 3 main steps, namely:

2.1.1 Preprocessing

The preprocessing of data consists of cleaning data, data transformation, dimension reduction, selection of feature subset and so on.

2.1.2 Build models and evaluate validity

Building a model and validation means conducting an analysis of the formed model and choosing the model that has the best performance, at this stage of research used the classification method. Classification is a method in data mining that is used to predict class labels in data [11].

2.1.3 Implementation

Implementation means applying a model to new data to form certain knowledge. The data mining process can be seen in Figure 1.

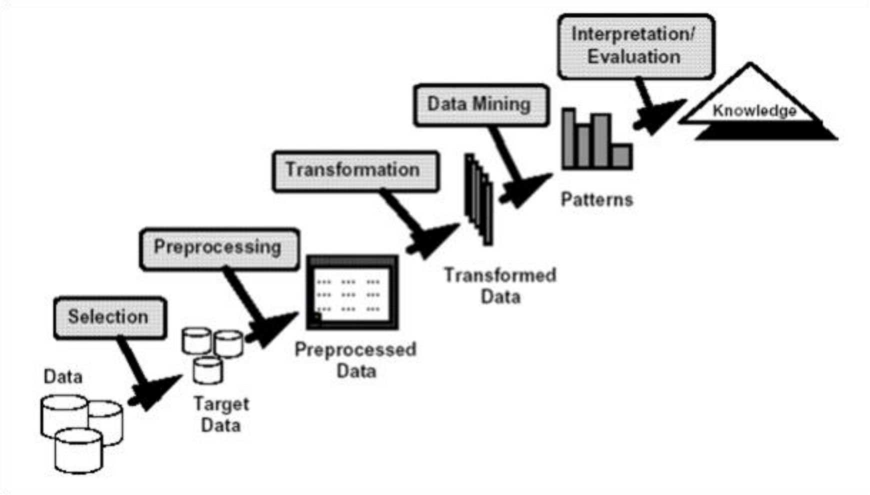


Figure 1. Data Mining Process

2.2 Z-Score Normalization

The normalization is a process in preprocessing stage by decompositing data of numeric attributes which can convert values in data into a certain range [12]. There are several methods that are usually applied in data normalization, including: min- max normalization, z-score normalization and normalization by decimal scaling. Z- score normalization maps a vi value from attribute E to v 'into a range that was previously unknown, can be seen in Equation 1.

$$v' = \frac{v_i - E_i}{std(E)} \tag{1}$$

Description:
v’ = result of normalization value.
v = the value to be normalized in attribute
E_i = the mean value of attribute
std(E) = standard deviation attribute E.

2.3 C4.5 Algorithm

Decision tree is a classification method that converts data into a tree as a rule representation [13]. In the decision tree there is a very famous classification algorithm, namely C4.5 algorithm. Algorithms is a way to solve problems using certain instructions to produce the output [14]. The C4.5 algorithm is an algorithm introduced by Quinlan which is an improvement from the ID3 algorithm. In ID3, the induction decision tree can only be performed on categorical features (nominal/ ordinal), while numeric types (internal / ratio) cannot be used. The C4.5 algorithm is also defined as an algorithm that uses gain ratio as a split attribute selection [15].

The stages form a decision tree using C4.5 algorithm: Prepare the training data from existing data recap and have been grouped in certain classes. Next, determine the root of the tree by calculating the highest gain value for each attribute. For conducting that step, calculate the entropy index first using Equation 2 below

$$Entropy(S) = \sum_{i=1}^{n} - p_i * log_2 p_i \tag{2}$$

Description:
S = Set of Case
n = Number of Partitions S
pi = The proportion of Si to S
Where *log2pi* can be calculated using Equation 3 below.

$$log(X) = \frac{\ln(X)}{\ln(2)} \tag{3}$$

For calculating the gain can use Equation 4 below.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|\ _i|}{|S|} * Entropy(S_i) \tag{4}$$

The criteria for choosing the C4.5 feature is the gain ratio, which can be formulated by the following Equation 5.

$$GainRatio(A) = \frac{\text{Gain}(A)}{\text{SplitEntropy}(A)} \tag{5}$$

For calculating Split Entropy can used Equation 6 as follows

$$SplitEntropy_A(S) = -\sum_{i+1}^{n} \frac{|S_i|}{|S|} * log_2 \frac{|S_i|}{|S|} \tag{6}$$

Description:
S = Set of Case
A = Attributs
n = The number of A Attributr Partition
$|S_i|$ = The Case Number in i Partition
$|S|$ = The Case Number

Repeat the steps of determining root by calculating the highest gain value until all records are filled. The process of partitioning the tree will stops when: (1) There is no attribute in the partition which partitioned again. (2) There is no record in an empty branch.

