Information Processing and Management 000 (2016) 1–14

Contents lists available at ScienceDirect

Information Processing and Management

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

A sampling based sentiment mining approach for e-commerce applications

G Vinodhini*, RM Chandrasekaran

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar 608002, India

ARTICLE INFO

Article history:
Received 20 February 2015
Revised 24 August 2016
Accepted 25 August 2016
Available online xxx

Keywords:
Sentiment
Opinion
Imbalance
Sampling
Ensemble

ABSTRACT

Emerging technologies in online commerce, mobile and customer experience have transformed the retail industry so as to enable the marketers to boost sales and the customers with the most efficient online shopping. Online reviews significantly influence the purchase decisions of buyers and marketing strategies employed by vendors in e-commerce. However, the vast amount of reviews makes it difficult for the customers to mine sentiments from online reviews. To address this problem, sentiment mining system is needed to organize the online reviews automatically into different sentiment orientation categories (e.g. positive/negative). Due to the imbalanced nature of positive and negative sentiments, the real time sentiment mining is a challenging machine learning task. The main objective of this research work is to investigate the combined effect of machine learning classifiers and sampling methods in sentiment classification under imbalanced data distributions. A modification is proposed in support vector machine based ensemble algorithm which incorporates both oversampling and undersampling to improve the prediction performance. Extensive experimental comparisons are carried out to show the effectiveness of the proposed method with several other classifiers used in terms of receiver operating characteristic curve (ROC), the area under the ROC curve and geometric mean.

1. Introduction

E-commerce has changed the landscape of the retail industry. In recent years, the customers are changing the nature of e-commerce by providing their experience in the form of online reviews. Online review analysis can help retailers in sales prediction and provides a guiding effect on purchase decisions of customers. The abundance of online reviews causes information overload (Xie, Zhou, & Sun, 2012; Goeuriot et al., 2011; Kessler & Nicolov, 2009; Zhang, Yu, & Meng, 2007). Sentiment mining has attracted the focus of the research community from last decade to mine the online reviews to identify the user sentiments. Sentiment mining carried out at the document level or sentence level is useful in many applications (Dang, Zhang, & Chen, 2010; O'Keefe & Koprinska, 2009; Prabowo & Thelwall, 2009;Ye, Zhang, & Law, 2009; Abbasi, Chen, Thoms, & Fu, 2008; Pang, Lee, & Vaithyanathan, 2002). These levels of sentiment mining are not enough for decision making of whether to purchase the product or not. The focus of this research work is on the next fine grain level of analysis, i.e. feature level sentiment mining (Hu & Liu, 2004; Scaffidi et al., 2007).

Also nowadays, imbalanced datasets are pervasive in real world applications. Handling imbalanced datasets become a very interesting research area within machine learning communities. Imbalanced datasets introduce a significant reduction

* Corresponding author.
E-mail address: g.t.vino@gmail.com (G. Vinodhini).

http://dx.doi.org/10.1016/j.ipm.2016.08.003
0306-4573/

Table 1
Summary of literature review.

S.no	Studies	Technique used	Feature	Data source	Performance
1.	Kubat and Matwin (1997)	Undersampling	Domain specific	UCI dataset	G-mean: 0.96
2.	Chawla et al. (2002)	Synthetic minority oversampling technique (SMOTE)	Domain specific	PIMA	AUC:0.73
3.	Guo and Viktor (2004)	Data boost	Domain specific	Real time	AUC:0.830.92
4.	Wang and Yao (2009)	Ensemble	Domain specific	UCI dataset	G-mean: 0.96
5.	Burns et al. (2011)	Undersampling	Linguistic feature	Movie reviews	G-mean: 0.72
6.	Li et al., 2011a)	Co- selection method	Stemmed terms	Product reviews	AUC: 0.89
7.	Li et al. (2012)	Selective sampling	Unigrams	Product reviews	G-mean: 0.78
8.	Wang et al., (2012)	Multi class -ensemble	Domain specific	UCI dataset	G-mean: 0.92

in performance of standard classifiers when they are invoked to learn data underlying concepts (Kim & Hovy, 2004; Zhang et al., 2007). Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. In this research work, attention is paid to the sentiment classification problem of imbalanced datasets (Wang, Sun, Ma, Xu, & Gu, 2014; Su, Zhang, Ji, Wang, & Wu, 2013; Li, Wang, & Chen, 2012; Oza & Tumer, 2008). The classifiers employed in this research work are constructed based on support vector machine (SVM). SVM is the interest in this study for its good classification accuracy reported in many real applications, in the imbalanced data context. The strategies proposed already to address the imbalanced classification (undersampling and oversampling) was not systematically evaluated in the sentiment classification.

