Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 690–701
Expert Systems with Applications
www.elsevier.com/locate/eswa

Imbalanced text classification: A term weighting approach

Ying Liu [a,*], Han Tong Loh [b], Aixin Sun [c]

[a] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China
[b] Department of Mechanical Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore 117576, Singapore
[c] School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore

Abstract

The natural distribution of textual data used in text classification is often imbalanced. Categories with fewer examples are under-represented and their classifiers often perform far below satisfactory. We tackle this problem using a simple probability based term weighting scheme to better distinguish documents in minor categories. This new scheme directly utilizes two critical information ratios, i.e. relevance indicators. Such relevance indicators are nicely supported by probability estimates which embody the category membership. Our experimental study using both Support Vector Machines and Naïve Bayes classifiers and extensive comparison with other classic weighting schemes over two benchmarking data sets, including Reuters-21578, shows significant improvement for minor categories, while the performance for major categories are not jeopardized. Our approach has suggested a simple and effective solution to boost the performance of text classification over skewed data sets.

Keywords: Text classification; Imbalanced data; Term weighting scheme

1. Introduction

1.1. Motivation

Learning from imbalanced data has emerged as a new challenge to the machine learning (ML), data mining (DM) and text mining (TM) communities. Two recent workshops in 2000 (Japkowicz, 2000) and 2003 (Chawla, Japkowicz, & Kolcz, 2003) at AAAI and ICML conferences, respectively and a special issue in ACM SIGKDD explorations (Chawla, Japkowicz, & Kolcz, 2004) were dedicated to this topic. It has been witnessing growing interest and attention among researchers and practitioners seeking solutions in handling imbalanced data. An excellent review of the state-of-the-art is given by Weiss (2004).

The data imbalance problem often occurs in classification and clustering scenarios when a portion of the classes possesses many more examples than others. As pointed out by Chawla et al. (2004) when standard classification algorithms are applied to such skewed data, they tend to be overwhelmed by the major categories and ignore the minor ones. There are two main reasons why the uneven cases happen. One is due to the intrinsic nature of such events, e.g. credit fraud, cancer detection, network intrusion, and earthquake prediction (Chawla et al., 2004). These are rare events presented as a unique category but only occupy a very small portion of the entire example space. The other reason is due to the expense of collecting learning examples and legal or privacy reasons. In our previous study of building a manufacturing centered technical paper corpus (Liu & Loh, 2007), due to the costly efforts demanded for human labeling and diverse interests in the papers, we ended up naturally with a skewed collection.

Automatic text classification (TC) has recently witnessed a booming interest, due to the increased availability of documents in digital form and the ensuing need to organize them (Sebastiani, 2002). In TC tasks, given that most test collections are composed of documents belonging to multiple classes, the performance is usually reported in terms of micro-averaged and macro-averaged scores (Sebastiani,

* Corresponding author. Tel.: +852 34003782.
E-mail address: mfyliu@polyu.edu.hk (Y. Liu).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.10.042

2002; Yang & Liu, 1999). Macro-averaging gives equal weights to the scores generated from each individual category. In comparison, micro-averaging tends to be dominated by the categories with more positive training instances. Due to the fact that many of these test corpora used in TC are either naturally skewed or artificially imbalanced especially in the binary and so called "one-against-all" settings, classifiers often perform far less than satisfactorily for minor categories (Lewis, Yang, Rose, & Li, 2004; Sebastiani, 2002; Yang & Liu, 1999). Therefore, micro-averaging mostly yields much better results than macro-averaging does.