The C4.5 algorithm has several weaknesses, including: (1) By a value of 0 or a value close to 0 it does not have any contribution to the classification and makes the tree size more complex. (2) Data that has noise tends to result in overfitting [16].

2.4 Adaptive Boosting (Adaboost)

Adaboost or Adaptive Boosting is a machine learning algorithm by Yoav Freud and Robert Schapire which is often used to improve the performance of certain algorithms from a set of strong or weak classifiers [17]. Adaboost can be combined with other algorithm classifiers to improve classification performance.

The method of the adaboost algorithm is as follows:

1. Initialize: weight of training sample$w\,_n^1 = 1/N$, which is n=1,...,N.
2. Do for t= 1, ...,T
3. Use component learn algorithm to train a classification component, h_t , to sample weight of training.
4. Training by minimizing error training or error function in $h_t\colon \varepsilon_t = \sum_{i=1}^{N} w_i^t\, , y_i \neq h_t(x_i)$
5. Update the weight sample of training $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(x_i)\}}{C_t}, i = \quad 1, \ldots, N\; C_t$ is the constant normalization.

Output: $f(x) \;= sign\,(\sum_{t=1}^{T} \alpha_t h_t(x))$ to make prediction using the last model. The stages of work flow can be seen in Figure 2.

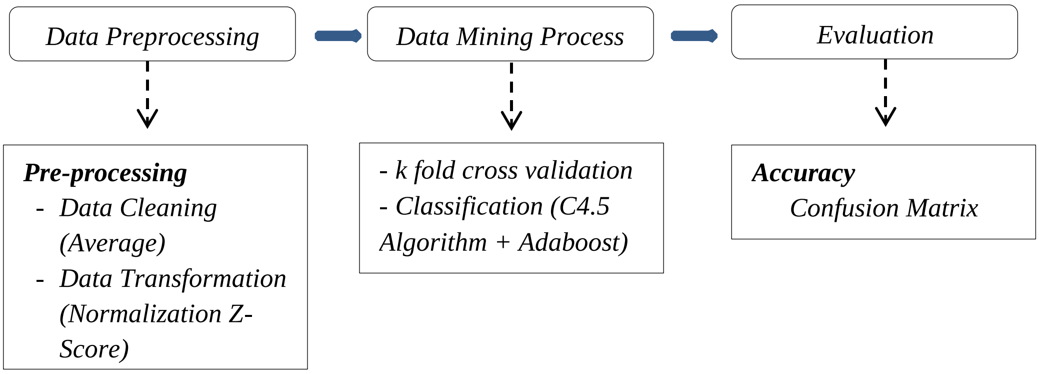

Data Preprocessing
Data Mining Process
Evaluation
Pre-processing
- Data Cleaning (Average)
- Data Transformation (Normalization Z-Score)
- k fold cross validation
- Classification (C4.5 Algorithm + Adaboost)
Accuracy
Confusion Matrix

Figure 2. The stage of workflow the Implementation C4.5 using Adaboost and Z-Score Normalization

3. RESULTS AND DISCUSSIONS

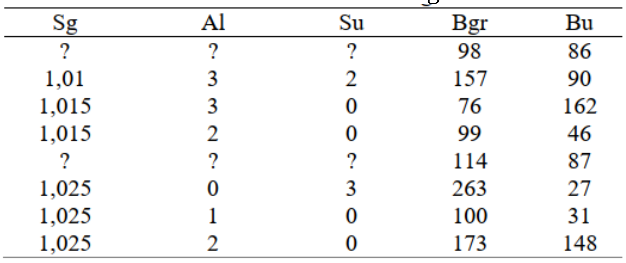
This study measures the accuracy of the C4.5 algorithm with the implementing of zscore normalization and adaboost using MatLab software. The data used in this study is the chronic kidney disease dataset obtained from the UCI repository of machine learning. The chronic kidney disease dataset consists of 400 data records which is divided into 24 attributes and 1 class attribute. The attributes consist of 11 numeric attributes and 14 nominal attributes.

This dataset had .arff format the it required to rewrite in the same form stored with the extension .xlsx. Before the classification process was conducted using the C4.5 algorithm, the data must be prepared in advance so it can be ready to be processed or well known as data pre-processing.

3.1 Handling Missing Value (Cleaning Data)

Cleaning data is a process of eliminating noise and handling data that has a missing value in a record. Data which has a missing value is usually symbolized by the question mark "?" in the data record. Therefore, it needs to be given the treatment or handling of missing value, by applying the average technique. The sample data consist missing values is shown in Table 1.

