Electronic Commerce Research and Applications 16 (2016) 66–76

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews

Feng Tian [a,b,*], Fan Wu [a,b], Kuo-Ming Chao [c], Qinghua Zheng [d], Nazaraf Shah [c], Tian Lan [a,b], Jia Yue [a,b]

[a] Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, China
[b] Shaanxi Key Lab of Satellite-Terrestrial Network Tech. R&D, Xi'an Jiaotong Univ., Xi'an, China
[c] Department of Computer Science and Technology, Coventry University, CV1 2JH, UK
[d] Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form 22 October 2015
Accepted 22 October 2015
Available online 30 October 2015

Keywords:
Classification methods
Imbalanced sample classification
Instance transfer methods
Product reviews
Topic sentence analysis

ABSTRACT

The increasing interest in sentiment classification of product reviews is due to its potential application for improving e-commerce services and quality of the products. However, in realistic e-commerce environments, the review-related data are imbalanced, and this leads to a problem in which minority class information tends to be ignored during the training phase of a classification model. To address this problem, we propose a *topic sentence-based instance transfer method to* process imbalanced Chinese product reviews by using an auxiliary dataset (source dataset). The proposed method incorporates a rule and supervised learning hybrid approach to identify a topic sentence of each product review and adds the feature set of the topic sentence to the feature space of sentiment classification. Next, to measure the transferability of instances in source dataset, a greedy algorithm based on information gain of top-N common features is used to select common features. Then, a common feature-based cosine similarity of instances between source dataset and target dataset is introduced to select the transferable instances. Furthermore, a *synthetic minority over-sampling technique* (Smote) based method is adopted to overcome feature space inconsistency between the source dataset and target dataset. Finally, we immigrate the instances selected in source dataset into target dataset to form a new dataset for the training of classification model. Two datasets collected from Jingdong and Dangdang are the target dataset and source dataset. The experimental results verify that, considering the ability of generalization, our proposed method helps a support vector machine (SVM) to outperform other classification methods, such as the J48, Naive Bayes, Random Forest and Random Committee methods, when applied to datasets produced by resampling and Smote.

1. Introduction

In the last few years, we have witnessed a surge of interest in opinions mining automated systems for online product reviews. There are many representative articles in the large research literature (e.g., Bagheri et al. 2013, Fu et al. 2013, Zhang et al. 2012, Zhang et al. 2014). The major supporting technology for opinion mining systems includes topic modelling and sentiment analysis. Researchers, engineers, and practitioners believe that the systems capable of automatically analyzing consumer sentiment expressed widely in online venues will help companies to understand how the consumers perceive their products and services. Many research efforts on sentiment analysis on product reviews have been carried out to enable companies to understand consumer's perception of the products and services.

Most of them rely on an assumption that the class distribution in the training datasets is balanced. However, in reality, the class distribution in collected product review data is usually imbalanced, so they called *imbalanced data*. The imbalanced data encountered in classification is a well-known problem, especially when the size of majority classes is above three times of the size of minority classes. This leads to a situation where minority class information gets ignored during the training phase of a classification model. A model trained from this kind of dataset that have low identification precision in minority classes exhibit overfitting of the majority class.

Some researchers have employed a sub-sampling strategy on imbalanced data to balance the class distribution of the dataset. This approach worsens the performance and generalization ability of the classification model trained on subsampled dataset. At the

* Corresponding author at: Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, P.R. China.

http://dx.doi.org/10.1016/j.elerap.2015.10.003
1567-4223/

same time, different products (or topics, to relate products to the more formal language of our research) from one data source may have an imbalanced distribution in emotion classes. This may form different feature spaces with diverse data distributions in emotion classification. That is, the imbalanced distribution of emotion classes with different topics represents different kind of interactions and mental states of the users.

The traditional methods for handling imbalanced classification problem rely on data level sampling, cost sensitive learning, features selection, feature weight adjustment and one-class learning approaches (Ogura et al. 2011). However, because these methods normally only rely on one dataset, the classification models that are trained on them have an over-fitting problem and lack the ability for generalization. For example, suppose that a balanced dataset is created from only one dataset according to a sampling strategy for training the classifiers. When a trained classifier is applied to a different real-world dataset for analysis, the classification performance is often degraded (He and Ma 2013).

Methodologies behind the classifiers that are trained on more than one auxiliary dataset have been widely adopted (Nguyen et al. 2011, Tommasi and Tuytelaars 2014, Gong et al. 2012; Heim et al. 2014, Hung and Lin 2011) in recent years in an attempt to address problems with insufficient and homogeneous data by adopting the knowledge transfer learning method (Pan and Yang 2010). A simple method may directly combine an auxiliary dataset and an original dataset into a single dataset to train the classifier. As the tasks of emotion detection are strongly domain and product- or topic-dependent. The feature distribution of each product will have its own characteristics. So we believe that such a method will destroy the innate and unique features that exist in different domains and will decrease recognition accuracy.

