Sentiment Mining Using SVM-Based Hybrid Classification Model

G. Vinodhini and R. M. Chandrasekaran

Abstract With the rapid growth of social networks, opinions expressed in social networks play an influential role in day-to-day life. A need for a sentiment mining model arises, so as to enable the retrieval of opinions for decision making. Though support vector machine (SVM) has been proved to provide a good classification result in sentiment mining, the practically implemented SVM is often far from the theoretically expected level because their implementations are based on the approximated algorithms due to the high complexity of time and space. To improve the limited classification performance of the real SVM, we propose to use the hybrid model of SVM and principal component analysis (PCA). In this paper, we apply the concept of reducing the data dimensionality using PCA to decrease the complexity of an SVM-based sentiment classification task. The experimental results for the product reviews show that the proposed hybrid model of SVM with PCA outperforms a single SVM in terms of classification accuracy and receiver-operating characteristic curve (ROC).

Keywords Sentiment · Opinion · Mining · Hybrid model · PCA

1 Introduction

With the rapid growth of e-commerce and large number of online reviews in digital form, the need to organize them arises. Various machine learning classifiers have been used in sentiment classification [8]. Many studies in machine learning communities have shown that combining individual classifiers is an effective technique for improving classification accuracy. There are different ways in which classifier can be combined to classify new instances. Dimension reduction plays an

G. Vinodhini (✉) · R. M. Chandrasekaran
Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram 608002, India
e-mail: g.t.vino@gmail.com

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_18, © Springer India 2014

important part in optimizing the performance of a classifier by reducing the feature vector size. Principal component analysis (PCA) can transform the original dataset of correlated variables into a smaller dataset of uncorrelated variables that are linear combinations of the original ones. Support vector machines (SVMs) have been recognized as one of the most successful classification methods for many applications including sentiment classification. Even though the learning ability and computational complexity of training in support vector machines may be independent of the dimension of the feature space, reducing computational complexity is an essential issue to efficiently handle a large number of terms in practical applications of text sentiment classification. In this study, we introduce a SVM-based hybrid sentiment classification model with PCA as dimension reduction for online product reviews using the product attributes as features. The results are compared with an individual statistical model, i.e., support vector machine. The rest of the paper is organized as follows. We provide an overview of the related work in Sect. 2. Section 3 presents an overview of the data source used. Methodology of the work is discussed in Sect. 4, which presents the feature reduction and classification methods used. The experimental results are discussed in Sect. 5 and conclusion in Sect. 6.

2 Related Work

Many studies on sentiment classification have used machine learning algorithms, with SVM and naive Bayes (NB) being the most commonly used. In comparison, SVM has outperformed other classifiers such as NB, centroid classifier, k-nearest neighbor, and window classifier [4–7, 9–11, 13]. So in this study, we considered SVM as baseline classifier. Feature selection is the most crucial task in sentiment mining [12]. Tan and Zhang [7] presented a sentiment categorization using four feature selection methods. The experimental results indicate that information gain performs the best for selecting the sentimental terms. Wang et al. [10] presented a hybrid method for feature selection based on the category-distinguishing capability of feature words and IG. Gamon [1] presented a feature reduction technique based on log-likelihood ratio to select the important attributes from a large initial feature vectors. Hatzivassiloglou and Wiebe [2] used words, bigrams, and trigrams, as well as the parts of speech as features in each sentence. In recent years, we witnessed the advance in machine learning methodology, such as SVM and PCA independently. However, the literature does not contribute much to sentiment classification using hybrid classification model with the combination of SVM and PCA. Even though Prabowo and Thelwall [5] used multiple classifiers in a hybrid manner, the hybrid combination of SVM and PCA is not experimented so far.