Table. 1 The Data with Missing Value



Sg	Al	Su	Bgr	Bu
?	?	?	98	86
1,01	3	2	157	90
1,015	3	0	76	162
1,015	2	0	99	46
?	?	?	114	87
1,025	0	3	263	27
1,025	1	0	100	31
1,025	2	0	173	148

The average calculation to replace the missing value data using average model as follows.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (07)$$

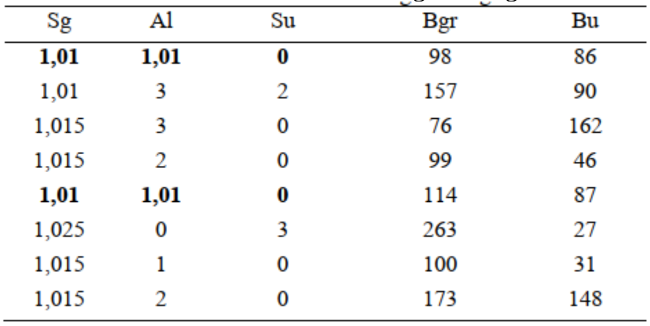
a. The value to replace the missing value of Sg attribute:
$\bar{x}(Sg) = \frac{\sum_1^{353} x(Sg)}{353} = \frac{359{,}145}{353} = 1{,}01$

b. The value to replace the missing value of Al attribute:
$\bar{x}(Al) = \frac{\sum_1^{354} x(Al)}{354} = \frac{360}{354} = 1{,}01$

c. The value to replace the missing value of Su attribute:
$\bar{x}(Su) = \frac{\sum_1^{351} x(Su)}{351} = \frac{158}{351} = 0{,}45 = 0$

The chronic kidney disease datasets with the handling of missing values is presented in Table 2.

Table 2. The Data After Handling Missing Value

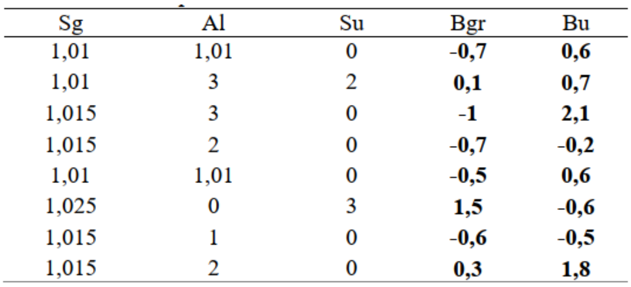


Sg	Al	Su	Bgr	Bu
1,01	**1,01**	**0**	98	86
1,01	3	2	157	90
1,015	3	0	76	162
1,015	2	0	99	46
1,01	**1,01**	**0**	114	87
1,025	0	3	263	27
1,015	1	0	100	31
1,015	2	0	173	148

3.2 Data Transformation Stage

The data transformation stage was conducted to normalize chronic kidney disease dataset using zscore normalization. It was transforming the numerical type data into patterns that coulb be identified to know the range values between dataset attributes so that data became simpler and had an even range of values between numeric attributes. The results of zscore normalization calculations can be seen in Table 3.

Table 3. The Implementation Result of Zscore Normalization

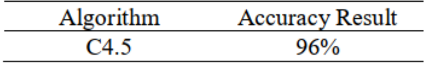


Sg	Al	Su	Bgr	Bu
1,01	1,01	0	**-0,7**	**0,6**
1,01	3	2	**0,1**	**0,7**
1,015	3	0	**-1**	**2,1**
1,015	2	0	**-0,7**	**-0,2**
1,01	1,01	0	**-0,5**	**0,6**
1,025	0	3	**1,5**	**-0,6**
1,015	1	0	**-0,6**	**-0,5**
1,015	2	0	**0,3**	**1,8**

3.2 Data Mining Stage

In this research the data distribution was conducted automatically by using k-fold cross validation with the default value k = 10. The testing result of the C4.5 algorithm by using the k-fold cross validation in the chronic kidney disease dataset can be seen in Table 4.

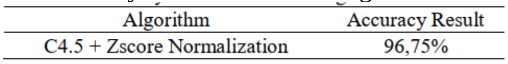
Tabel 4. The Accuracy Result of C4.5



Algorithm	Accuracy Result
C4.5	96%

a. The Implementation C4.5 Algorithm with Zscore Normalization
The implementation of classification was conducted by applying the C4.5 algorithm and zscore normalization method. The testing result of the C4.5 algorithm and zscore normalization as a pre-processing process by using the k-fold cross validation in chronic kidney disease datasets can be seen in Table 5.

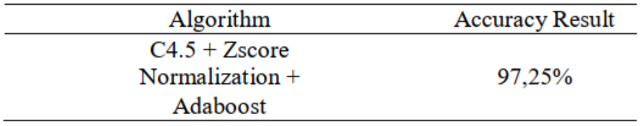
Table 5. The Accuracy Result of C4.5 Using Zscore Normalization



Algorithm	Accuracy Result
C4.5 + Zscore Normalization	96,75%

b. The Implementation C4.5 Algorithm and Adaboost with Normalization Zscore in Pre-Processing
This classification conducted by implementing the C4.5 algorithm and adaboost as ensemble learning. Before the mining process was conducted, the pre-processing was processed using zscore normalization. In this adaboost, the training set used for each classifier was selected based on the performance of the previous classifier. The distribution data was conducted by using k-fold cross validation with the default value k = 10. After the data was divided into training and testing data, then the data was processed with the C4.5 then boosting. The accuracy result was obtained and can be seen in Table 6.