We are taking on the task of topic sentence-based instance transfer in this research. Our approach is to sample similar instances from the auxiliary dataset in order to deal with imbalanced sentiment classification of target dataset of product reviews. This can be classified as one of data level sampling approaches.

Fig. 1 illustrates the core idea of this research on instance transfer for providing a solution to the problem of imbalanced sentiment classification of product reviews.

We begin by defining some key language for this research. Suppose there are two datasets: a target dataset (T) and a source dataset (S), and dataset T can have a different number of instances in each class. Further assume that datasets S and T have the same classes of the sentiment analysis. The goal of instance transfer involves the following process.

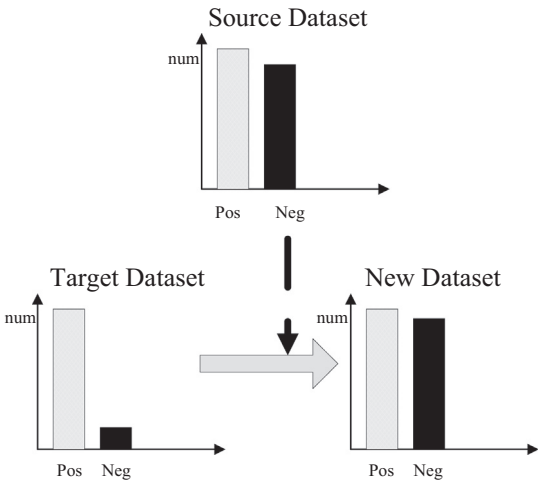

Fig. 1. An instance transfer for imbalanced emotion classification.

In order to achieve the training task of sentiment classification model in T, it chooses the transferable instances of same class from S and transfers them to the corresponding class in dataset T to create a new target dataset D', while it ensures that different classes in dataset D' have a similar data size. This helps to improve the performance of the classification model that is trained on dataset D'. The figure shows that both of datasets T and S have two same classes to be recognized, known as Pos (Positive) and Neg (Negative). After instance transfer, the instances of these classes in new dataset D' have a similar number.

The challenges of implementing this core idea are as follows: (1) how to measure the transferability of instances in S, and (2) how to homogenize the feature space of these instances with that of T. The similarity between feature space $\Omega(F|T)$ in T and feature space $\Omega(F|S)$ in S is adopted to evaluate the transferability of each instance in S. If $\Omega(F|T) = \Omega(F|S)$, then instance transfer becomes a simple task to be solved as they have direct transferability. However, in general, datasets S and T not only have common words in the unigram sets or phases in the bigram set, but also have their own innate and unique words in the unigram set or phases in the bigram set. This leads to the issue of feature space inconsistency between T and S which can be represented as $\Omega(F|T) \neq \Omega(F|S)$.

We use two datasets collected from two famous Chinese e-commerce portals, Jingdong (www.jd.com) and Dangdang (www.dangdang.com), and are named as JingDong and Dangdang, respectively in this research. The feature space of both datasets is one or many types of N-gram features, such as the unigram and bigram of the product reviews corpora. In these two corpora, the products (as topics) of Jingdong only include Laptop and PC, while the topics of Dangdang only includes digital product accessories. The number of items of the unigram and bigram in the feature sets, JingDong and DangDang, are 1385 and 1258, and most of the items are different.

Inspired by the idea of topic sentences, Baxendale (1958) and Paice (1980) provide a strong indication of overall subject in each product review, this research proposes a topic sentence-based instance transfer method for imbalanced emotion classification of Chinese product reviews. The contributions of the proposed approach are as follows:

(1) Introduce a concept topic sentence for each product review. An algorithm for identifying a topic sentence for each product review is proposed based on features of title, first sentence or last sentence of the review
(2) Introduce new feature spaces, based on two feature sets, features of topic sentences and features of the whole body of each review. A *feature set* of a topic sentence includes syntax features and the frequency of emotion words and relevant nouns, as shown in Table 1.
(3) Propose a feature selection strategy for transferable instances, which is a greedy algorithm based on a function of extracting the proportion of sum of the information gain of top-N common features between the T and S datasets. This strategy helps to choose a set of common features, which contribute towards improvement of imbalanced data classification.
(4) Introduce a Smote-based method (Chawla 2003) for processing feature space inconsistency in order to overcome the inconsistency problem between feature spaces of T and the instances transferred from dataset S.
(5) Generate a training dataset by immigrating instances depending on emotion class distribution of both T and S.