Our contribution in this work shows the effectiveness of the proposed hybrid classification method in sentiment mining of product reviews. Though SVM can efficiently deal with high-dimensional data, the linear dependence between different variables of the sample influence the generalization of SVM method. On the contradiction, PCA can deal with linear dependence between variables effectively and reduce dimension of the input samples and strengthen the ability of SVM to approximate to a nonlinear function. PCA is a suitable dimension reduction method for SVM classifiers because SVM is invariant under PCA transform. Though SVM has been proved to provide a good classification result in sentiment mining, the practically implemented SVM is often far from the theoretically expected level because their implementations are based on the approximated algorithms due to the high complexity of time and space. To improve the limited classification performance of the real SVM, we propose to use the hybrid model of SVM and PCA. Sentiment analysis is conducted at feature-based sentence level.

3 Data Source

The dataset used contains review sentences of products, which were labeled as positive, negative, or neutral. We collected the review sentences from the publicly available customer review dataset. This dataset can be downloaded from http://www.cs.uic.edu/~liub/FBS/FBS.html. This dataset contains annotated customer reviews of five different products. From those five products, we have selected reviews of digital camera, mobile phone, and music player. For our classification problem, we have considered only positive reviews and negative reviews. The product attributes discussed in the review sentences are collected. Unique unigram product features alone are grouped, which results in a final list of product attributes (features) for each product. In terms of these, the descriptions of review dataset model (model I) to be used in the experiment are given in Table 1.

4 Methodology

The following steps are involved in this work to develop the prediction system.

4.1 Create a Word Vector Model

A word vector representation of review sentences is developed for model using the unigram features. To create the word vector list, the review sentences are

Table 1 Description of dataset

Product	No. of reviews	Feature	No. of features	Positive reviews	Negative reviews
Camera	500	Unigrams only	95	365	135
Mobile phone	456	Unigrams only	86	248	208
Music player	328	Unigrams only	58	174	154

preprocessed. The data preprocessing includes tokenization, stopping word removal, and stemming. After preprocessing, the reviews are represented as unordered collection of words and the features are modeled as a bag of words. A word vector is developed for model using the respective features based on the term binary occurrences. The binary occurrences of the each feature word (n) in the processed review sentences (m) result in a word vector X of size $m \times n$ for model I.

4.2 Dimension Reduction Using PCA

PCA is a common technique for finding patterns in high-dimensional data. PCA algorithm works by calculating the covariance matrix, eigenvalues, and eigenvector. Then, the dimensionality of the data is reduced. A standardized transformation matrix is developed, and finally, the reduced components are obtained. These new components are called principal components (PC). Using Weka, the PC for the model with their unigram features are identified. The PC with variance less than 0.95 are obtained. A word vector model is recreated using review sentences and the reduced PC. The description of principle components obtained for model is shown in Table 2.

4.3 Construct SVM Model as Baseline

Support vector machine belongs to a family of generalized linear classifiers. It is a supervised machine learning approach used for classification to find the hyperplane maximizing the minimum distance between the plane and the training points. The basic idea behind the training procedure for binary classification is to find a hyper plane that separates the document vectors in one class from those in the other. Also the separation margin is as larger as possible. The SVM model is employed using Weka tool. The kernel type chosen is the polynomial kernel with default values for kernel parameters such as cache size and exponent.

Table 2 Description of principle components

Product	Camera	Mobile	Music player
No. of components	PC1–PC57	PC1–PC45	PC1–PC24
Variance	<0.95	<0.95	<0.95
Standard deviation	0.67	0.62	0.07
Proportion of variance	0.003	0.002	0.002
No. of features (original)	95	86	58
No. of principle components (reduced)	57	45	24
No. of reviews	500	456	328
Positive reviews	365	248	174
Negative reviews	135	208	154