Table 6. The Accuracy Result of C4.5 using Zscore Normalization and Adaboost



Algorithm	Accuracy Result
C4.5 + Zscore Normalization + Adaboost	97,25%

4. CONCLUSION

The implementation of classification was conducted by applying the C4.5 algorithm which obtained an accuracy of 96%. While the accuracy results by applying the C4.5 algorithm and zscore normalization obtained an accuracy of 96.75%. Then the best accuracy of the C4.5 algorithm was obtained by applying the zscore normalization method with boosting obtained an accuracy of 97.25%. Its accuracy result was higher and occur the increased accuracy by 1.25% compared to the accuracy results of the C4.5 algorithm.

REFERENCES

[1] Sugiharti, E. & Muslim, M.A. (2016). On-line Clustering of Lecturers Performance of Computer Science Department of Semarang State University Using K-Means Algorithm. Journal of Theoretical and Applied Information Technology, 83(1): 64-71.

[2] Tamilselvi, R., Sivasakthi, B., & Kavitha, R. (2015). An Efficient Preprocessing and Postprocessing Techniques in Data Mining. International Journal of Research in Computer Applications and Robotics,3(4): 80-85.

[3] Saranya, C., & Manikandan, G. (2013). A Study on Normalization Techniques for Privacy Preserving Data Mining. International Journal of Engineering and Technology (IJET), 5(3): 2701-2704.

[4] Goyal, H., Sandeep, Venu, Pokuri, R., Kathula, S., Battula, N. (2014). Normalization of Data in Data Mining. International Journal of Software and Web Science (IJSWS). 32-33.

[5] Han, J., Kamber, M. & Pei, J. (2011). Data Mining Concepts and Techniques, 3rd ed. USA: Morgan Kaufmann Publisher.

[6] Muslim, M.A., Herowati, A.J., Sugiharti, E., & Presetiyo, B. (2018). Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease. Journal of Physics: Conference Series, 983(1).

[7] Bala, S. & Kumar, K. (2014). A Literature Review on Kidney Disease Prediction using Data Mining Classification Techniques. International Journal of Computer Science and Mobile Computing, 3(7): 960-967.

[8] Sujatha, R. & Ezhilmaran. (2016). Performance Analysis of Data Mining Classification Techniques for Chronic Kidney Disease. International Journal of Pharmacy & Technology, 8(2): 13032-13037.

[9] Celik, E., Atalay, M., & Kondiloglu, A. (2016). The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods. International Journal of Intelligent Systems and Applications in Enggineering, 4(1): 27-31.

[10] Chary, N., & Rama, B. (2017). A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining. International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS), 3(1): 91-95.

[11] Handarko, J.L. & Alamsyah. (2015). Implementasi Fuzzy Decision Tree untuk Mendiagnosa Penyakit Hepatitis. Unnes Hournal of Mathematic, 4(2): 1-9.

[12] Mishra, A.K., Choudhary, A., & Choundhary, S. (2016). Normalization and Transformation Technique Based Efficient Privacy Preservation In Data Mining. International Journal of Modern Engineering and Research Technology, 3(2): 5- 10.

[13] Muzakir, A., & Wulandari, R.A. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. Scientific Journal of Informatics, 3(1): 19-26.

[14] Sampurno, G.I., Sugiharti, E., & Alamsyah, A. (2018). Comparison of Dynamic Programming Algorithm and Greedy Algorithm on Integer Knapsack Problem in Freight Transportation. Journal of Soft Computing Exploration, 5(1): 49.

[15] Dai, W., & Ji, W. (2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm. International Journal of Database Theory and Application, 7(1): 49- 60.

[16] Muslim, M.A., Rukmana, S.H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018). Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis. Journal of Physics: Conference Series, 983(1).

[17] Korada, N.K., Kumar, N.S.P., & Deekshitulu, Y.V.N.H. (2012). Implementation of Naïve Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System. International Journal of Information Sciences and Techniques (IJIST), 2(3): 63-75.