1.2. Related work

There have been several strategies in handling imbalanced data sets in TC. Here, we only focus on the approaches adopted in TC and group them based on their primary intent. The first approach is based on sampling strategy. Yang (1996) has tested two sampling methods, i.e. proportion-enforced sampling and completeness-driven sampling. Her empirical study using the ExpNet system shows that a global sampling strategy which favors common categories over rare categories is critical for the success of TC based on a statistical learning approach. Without such a global control, the global optimal performance will be compromised and the learning efficiency can be substantially decreased. Nickerson, Japkowicz, and Milios (2001) provide a guided sampling approach based on a clustering algorithm called Principal Direction Divisive Partitioning to deal with the between-class imbalance problem. It has shown improvement over existing methods of equalizing class imbalances, especially when there is a large between-class imbalance together with severe imbalance in the relative densities of the subcomponents of each class. Liu's recent efforts (Liu, 2004) in testing different sampling strategies, i.e. under-sampling and over-sampling, and several classification algorithms, i.e. Naïve Bayes, k-Nearest Neighbors (kNN) and Support Vector Machines (SVMs), improve the understanding of interactions among sampling method, classifier and performance measurement.

The second major effort emphasizes cost sensitive learning (Dietterich, Margineantu, Provost, & Turney, 2000; Elkan, 2001; Weiss & Provost, 2003). In many real scenarios like risk management and medical diagnosis, making wrong decisions are usually associated with very different costs. A wrong prediction of the nonexistence of cancer, i.e. false negative, may lead to death, while the wrong prediction of cancer existence, i.e. false positive, only results in unnecessary anxiety and extra medical tests. In view of this, assigning different cost factors to false negatives and false positives will lead to better performance with respect to positive (rare) classes (Chawla et al., 2004). Brank, Grobelnik, Milic-Frayling, and Mladenic (2003) have reported their work on cost sensitive learning using SVMs on TC. They obtain better results with methods that directly modify the score threshold. They further propose a method based on the conditional class distributions for SVM scores that works well when only very few training examples are available.

The recognition based approach, i.e. one-class learning, has provided another class of solutions (Japkowicz, Myers, & Gluck, 1995). One-class learning aims to create the decision model based on the examples of the target category alone, which is different from the typical discriminative approach, i.e. the two classes setting. Manevitz and Yousef (2002) have applied one-class SVMs on TC. Raskutti and Kowalczyk (2004) claim that one-class learning is particularly helpful when data are extremely skewed and composed of many irrelevant features and very high dimensionality.

Feature selection is often considered an important step in reducing the high dimensionality of the feature space in TC and many other problems in image processing and bioinformatics. However, its unique contribution in identifying the most salient features to boost the performance of minor categories has not been stressed until some recent work (Mladenic & Grobelnik, 1999). Yang and Pedersen (1997) has given a detailed evaluation of several feature selection schemes. We noted the marked difference between micro-averaged and macro-averaged values due to the poor performances over rare categories. Forman (2003) has done a very comprehensive study of various schemes for TC on a wide range of commonly used test corpora. He has recommended the best pair among different combinations of selection schemes and evaluation measures. The recent efforts from Zheng, Wu, and Srihari (2004) advance the understanding of feature selection in TC. They show the merits and great potential of explicitly combining positive and negative features in a nearly optimal fashion according to the imbalanced data.

Some recent work simply adapting existing machine learning techniques and not even directly targeting the issue of class imbalance have shown great potential with respect to the data imbalance problem. Castillo and Serrano (2004) and Fan, Yu, and Wang (2004) have reported the success using an ensemble approach, e.g. voting and boosting, to handle skewed data distribution. Challenged by real industry data with a huge number of records and an extremely skewed data distribution, Fan's work shows that the ensemble approach is capable of improving the performance on rare classes. In their approaches, a set of weak classifiers using various learning algorithms are built up over minor categories. The final decision is reached based on the combination of outcomes from different classifiers. Another promising approach which receives less attention falls into the category of semi-supervised learning or weakly supervised learning (Blum & Mitchell, 1998; Ghani, 2002; Goldman & Zhou, 2000; Lewis & Gale, 1994; Liu, Dai, Li, Lee, & Yu, 2003; Nigam, 2001; Yu, Zhai, & Han, 2003; Zelikovitz & Hirsh, 2000). The basic idea is to identify more positive examples from a large amount of unknown data. These approaches are especially