4.4 Construct Hybrid SVM Model

A hybrid classifier is a collection of several classifiers whose individual decisions are combined in such a way to classify the test examples. It is known that combined model often shows much better performance than the individual classifiers used. SVM has been known to show a good generalization performance and is easy to learn exact parameters for the global optimum. Due to these advantages, their combination may not be considered as a method for increasing the classification performance. However, when implementing SVM practically, approximated algorithms have been used in order to reduce the computation complexity of time and space. Thus, a single SVM may not learn exact parameters for the global optimum. Sometimes, the support vectors obtained from the learning are not sufficient to classify all unknown test examples completely. So, we cannot guarantee that a single SVM always provides the global optimal classification performance over all test examples. Moreover, the dimension of the feature space does not influence the computational complexity of training or testing due to the use of the kernel function. But the computational complexity of SVM training depends on the dimension of the input space. Therefore, more efficient testing and training is expected from dimension reduction. To overcome these limitations, we propose to use hybrid model of support vector machines with PCA as dimension reduction technique. In hybrid model, each individual SVM is trained independently using the randomly chosen dimension-reduced training samples via a bootstrap technique and then aggregated. Each classifier is trained on a sample of examples taken from the training set and thereby produces a combined model that often performs better than the single model built from the original single training set. The algorithm is shown in Fig 1.

Fig. 1 Hybrid algorithm

```
Input: Data set $D = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$;
Base learning algorithm $B$; //SVM
Number of learning rounds $R$. //10
Process: for $i = 1,\cdots,R$:
/* Generate a bootstrap sample from $D$*/
$Di$ = Bootstrap ($D$);
/* Train base learner $hi$ from the bootstrap sample */
$hi = B(Di)$;
end.
/* the value of $l$(a) is 1 if a is true and 0 */
Output : $O(x) = \arg\max y \in Y \sum_{t=1}^{T} l(y = h_i(x))$
```

4.5 Aggregating Support Vector Machines

After training, we need to aggregate several independently trained SVMs in an appropriate combined manner. Majority voting is to the simplest method for combining several SVMs. The results of individual SVMs are aggregated by determining the large number of SVMs whose decisions are known. The LSE-based weighting treats several SVMs in the SVM ensemble with different weights. Often, the weights of several SVMs are determined in proportional to their accuracies of classifications. The double-layer hierarchical combining uses another SVM to aggregate the outputs of several SVMs in the SVM ensemble. So, this combination consists of double-layer SVMs hierarchically where the outputs of several SVMs in the lower layer are fed into a super SVM in the upper layer.

5 Results

For hybrid models, we trained groups of 10 individual SVM models, each with different bootstrapped training subsets. The results of the individual classification model in the hybrid were combined using various methods such as the majority voting, the LSE-based weighting, and the double-layer hierarchical combining to produce the final result. We used tenfold cross-validation to compare classification accuracy. The performance evaluation is done using accuracy as a measure (Table 3.). Experiment is conducted on the customer reviews of digital camera, mobile phone, and music player (Sect. 3).

Receiver-operating characteristic (ROC) curve is an alternative technique for selecting classifiers based on their performance in which true-positive rate (TP rate) is plotted along the y-axis and false-positive rate (FP rate) is plotted along the y-axis. Since SVM is a discrete classifier, it is represented by only one point on an ROC graph. So a very approximate ROC curve is constructed by connecting this point with the points denoting both default classifiers [3].

Figures 2, 3 and 4 show ROC graph to compare the performance of baseline (SVM) method and hybrid method for three different product reviews. The hybrid method performed well and often outperformed the single SVM in terms of ROC curves.

Table 3 Classification accuracy

	Accuracy (%)		
Method	Camera	Mobile phone	Music player
Single SVM	86.99	85.34	88.73
Hybrid SVM—Majority voting	87.55	86.05	90.51
Hybrid SVM—LSE-based weighting	87.82	86.98	90.87
Hybrid SVM—Hierarchical SVM	88.01	87.57	91.08