JOSCEX Journal of Soft Computing Exploration

Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms

Afifah Ratna Safitri[1], Much Aziz Muslim[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Aug 1, 2020
Revised Aug 14, 2020
Accepted Sept 5, 2020

Keywords:

Naive Bayes;
Customer Churn;
SMOTE;
Genetic Algorithms;

ABSTRACT

With increasing competition in the business world, many companies use data mining techniques to determine the level of customer loyalty. The customer data used in this study is the german credit dataset obtained from UCI. Such data have an imbalance problem of class because the amount of data in the loyal class is more than in the churn class. In addition, there are some irrelevant attributes for customer classification, so attributes selection is needed to get more accurate classification results. One classification algorithm is naive bayes. Naive Bayes has been used as an effective classification for years because it is easy to build and give an independent attribute into its structure. The purpose of this study is to improve the accuracy of the Naive Bayes for customer classification. SMOTE and genetic algorithm do for improving the accuracy. The SMOTE is used to handle class imbalance problems, while the genetic algorithm is used for attributes selection. Accuracy using the Naive Bayes is 47.10%, while the mean accuracy results obtained from the Naive Bayes with the application of the SMOTE is 78.15% and the accuracy obtained from the Naive Bayes with the application of the SMOTE and genetic algorithm is 78.46%.

Corresponding Author:

Afifah Ratna Safitri
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: afifahratna25@students.unnes.ac.id

1. INTRODUCTION

The rapid development of technology, information systems, and science has resulted in increasingly tight competition in the business world. In the business world, customers are the main asset. Therefore, various ways have been taken by the company so that customers do not stop subscribing. The term that is often used for customers who stop subscriptions with one service provider and become a customer of another service provider is called customer churn [1]. Customer churn occurs because of customer dissatisfaction [2]. This happened in various industries including insurance, banking, and the telecommunications industry [3]. To prevent this from happening, one of the models used by the company is Customer Relationship Management (CRM) [4].

Journal of Soft Computing Exploration, Vol. 6, No. 1, May 2019 2 The concept of Customer Relationship Management (CRM) leads to the importance of maintaining customers and building long-term relationships with customers to keep customers from moving to the company's competitors [5]. The transfer of customers from one provider to another is due to better rates or services, or because of the different benefits offered by the company's competitors when registering [6].

With the increasing competition and diversity of offerings in the industrial market, many companies utilize data mining techniques to determine customer churn rates [6]. Data mining is an activity to find interesting patterns from a large number of data [7]. Data mining has been applied to many fields because of its ability to analyze large amounts of data and fast time [8]. Data mining has several techniques such as estimation, classification, association, and clustering [9]. Companies need customer classifications to determine the level of customer loyalty. Classification is the most important part in data mining [10]. Classification is a data mining technique that serves to predict classes in a data [11]. One classification algorithm is Naive Bayes. Naive Bayes has been used as an effective classification for years. Because Naive Bayes is easy to build and can handle a number of independent variables randomly, either continuously or categorically [12].

In the field of machine learning and data mining the classification of unbalanced data is a problem that often occurs. Data imbalances have a negative impact on classification results where minority classes are often incorrectly classified as the majority class [11]. The problem of class imbalance is a problem where data experiences significant differences between classes, where loyal classes are greater than the churn class. The problem of class imbalance can be overcome by using the Synthetic Minority Over Sampling Technique (SMOTE) method. The SMOTE method is often used to overcome class imbalance problems because the SMOTE method does not reduce the amount of data, so that no information is lost [13].

Classification on high dimensional data will reduce accuracy. High dimensional data is data that has a large number of attributes. To improve classification accuracy on high-dimensional data can be used attribute selection methods that function to understand the relevant attributes [14]. One algorithm that can be used for attribute selection is a genetic algorithm. Genetic algorithms are chosen because they can reduce attributes in high dimensional data. So that data that initially has many attributes is reduced to a few fewer attributes, without reducing information from the data [15]. The concept of genetic algorithms is to search for solutions based on the evolutionary process [16].

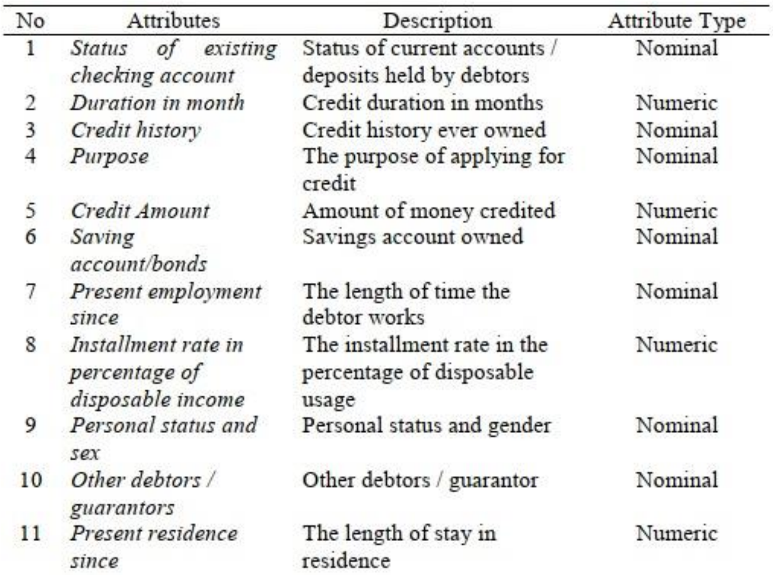
This study uses the German credit dataset. The dataset used in this study was taken from the UCI Machine Learning Repository. The purpose of this study is to improve the accuracy of the Naive Bayes algorithm by applying the SMOTE algorithm and attribute selection of Genetic Algorithms in classifying customers by seeing an increase in accuracy before and after the application of SMOTE and Genetic Algorithms.

