Decision Support Systems 57 (2014) 77–93

Contents lists available at ScienceDirect

Decision Support Systems

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

Sentiment classification: The contribution of ensemble learning

Gang Wang [a,b,c,*], Jianshan Sun [c,d], Jian Ma [c], Kaiquan Xu [e], Jibao Gu [d]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, PR China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui, PR China
[c] Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[d] School of Management, University of Science and Technology of China, Hefei, Anhui, PR China
[e] Department of Electronic Commerce, School of Business, Nanjing University, Nanjing, Jiangsu 210093, PR China

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form 1 August 2013
Accepted 5 August 2013
Available online 15 August 2013

Keywords:
Sentiment classification
Ensemble learning
Bagging
Boosting
Random Subspace

ABSTRACT

With the rapid development of information technologies, user-generated contents can be conveniently posted online. While individuals, businesses, and governments are interested in evaluating the sentiments behind this content, there are no consistent conclusions on which sentiment classification technologies are best. Recent studies suggest that ensemble learning methods may have potential applicability in sentiment classification. In this study, we conduct a comparative assessment of the performance of three popular ensemble methods (Bagging, Boosting, and Random Subspace) based on five base learners (Naive Bayes, Maximum Entropy, Decision Tree, K Nearest Neighbor, and Support Vector Machine) for sentiment classification. Moreover, ten public sentiment analysis datasets were investigated to verify the effectiveness of ensemble learning for sentiment analysis. Based on a total of 1200 comparative group experiments, empirical results reveal that ensemble methods substantially improve the performance of individual base learners for sentiment classification. Among the three ensemble methods, Random Subspace has the better comparative results, although it was seldom discussed in the literature. These results illustrate that ensemble learning methods can be used as a viable method for sentiment classification.

1. Introduction

With the rapid development of information technologies, user-generated contents can be easily posted online [1]. The sheer volume and exponential growth of this information provide potential value to governments, businesses, and users themselves. For instance, governments can evaluate online citizen-generated texts to assess public sentiment for making policies. Furthermore, many customer-generated reviews of products and services have become valuable sources for market analysis; these reviews are used to set business strategy of E-commerce websites, such as Amazon.com and Epinion.com [50]. Online users can also benefit from reading others' opinions through recommender systems.

There is an inherent property called sentiment involved in the vast majority of online-generated content. Sentiment is an opinion or feeling you have about something [12]. In this study of sentiment classification, we focus on attempts to identify the sentiment polarity of a given text, which is traditionally classified as either positive or negative. Analyzing and predicting the polarity of the sentiment plays an important role in understanding social phenomena and general society trends [6].

Accordingly, sentiment classification has become a popular research topic [1,4,6]. The sentiment classification problem was initially tackled granularly at the levels of document, sentence, clause, phrase, and word, depending on the specific objectives of applications. Heuristic-based methods and machine learning approaches were frequently employed in previous research. Heuristic-based methods were primarily used in conjunction with linguistic characters and semantic features. For example, Turney [38] used mutual information with predefined sentiment words to score other phrase tags, therefore identifying the sentiment of documents. In parallel, many studies focused on using machine learning algorithms to classify sentiment. For instance, Support Vector Machines (SVM) and Naive Bayes (NB) are commonly used to identity sentiment, due to their predictive power. Pang et al. [29] conducted an empirical study in sentiment classification, concluding that SVM outperformed other classifiers such as NB. In recent years, there has been a growing interest in using ensemble learning techniques, which combine the outputs of several base classification techniques to form an integrated output, to enhance classification accuracy [43,48]. However, compared with other research domains, related work about ensemble methods contributing to sentiment classification are still limited and more extensive experimental work is needed in this area.

To fill this research gap, this paper makes a comparative study of the effectiveness of ensemble learning for sentiment classification and demonstrates that three popular ensemble methods (Bagging [5], Boosting [33] and Random Subspace [19]) can be useful. Research

* Corresponding author at: Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. Tel.: +852 9799 0955; fax: +852 2788 8694.
E-mail address: wgedison@gmail.com (G. Wang).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.08.002

Table 1
Selected previous studies in ensemble learning for sentiment analysis.