Fig. 2 ROC curve for camera reviews

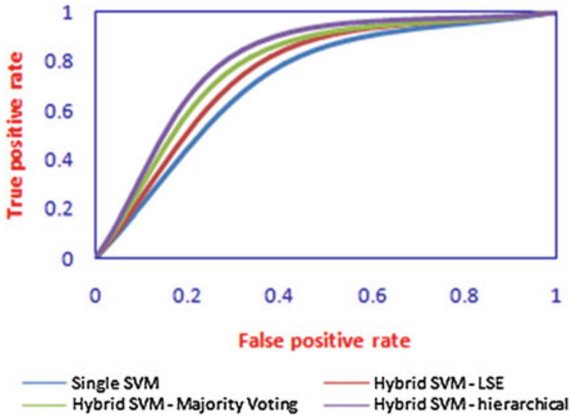


Fig. 3 ROC curve for mobile phone reviews

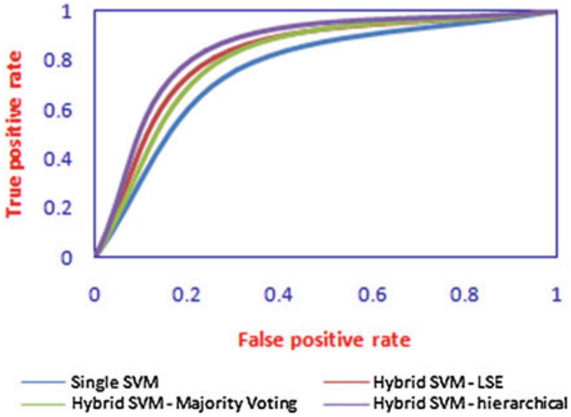
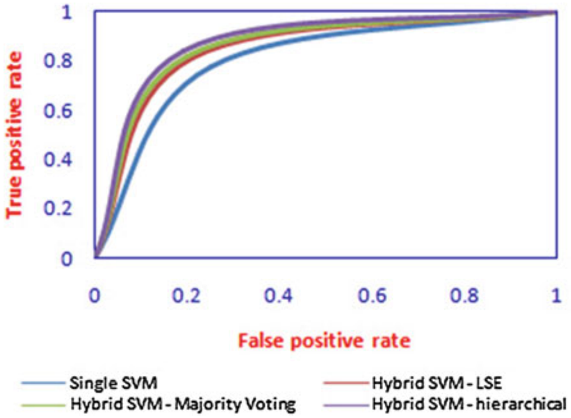


Fig. 4 ROC curve for music player reviews



6 Conclusion

We evaluated the classification performance of the proposed hybrid SVM over three different product reviews such as digital camera, mobile phone, and music player. The hybrid SVM with PCA outperforms a single SVM for all products in terms of classification accuracy and ROC curve. For three different aggregation methods, the classification performance is superior in the order of the double-layer hierarchical combining, the LSE-based weighting, and the majority voting. Thus, the hybrid model seems to be the best choice if time and computational resources are not an issue. In the future, we will consider other classifier combinations and aggregation schemes.

References

1. Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of the 20th International Conference on Computational Linguistics (p. 841).
2. Hatzivassiloglou, V., Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In International Conference on Computational Linguistics (COLING-2000).
3. Matjaz Majnik, Zoran Bosni, ROC Analysis of Classifiers in Machine Learning: A Survey,Technical report MM-1,2011.
4. Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP'.
5. Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3, 143–157.
6. Rui Xia, Chengqing Zong, Shoushan Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 181 (2011) 1138–1152.
7. Tan, S. B., & Zhang, J. (2008). An Empirical study of sentiment analysis for Chinese documents. Expert Systems with Application, 34(4), 2622–2629.
8. Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36, 10760–10773.
9. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics, (pp. 417–424). Philadelphia, PA.
10. Wang, S. G., Wei, Y. J., Zhang, W., Li, D. Y., & Li, W. (2007). A hybrid method of feature selection for chinese text sentiment classification [C]. In Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (pp. 435–439). IEEE Computer Society.
11. Wilson,T., Wiebe, J., Hoffman, P. (2005). Recognizing contextual polarity in phrase level sentiment analysis. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 347–354). British Columbia, Canada.
12. Yang, Y. M., & Pedersen, J. O. (1997). A Comparative study on feature selection in text categorization. ICML, 412–420.
13. Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, Sentiment classification of Internet restaurant reviews written in Cantonese, Expert Systems with Applications 38 (2011) 7674–7682.