2. METHOD

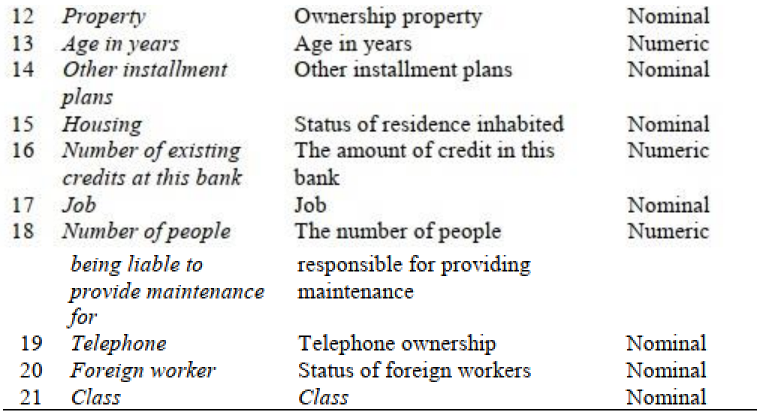
2.1 Dataset

The data used in this study are German Credit Data taken from the UCI Machine Learning Repository. The German Credit Data Collection has 20 attributes and 1000 instances. This dataset has 13 nominal type attributes and 7 numeric type attributes. This dataset has 1 class attribute of nominal type consisting of good and bad or loyal and churn. The description of the attributes of the German credit dataset can be seen in Table 1.

Table 1. German Credit Datasets Attributes



No	Attributes	Description	Attribute Type
1	*Status of existing checking account*	Status of current accounts / deposits held by debtors	Nominal
2	*Duration in month*	Credit duration in months	Numeric
3	*Credit history*	Credit history ever owned	Nominal
4	*Purpose*	The purpose of applying for credit	Nominal
5	*Credit Amount*	Amount of money credited	Numeric
6	*Saving account/bonds*	Savings account owned	Nominal
7	*Present employment since*	The length of time the debtor works	Nominal
8	*Installment rate in percentage of disposable income*	The installment rate in the percentage of disposable usage	Numeric
9	*Personal status and sex*	Personal status and gender	Nominal
10	*Other debtors / guarantors*	Other debtors / guarantor	Nominal
11	*Present residence since*	The length of stay in residence	Numeric



12	*Property*	Ownership property	Nominal
13	*Age in years*	Age in years	Numeric
14	*Other installment plans*	Other installment plans	Nominal
15	*Housing*	Status of residence inhabited	Nominal
16	*Number of existing credits at this bank*	The amount of credit in this bank	Numeric
17	*Job*	Job	Nominal
18	*Number of people being liable to provide maintenance for*	The number of people responsible for providing maintenance	Numeric
19	*Telephone*	Telephone ownership	Nominal
20	*Foreign worker*	Status of foreign workers	Nominal
21	*Class*	*Class*	Nominal

2.2 Experiment

In this study several algorithms were used to obtain a model to improve the accuracy of the Naive Bayes algorithm by using SMOTE and Genetic Algorithms. The experimental stages carried out in this study can be seen in Figure 1.

