Polarity Classification of Imbalanced Microblog Texts

1710250 XIANG YUNMIN

Sentiment analysis is a process to analyze opinion or emotion in texts. Polarity classification is one of the major problems in sentiment analysis. It is a task to classify a given text into negative, positive, or neutral. Many researchers have devoted to studies of the polarity classification. Especially, the polarity classification of texts on microblog such as Twitter is paid much attention, since users actively express their opinion on social media. However, most datasets used in past studies are balanced, in which the number of samples of each class is almost the same. However, the distribution of the polarity of texts is actually imbalanced in real social media, that is the number of neutral samples are much more than other classes. Supervised machine learning usually performs poorly on imbalanced data, since a classifier tends to judge minority samples as majority class. However, detection of minority samples (i.e. positive and negative) is important because they provide useful information for users.

Over-sampling is a technique to train an accurate classifier from an imbalanced data. It increases an amount of minority samples so that the distribution of the classes is well balanced. Among various over-sampling methods, SMOTE and ADASYN are widely used. Supposing that each sample in a dataset is represented as a feature vector, SMOTE synthesizes new minority samples by randomly choosing a vector that is on a line between two existing minority samples in vector space. ADASYN is an extended version of SMOTE. It focuses on the fact that the samples nearby the other classes are difficult to be classified. Therefore, ADASYN generates more synthetic samples from minority samples near the borderline.

The goal of this research is to train an accurate model that can classify polarity of a given text in an imbalanced data set. We focus on the polarity classification of texts in Twitter. We conducted a preliminary survey to reveal the distribution of the polarity in Twitter, and confirmed that 86% of tweets were neutral. It means that training a classier from an imbalance data is an important problem for the polarity classification of tweets. This thesis proposes several methods to extend SMOTE and ADASYN to improve the performance of the polarity classification in microblog.

First, a novel over-sampling method called Amount Control Over-sampling (ACO) is proposed. One of the problems of SMOTE and ADASYN is that the synthetic samples are artificially generated and not real samples at all. The excessive number of synthetic samples may lead the poor performance of the polarity classification. Therefore, we propose ACO to control or optimize

the number of synthetic samples. The basic idea of ACO is to optimize the balance parameter *bal* on the development data. It is defined as the proportion of the minority samples to the majority samples in a new (over-sampled) data set. The balance parameter is optimized on a development data. First, for a given *bal*, the training data is balanced by SMOTE or ADASYN. Next, a classifier is trained on the balanced training data and applied to determine polarity labels of samples in the development data. The above procedure is repeated by changing the value of *bal*. The optimized *bal* is chosen so that F1-measure on the development data become the highest.

A polarity word is a word that expresses positive or negative opinions such as "good" and "bad". Many studies proved that polarity words were effective features for the polarity classification. Therefore, we propose an over-sampling method that considers the importance of polarity words. The core idea of this method is to generate more samples from those samples that include polarity words. A weight parameter named *wp* is defined as the weight of samples including polarity words. More precisely, *wp* is a ratio of the number of synthesized samples generated from a minority sample with polarity words to that from a sample without polarity words. It is optimized on the development data. This method is called Polarity Oriented Over-sampling (POO). In addition, since the computational costs of determining *wp* by using trial and error on a development data is high, we propose a method to automatically determine the parameter *wp*. We measure the intensity of the sentiment expressed in a tweet by calculating average of sentiment scores of the words. We use a sentiment lexicon to get sentiment scores of words. The higher the intensity of the sentiment of a sample is, the greater the parameter *wp* is set. It enable us to synthesize more samples from a minority sample that expresses strong sentiment. This method is called as Polarity Intensity Oriented Over-sampling (PIOO).

Support Vector Machine (SVM) is used to train a polarity classifier in this study. Word embedding is used to obtain a feature vector of a tweet. A weighted sum of word vectors is defined as a feature vector of a tweet. Skip-gram is applied to train word embedding.

Several experiments are conducted to evaluate our methods. An imbalanced data set is constructed by adding neutral tweets to SemEval 2017 data set. First, our proposed ACO is evaluated. F1-measure of SMOTE+ACO is 48.74% and 52.79% for the negative and positive classification, which are 11.35 and 6.27 points better than SMOTE, respectively. F1-measure of ADASYN+ACO is 51.44% and 53.78% for the negative and positive classification, which are 11.07 and 10.04 points better than ADASYN, respectively. These results indicate that ACO, the method to optimize the number of synthesized samples, is effective. Next, our proposed POO is evaluated. F1-

measure of SMOTE+POO is 54.07% and 57.94% for the negative and positive classification, which are 5.33 and 5.15 points better than SMOTE+ACO, respectively. F1-measure of ADASYN+POO is 54.93% and 65.03% for the negative and positive classification, which are 3.49 and 11.25 points better than ADASYN+ACO, respectively. Therefore, POO can contribute to boost the performance of the polarity classification. Finally, our proposed PIOO is evaluated. F1-measure of ADASYN+PIOO is 52.84% and 63.49% for the negative and positive classification, which are 2.09 and 1.34 points worse than ADASYN+POO, respectively.

According to the results of our experiments, PIOO could not contribute to improve the performance. We will explore anther solution to automatically determine $wp$ with less computational costs than POO in future. We also notice that several errors are caused by ignoring the semantic features of sentence. Not only words or word embedding but also semantic relations in a tweet should be used as features for the polarity classification.