Information Processing and Management 53 (2017) 814–833

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification

Aytuğ Onan[a,*], Serdar Korukoğlu[b], Hasan Bulut[b]

[a] Celal Bayar University, Department of Computer Engineering, 45140, Muradiye, Manisa, Turkey
[b] Ege University, Department of Computer Engineering, 35100, Bornova, Izmir, Turkey

ARTICLE INFO

Article history:
Received 16 May 2016
Revised 26 December 2016
Accepted 16 February 2017

Keywords:
Ensemble pruning
Consensus clustering
Multi-objective evolutionary algorithm
Sentiment classification

ABSTRACT

Sentiment analysis is a critical task of extracting subjective information from online text documents. Ensemble learning can be employed to obtain more robust classification schemes. However, most approaches in the field incorporated feature engineering to build efficient sentiment classifiers.

The purpose of our research is to establish an effective sentiment classification scheme by pursuing the paradigm of ensemble pruning. Ensemble pruning is a crucial method to build classifier ensembles with high predictive accuracy and efficiency. Previous studies employed exponential search, randomized search, sequential search, ranking based pruning and clustering based pruning. However, there are tradeoffs in selecting the ensemble pruning methods. In this regard, hybrid ensemble pruning schemes can be more promising.

In this study, we propose a hybrid ensemble pruning scheme based on clustering and randomized search for text sentiment classification. Furthermore, a consensus clustering scheme is presented to deal with the instability of clustering results. The classifiers of the ensemble are initially clustered into groups according to their predictive characteristics. Then, two classifiers from each cluster are selected as candidate classifiers based on their pairwise diversity. The search space of candidate classifiers is explored by the elitist Pareto-based multi-objective evolutionary algorithm.

For the evaluation task, the proposed scheme is tested on twelve balanced and unbalanced benchmark text classification tasks. In addition, the proposed approach is experimentally compared with three ensemble methods (AdaBoost, Bagging and Random Subspace) and three ensemble pruning algorithms (ensemble selection from libraries of models, Bagging ensemble selection and LibD3C algorithm). Results demonstrate that the consensus clustering and the elitist pareto-based multi-objective evolutionary algorithm can be effectively used in ensemble pruning. The experimental analysis with conventional ensemble methods and pruning algorithms indicates the validity and effectiveness of the proposed scheme.

© 2017 Elsevier Ltd. All rights reserved.

* Corresponding author.
E-mail addresses: aytug.onan@cbu.edu.tr (A. Onan), Serdar.korukoglu@ege.edu.tr (S. Korukoğlu), hasan.bulut@ege.edu.tr (H. Bulut).

http://dx.doi.org/10.1016/j.ipm.2017.02.008
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Ensemble learning is an important research direction of pattern recognition and machine learning. The main idea behind ensemble learning is to combine the predictions of multiple classification algorithms so that a more robust and accurate classification model can be constructed (Dietterich, 2000). With the use of ensemble learning, remarkable improvement in generalization ability can be achieved. In addition, the variance and bias of classification and the dependency of the results to the odd characteristics of a single training set can be reduced (Kuncheva, 2014). Ensemble learning can be successfully utilized for supervised learning (such as classification and regression) and for unsupervised learning (such as cluster analysis) (Strehl & Ghosh, 2003).

Sentiment analysis (also known as opinion mining) is a subfield of natural language processing and text mining, which aims to classify text documents as positive, negative and neutral. The Web is a rich, widely distributed source of information with a progressively expanding volume of data (Bhatia & Khalid, 2008). The information available on the Web can provide valuable information to governments, business organizations and individual decision makers. Public sentiment toward policies, product and services can be beneficial to organizations (Wang, Sun, Ma, Xu, & Gu, 2014). The identification of subjective information is very important to generate structured knowledge that will serve crucial information to decision support systems and individual decision makers (Fersini, Messina, & Pozzi, 2014). Hence, sentiment analysis is an important research direction. Recent research contributions on sentiment analysis indicate that the predictive performance of sentiment classification can be greatly improved with the use of ensemble learning methods (Fersini et al., 2014; Wang et al., 2014; Xia, Zong, & Li, 2011).