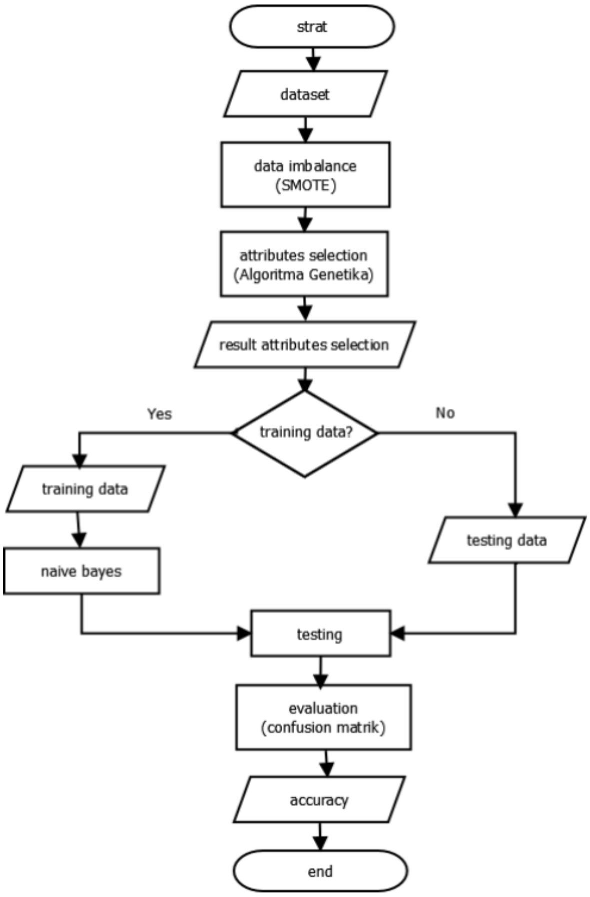


Figure 1. Experimental Stages of the Naive Bayes Method by Applying SMOTE and Genetic Algorithms

As seen in the picture above, the method used in this study is the application of the SMOTE algorithm and Genetic Algorithm to the Naive Bayes Classifier. The dataset that has been done by class balancing and attributes selection is then divided into training data and testing data for classification using the Naive Bayes Classifier. Data evaluation is done using cofussion matrix to calculate classification accuracy. The stages of each method can be seen as follows:

2.2.1 SMOTE

SMOTE is a technique used to expand minority sample data areas. This technique is made by making synthetic data for minority classes. Making synthetic data for minority classes in more detail can be seen as follows:

1. Enter the dataset and the amount of additional data that will be created. In this system, new minority class datasets generated as many as 300 new data.
2. Selecting minority class data, where in this dataset the minority class data is churn class data.
3. Separating minority data (churn) and majority class data (loyal). After the minority class data and the majority are separate and then remove the majority (loyal) class data.
4. Randomly select a minority dataset (churn) and calculate the selected k-nearest neighbor data. The k value used to calculate the k-nearest neighbor here is 3.
5. After that make new data based on randomly selected data and k- nearest neighbor by multiplying the distance that has been obtained in the fourth step with numbers chosen randomly between 0 and 1, then add the value of the original vector feature.
6. Repeat step 2 until the amount of new data corresponds to the number of additions to the desired data, where in this dataset the desired amount of new data is 300 data churn. 7. After all stages have been completed, 300 new minority data will be known so that there are 1300 sample data.

2.2.2 Genetic Algorithms

The stages of Genetic Algorithms can be seen as follows:

```
Awaken teh initial population of chromosome N.
Loop until the stop condition is fulfilled
    Loop for N chromosome N
          Individual = Decode (chromosome)
           Fitness    = Evaluation (individual)
    End
    Make one or two of the best chromosome copies
    Loop until you get a new N chromosome
          Select two chromosome as parents P1 and P2
          [parent1, parent2] = Recombination (P1, P2)
          [child1, child2]     = Mutation (child1, child2)
    End
    Change the old N chromosome with the new N chromosome.
End
```

2.2.3 Naïve Bayes Classifier

The stages of the Naive Bayes algorithm in classifying datasets are as follows:

1. Read training data.
2. Calculating probability in the following way:
   a. Calculates the average of each parameter with the following formula:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Information:
μ : mean
xi : sample value i
n : number of samples

   b. Calculates the standard deviation of each parameter with the following formula:

$$\sigma^2 = \frac{1}{n-1} \sum (x_i - \mu)^2 \tag{2}$$

Information:
σ : Standard deviation, expresses the variance of all attributes
n : Amount of data in the same class
x_i : Value of attribute to i
μ : mean

c. Look for probability values using the formula:

$$P(X_i = x_i \mid Y = y_J) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{3}$$

Information:

σ : Standard deviation, expresses the variance of all attributes
x_i : Value of attribute to i
μ : mean
X_i : attribute to i
Y : Class sought
y_j : Y syb-class searched

3. Repeat step 2 until the probability of all parameters is calculated.
4. The calculation process will stop when the probability value of all parameters of each attribute has been calculated.

3. RESULT AND DISCUSSION

3.1 Results

This study uses a system created with the PHP programming language that is applied to German Credit datasets. Accuracy results obtained on the application of Naive Bayes without the pre-processing process which is equal to 73%. Whereas, the results of the average accuracy of ten executions obtained using SMOTE and the attibutes selection of Genetic Algorithms on Naive Bayes is 80.948%.

3.2 Discussion

Based on the results of the application of the SMOTE algorithm and the attributes selection of Genetic Algorithm in the Naive Bayes algorithm that has been carried out, it can be seen that the accuracy for determining customer churn using the German Credit dataset is taken from the UCI Machine Learning Repository. Data previously obtained has passed the pre-processing stage, namely the class balancing stage and attibutes selection stage.