Study	Year	Feature set	Base learner	Ensemble methods	Dataset
Wilson et al. [45]	2006	N-gram, syntactic features	DT	Boosting	MPQA dataset
Tsutsumi et al. [37]	2007	N-gram	SVM, ME, Scoring	Stacking	Movie review dataset
Abbasi et al. [2]	2008	N-gram, lexicon	SVM	SVRCE	Two web forum datasets
Lu & Tsou [26]	2010	N-gram, lexicon	NB, ME, SVM, Scoring	Stacking	NTCIR opinion dataset
Whitehead & Yaeger et al. [43]	2010	N-gram	SVM	Bagging, Boosting and Random Subspace	Five product review datasets
Xia et al. [48]	2011	POS and word-relation based features	NB, ME, SVM	Stacking	Five product review datasets
Su et al. [34]	2012	N-gram	NB, CB, KNN, ME, SVM	Stacking	Three product review datasets
Li et al. [24]	2012	N-gram, lexicon	SVM, KNN, Scoring	Stacking	Chinese review dataset

in many areas has shown the advantages of ensemble methods both theoretically and empirically [30,51]. In ensemble methods, learners composing an ensemble are usually called base learners. In Bagging, the base learners are constructed using random independent bootstrap replicates from a training dataset, and the final result is calculated by a simple majority vote [5,51]. In Boosting, the base learners are constructed on weighted versions of the training set, which are dependent on previous base learners' results and the final result is calculated by a simple vote or a weighted majority vote [33,51]. In Random Subspace, the base learners are constructed in random subspaces of the feature space [19,51].

We employed ten public sentiment analysis datasets to verify the effectiveness of these three ensemble methods when using five base learners (NB, Maximum Entropy (ME), Decision Tree (DT), K Nearest Neighbors (KNN), and SVM). Based on a total of 1200 comparative group experiments, empirical results show that ensemble learning methods achieve better performances than base learners. Among the three ensemble methods, Random Subspace has the better comparative results except with NB as base learner, although it was seldom discussed in the literature. In addition, RS-SVM had the highest average accuracy in 6 datasets and similar results with other methods in the other 4 datasets. These results illustrate that ensemble learning methods can be used as a viable method for identifying sentiment polarities.

The main contribution of this paper is to verify the effectiveness of using ensemble learning for sentiment classification. The remainder of the paper is organized as follows. In Section 2, we survey the related work about sentiment classification. The details of three different types of ensemble methods are introduced in Section 3. Section 4 presents the design and methodology used in the experiments, while the results are analyzed in Section 5. Section 6 discusses conclusions and future research directions.

2. Literature review

Since the late 1990s, sentiment classification has been a hot research topic in the areas of data mining, information retrieval, and natural language processing [4,28]. Many researchers have investigated sentiment classification from different perspectives. Due to the linguistic characteristics involved, sentiment analysis is done at different levels of text units. A word, phrase, clause, sentence, or document may become the text unit in analysis [28]. In order to capture the sentiment of individual words or phrases, a measure of the strength of sentiment polarity is often defined to quantify how strongly a word or phrase is judged to be positive or negative [9,22,35,38]. Furthermore, Thet et al. [36] computed the sentiment of a clause from individual word sentiment scores, considering the grammatical dependency structure of the clause. Other studies used sentence-level attempts to classify the positive or negative sentiments for each sentence [49,50]. The greatest amount of work has been done on document level polarity categorization [1,4,11,29,43,48]. This is also the focus level of our study. The techniques for sentiment classification in prior research can be classified into heuristic-based methods and machine learning methods.

2.1. Heuristic-based methods for sentiment classification

By means of predefined lexicons and calculation rules, heuristic-based methods generally classify text sentiments based on the total number of derived positive or negative sentiment features [28]. For example, Hatzivassiloglou and McKeown [18] considered that adjectives are more predictive of sentiment classification and predicted the sentiment of adjectives by inspecting them in conjunction with “and,” “or,” “but,” “either/or,” and “neither/nor.” However this approach may overestimate the importance of adjectives and underestimate some

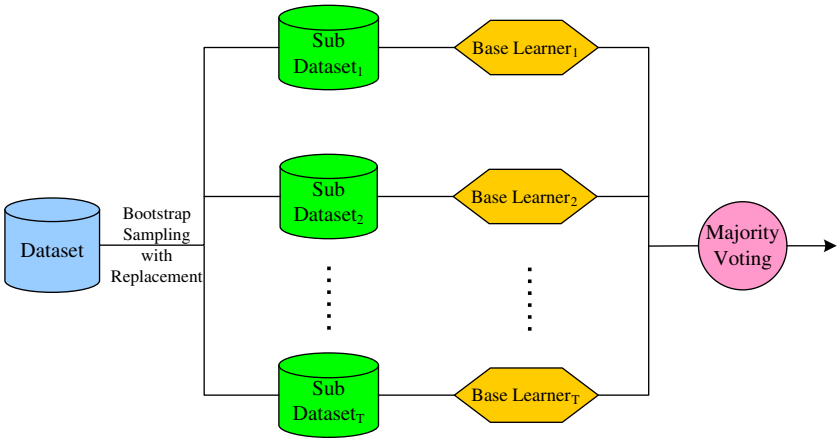

Fig. 1. The Bagging process.