The ensemble learning process consists of three main phases, which are the ensemble generation, the ensemble pruning and the ensemble integration (Mendes-Moreira, Soares, Jorge, & De Sousa, 2012; Roli, Giacinto, & Vernazza, 2001). In the ensemble generation phase, the classification algorithms to be utilized in the classifier ensemble are generated. The learning algorithms can be generated either homogeneous or heterogeneous. In homogeneous classifier ensembles, the same learning algorithm is utilized. In this scheme, the diversity is achieved by taking different parameter values, by randomization of the learning process, by differentiation of the training subsets and/or by taking different input attributes (Tsoumakas, Partalas, & Vlahavas, 2008). Bagging and Boosting algorithms are two well-known representatives for homogenous classifier ensemble methods. In contrast, heterogeneous classifier ensembles are generated by using different learning algorithms. In this scheme, high diversity of the ensemble is expected. In the ensemble integration, the predictions of multiple learning algorithms are combined by using an ensemble combination rule, such as majority voting and stacked generalization.

The ensemble pruning (also known as selective ensemble, ensemble thinning and ensemble selection) is the process of obtaining a subset of classifiers from the classifier ensemble so that the predictive performance and computational efficiency of the ensemble have been enhanced. It has been empirically validated that the utilization of some classification algorithms, rather than all available classifiers, enhances the predictive performance of the classifier ensemble (Zhou, Wu, & Tang, 2002). One of the critical issues in developing classifier ensembles is to provide high diversity (Gashler, Giraud-Carrier, & Martinez, 2008). A diverse subset of classifiers can be obtained with the ensemble pruning. The ensemble pruning methods can be broadly divided into five groups, as exponential search, randomized search, sequential search, ranking based pruning and clustering based pruning methods (Mendes-Moreira et al., 2012). In exponential search methods, all possible subsets of the classification algorithms within the ensemble are taken into consideration. This requires enumerating 2^k-1 possible subsets for a classifier ensemble containing k classification algorithms, which makes the search space large. In randomized search methods, the metaheuristic search algorithms are utilized to explore more effectively the search space of possible classifiers. The metaheuristic algorithms utilized in the ensemble pruning include genetic algorithms, tabu search, population based incremental learning and stochastic search (Ruta & Gabrys, 2001; Zhou & Tang, 2003; Partalas et al., 2006). In sequential search methods, the search can be forward, backward or forward-backward. In the forward methods, the search starts with an empty classifier ensemble. In each iteration, a learning algorithm is added to the ensemble. In contrast, classifier ensemble starts with all learning algorithms and the learning algorithms are iteratively eliminated from the ensemble in the backward methods. In ranking based pruning, the learning algorithms of the classifier ensemble are ranked based on a particular evaluation criterion. The k learning algorithms with the highest evaluation value are included in the pruned ensemble, whereas the other algorithms are eliminated. In clustering based pruning methods, a clustering algorithm (such as k-means) is utilized to group the classification algorithms within the ensemble into clusters based on the predictive performance of the algorithms. In this scheme, a number of classifiers are selected from each cluster to construct the pruned classifier ensemble (Mendes-Moreira et al., 2012).

The identification of an appropriate subset of classifiers is an NP-complete problem and it is a computationally intensive task. Since exponential methods require enumerating all possible subsets, these methods are only suitable for a classifier ensemble with a few classifiers (Martinez-Munoz & Suarez, 2006; Tamon & Xiang, 2000). The ranking based pruning methods require computing an evaluation criterion for each classifier. Since the decision of classifiers to be included within the ensemble is based on the evaluation criterion, the ranking based methods are computationally efficient techniques. However, the predictive performance of classifier ensembles obtained by ranking based pruning is relatively low (Pinto, 2013). The clustering based pruning methods may suffer from the cluster instability (Lin et al., 2014). There are tradeoffs in selecting the ensemble pruning methods.

The development and application of hybrid algorithms is a promising research interest in machine learning. In addition, recent research contributions in the ensemble pruning indicate that an effective ensemble pruning scheme can be built

Download English Version:

https://daneshyari.com/en/article/4966467

Download Persian Version:

https://daneshyari.com/article/4966467

Daneshyari.com