At the stage of class balancing is done by applying the SMOTE algorithm. The SMOTE algorithm is applied to make new data more balanced. German Credit's initial dataset has 1000 samples with 700 loyal (good) classes and 300 churn (bad) classes. Therefore it is necessary to balance the class by creating new data in the churn class. The new dataset of the SMOTE algorithm results in 300 churn class data, so there are 1300 new sample data. This is done so that data can be classified optimally. The attribute selection stage is done by selecting attributes in the data used. In this attribute selection stage there is a dimension reduction in the data in order to optimize attributes that will affect the accuracy of the Naive Bayes algorithm. Dimension reduction in attributes is done by using Genetic Algorithms. Removal of attributes is done one by one from attributes that have the smallest fitness value and will be mining. The process of selecting attributes and mining will stop when the results of the accuracy have exceeded the specified minimum limit.

After going through the pre-processing stage, new data will go through the classification process using the Naive Bayes algorithm. From the results obtained, there is an increase in the accuracy of the Naive Bayes algorithm and the Naive Bayes algorithm by applying the SMOTE algorithm and attibutes selection of Genetic Algorithms.

4. CONCLUSION

In this study, testing the Naive Bayes algorithm by applying the SMOTE algorithm and attribute selection of Genetic Algorithms is done using the German Credit dataset taken from the UCI Machine Learning Repository to classify churn and loyal customers. Accuracy results obtained on the application of the Naive Bayes algorithm without the pre-processing process that is equal to 73%. Meanwhile, the average accuracy of ten executions obtained using the SMOTE algorithm in the Naive Bayes algorithm is 74.918% and the results of the average accuracy of ten executions obtained using the SMOTE algorithm and the attributes selection of the Genetic Algorithm of the Naive Bayes algorithm is 80.948%.

REFERENCES

[1] V. Mahajan, R. Misra, R. Mahajan. “Review on factors affecting customer churn in telecom sector”, International Journal of Data Analysis Techniques and Strategies, 9(2), pp. 122-144, 2017.

[2] A. A. Q. Ahmed, D. Maheswari. “Churn prediction on huge telecom data using hybrid firefly based classification”, Egyptian Informatics Journal, 18(3), pp. 215-220, 2017.

[3] R. Hejazinia, M. Kazemi. “Prioritizing Factors influencing customer churn”, Interdisciplinary Journal of Contemporary Research in Business, 5(12), pp. 227-236, 2014.

[4] P. K. Banda, S. Tembo. “Application of System Dynamics to Mobile Telecommunication Customer Churn Management”, Journal of Telecommunication, Electronic and Computer Engineering, 9(3), pp. 67-76, 2017.

[5] H. S. Soliman. “Customer Relationship Management and Its Relationship to the Marketing Performance”, International Journal of Business and Social Science, 2(10), pp. 166-182, 2011.

[6] I. Brandusoiu, G. Toderean. “Churn prediction in the telecommunications sector using support vector machines”, Annals of the Oradea University, 22(1), pp. 19-22, 2013.

[7] M. A. Muslim, A. J. Herowati, E. Sugiharti, B. Prasetiyo. “Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease”, Journal of Physics: Conf. Series, 983, pp. 1-9, 2017.

[8] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, S. Alimah “Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis”, Journal of Physics: Conf. Series, 983, pp.1-7, 2017.

[9] P. Sinha, P. Sinha. “Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM”, International Journal of Engineering Research & Technology (IJERT), 4(12), pp. 608-612, 2015.

[10] Makhtar, S. Nafis, M. A. Mohamed, M. K. Awang, M. N. A. Rahman, M. M. Deris. “Churn Classification Model for Local Telecommunication Company Based on Rough Set Theory”, Journal of Fundamental and Applied Sciences, 9(6S), pp. 854-868, 2017.

[11] H. Lee, J. Kim, S. Kim. “Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions”, International Journal of Fuzzy Logic and Intelligent Systems, 17(4), pp. 229-234, 2017.

[12] M. H. A. Elhebir, A. Abraham. “A Novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification”, International Journal of Computer Information Systems and Industrial Management Applications, 7, pp. 189-195, 2015.

[13] M. Anis, M. Ali. “Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets”, European Scientific Journal, 13(33), pp. 341-353, 2017.

[14] L.Marlina, M. A. Muslim, A. P. U. Siahaan, “Data Mining Classification Comparison (Naive Bayes and C4.5 Algorithms)”, International Journal of Engineering Trends and Technology (IJETT), 38(7), pp. 382-383, 2016.

[15] C. Kirui, L. Hong, W. Cheruiyot, H. Kirui. “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining”, International Journal of Computer Science Issues, 10(1), pp. 165-172, 2013.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”, Journal of Artificial Intelligence Research, 16, pp. 321-357, 2002.