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.

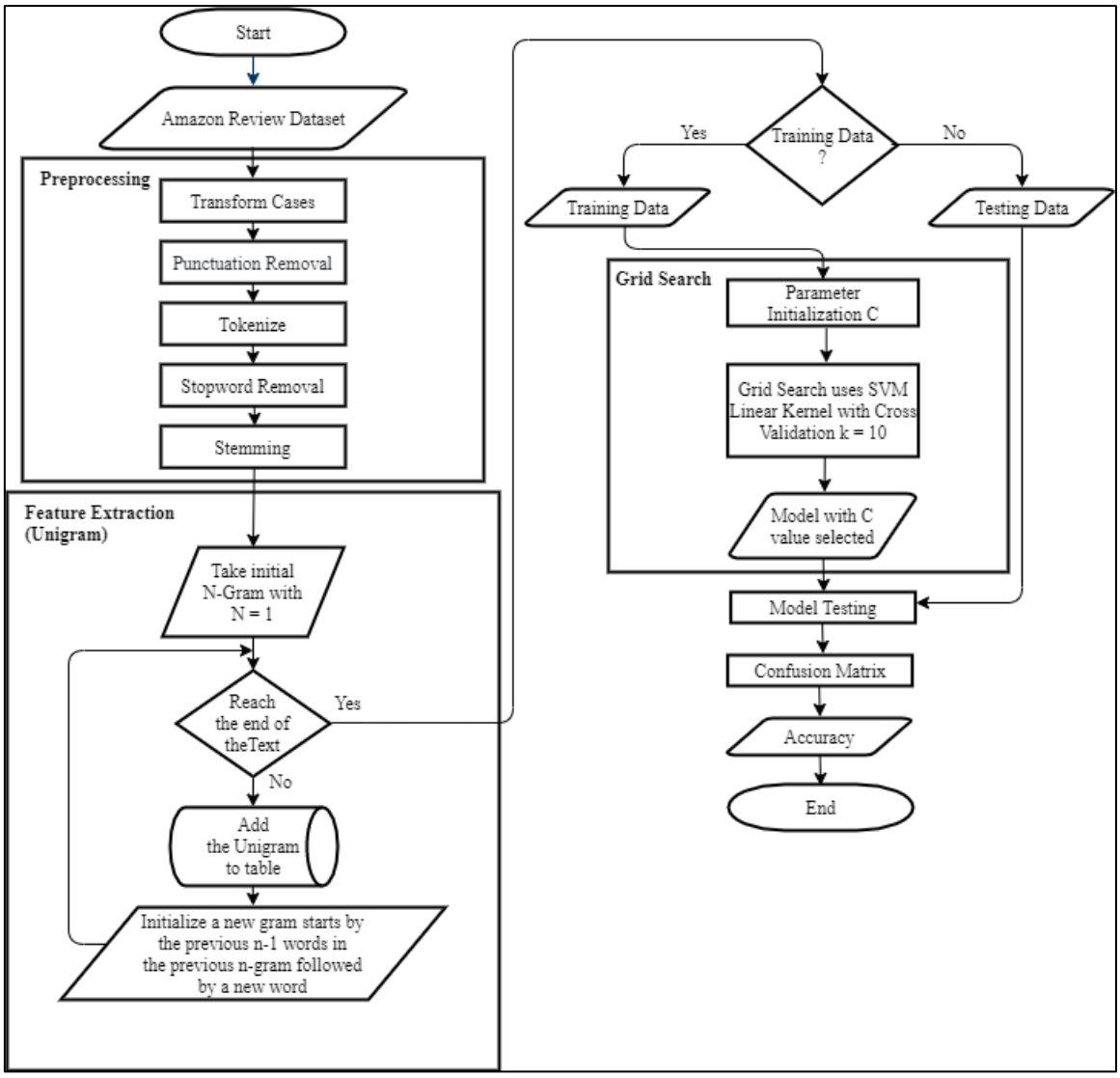

Figure 1. Flowchart SVM algorithm with Grid Search and Unigram

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.

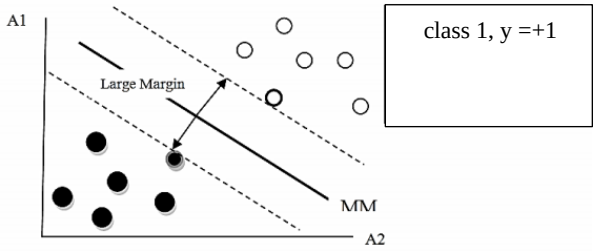

Figure 2. Separating 2 Data Classes with Maximum Margin

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 ($f11+f00$) and numbers of wrongly predicted data ($f10 + f01$). 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.

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

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%

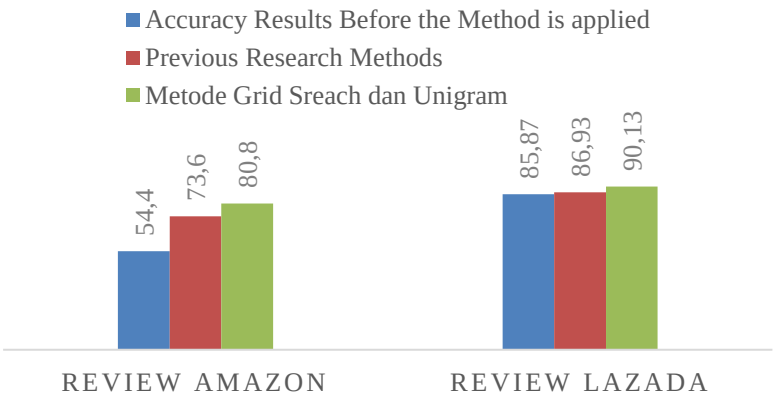

Figure 4. Comparison with previous work

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.