In this research, a modification is proposed in the SVM based ensemble algorithm. The proposed approach incorporates both oversampling and undersampling to improve the prediction performance. A comparative study on the effectiveness of the proposed method and sampling strategies using SVM classifier in the context of imbalanced sentiment classification is also conducted. The effectiveness of the experimental comparisons carried out is analyzed in terms of suitable metrics.

2. Related work

In sentiment classification literature, many classification methods were tested only on balanced datasets (Xia, Zong, & Li 2011; Melville, Gryc, & Lawrence, 2009, Ye et al., (2009), Tan & Zhang, 2008). Prior studies on sentiment classification have shown that ensemble methods performed better than single machine learning techniques in balanced data distribution (Li et al., 2012; Tsutsumi, Shimada, & Endo, 2007). There has been little discussion on the effects of learning subjective aspects from imbalanced data, although it is typical of the product domain to have substantially more positive than negative reviews (Vinodhini & Chandrasekaran, 2014a, b; Li et al., 2012; Burns, Bi, Wang, & Anderson, 2011; Wang et al., 2011).

The issue of class imbalance is addressed so far in two different ways in the machine learning literature. The first method is data level handling which is classifier independent. The second method is algorithmic level handling which involves modifying the classifiers. Kubat and Matwin (1997) suggested that a dataset can be balanced by undersampling the majority class without changing the original sample size of the minority class. The major limitation of undersampling approach is that it results in information loss for the majority class. Japkowicz and Stephen (2002) proposed oversampling approach which duplicates the minority class instances. Japkowicz et al., showed that undersampling produces better results than oversampling in many cases. The limitation of oversampling is that it introduced an unnatural bias in favor of the minority class. Chawla, Bowyer, Hall, and Philip Kegelmeyer (2002) introduced using synthetic examples to augment the minority class. It is proved to be better than oversampling with replacement. Though this method creates noise for the classifiers which could result in a reduction of performance, it is widely used to solve the problem of skewed data. In modifying the classifiers to adapt datasets to deal with the imbalanced data problem, various works have been carried out using different kinds of classifiers (Chen, Liaw, & Breiman, 2004; Guo & Viktor, 2004). In terms of SVMs, several attempts have been made to improve their class prediction accuracy. Many researchers proved that SVMs is capable to solve the problem of skewed vector spaces without introducing noise (Akbani, Kwek, & Japkowicz, 2004; Morik, Brockhausen, & Joachims, 1999).

Burns et al. (2011) addressed sentiment classification of imbalanced data. Their experiments do not involve SVM. Li, Wang, Zhou, and Lee (2011a, b) adopted the popular random undersampling method. The major drawback of random undersampling is that it discards potentially useful data that could be important for the learning process. Wang et al., (2011) proposed combining multiple classifiers trained from multiple instances of undersampled data. Table 1 summarizes the literature survey done. Many existing studies in sentiment mining used supervised learning either in document level or sentence level. Only a few studies employed supervised learning in feature level sentiment mining (Dang et al., 2010; O'Keefe & Koprinska, 2009; Prabowo & Thelwall, 2009; Ye et al., 2009; Abbasi et al., 2008; Pang et al., 2002). Most of the studies assume a balanced distribution of positive and negative samples, which may not be true in reality. This motivates us to investigate on feature level imbalanced sentiment classification.

Efficient features need to be extracted for a machine learning algorithm for better sentiment classification. Information gain and mutual information feature selection methods are used to eliminate the noise and irrelevant features from the feature vector. Yatsko (2014) suggested the concept of logarithmic equalizing as a normalizing factor and the idea of zonal correlation analysis of text classification. His approach is for automatic text classification based on analyzing the deviation

Please cite this article as: G. Vinodhini, R. Chandrasekaran, A sampling based sentiment mining approach for e-commerce applications, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.003