Computers and Electrical Engineering 61 (2017) 67–79

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Cost-sensitive learning for social network analysis and network measurement☆

Zhang Xing[a], Meili Wang[b], Zhang Yang[b,*], Ning Jifeng[b]

[a] College of Mechanical and Electronic Engineering, Northwest A&F University, Yang Ling 712100, PR China
[b] College of Information Engineering, Northwest A&F University, Yang Ling 712100, PR China

ARTICLE INFO

Article history:
Received 31 October 2016
Revised 21 May 2017
Accepted 31 May 2017

Keywords:
Social network
Network measurement
N-dependence estimators
Cost-sensitive learning

ABSTRACT

Recently, the application of data mining techniques to social network analysis and network measurement has received considerable attention. In this work, an aggregating N-dependence estimator (ANDE)-based cost-sensitive classification algorithm (CS_ANDE) was proposed for use with the unbalanced data commonly observed in social networks and network measurements. First, a one-dependence estimator was adopted to determine the approximate cost of misclassification. Second, multiple classifiers were constructed to minimize the misclassification cost. Subsequently, these classifiers were used to re-label samples. Ultimately, a CS_AODE classifier was obtained by learning these re-labeled samples. Consistent with two-dependence estimators, a CS_A2DE classifier was acquired. The CS_AODE and CS_A2DE classifiers were empirically evaluated against MetaCost and AODE for UCI datasets, and the results indicated that CS_AODE and CS_A2DE significantly outperformed the other classifiers and that the performance was stable under different parameters.

1. Introduction

With the emergence of social networks and network measurement capabilities, the interest in and reliance on network information has increased. Social and other network datasets have continued to grow every day, thus requiring automated information processing for their analysis. Interestingly, data mining techniques require massive data to build a best-fit model. Thus, data mining is the perfect tool to analyze social networking sites.

During social network and network measurement data mining, the following situations are frequently encountered. Often, only a small number of samples will be significant when a large quantity of information samples are processed. For example, in topic detection, the overwhelming majority of samples might be unrelated, with only an extremely small number of related samples reflecting the topic of interest. For instance, a training sample set might contain 9800 unrelated samples and 200 related samples. Therefore, even if all 200 related samples are predicted as unrelated samples, the classification accuracy of the corresponding classifier might be as high as 98%. Clearly, such a classifier is illogical. Data with uncoordinated quantities and weights are referred to as unbalanced data. To process such data, which are commonly observed in social networks and network measurements, a cost-sensitive classification algorithm is presented.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Zhihan Lu.

* Corresponding author.
E-mail addresses: zhangxing@nwsuaf.edu.cn (Z. Xing), zhangyang@nwsuaf.edu.cn, zhangyang_ce@126.com (Z. Yang).

http://dx.doi.org/10.1016/j.compeleceng.2017.05.032
0045-7906/

When using the cost-sensitive classification algorithm, different classification errors can be endowed with different costs; therefore, the classification result is robust to samples with a small quantity and a high weight. A Bayes-based algorithm is commonly used in cost-sensitive learning and effective when the Naive Bayes' attribute independence assumption can be established. However, in many practical applications, this assumption is not valid. To relax the Naive Bayes' attribute independence, an aggregating N-dependence estimator (ANDE)-based cost-sensitive classification algorithm (CS_ANDE) is proposed. First, the computing pattern of misclassification costs is redefined, and an approximate value of the misclassification cost is obtained by relaxing the Bayesian assumption derived from the ANDE method. Second, multiple classifiers are constructed by misclassification cost minimization and then employed to re-label samples. Finally, a cost-sensitive classifier is obtained by learning the re-labeled samples. Relevant experiments show that this algorithm can be used to effectively reduce misclassification costs and process unbalanced data more meaningfully.

The paper is organized as follows: Section 2 provides a review of the related work, while Section 3 presents the problem definition of this paper. Section 4 explains the theoretical structure of the cost-sensitive N-dependence estimators and the CS_ANDE algorithm. Next, Section 5 discusses the experiments, and Section 6 presents the conclusions.

2. Related work

2.1. Cost-sensitive learning

Traditional machine learning techniques are designed to minimize error rates. Nevertheless, different errors generate different outcomes and losses in practice. In addition, classifiers oriented to classification error rate minimization can only reduce the number of errors rather than the total loss. In consideration of this challenge, a cost-sensitive algorithm has been developed and applied for medical diagnoses, network intrusion detection and software development [1].

Cost can be broadly interpreted and may refer to monetary costs or time loss. In a detailed analysis, Turney identified nine types of cost [2]. Owing to the variety of possible costs, cost-sensitive learning involves multiple subproblems because these costs always occur simultaneously and are correlated. Therefore, the problem of cost-sensitive learning becomes especially complex. For example, in medical diagnostics, the cost to acquire a certain test result must be considered in conjunction with the cost of possible misclassification [3].

Among the various subproblems associated with cost-sensitive learning, the misclassification cost is the most common in practice. Thus, it is a topic of great interest and the focus of this paper. Many approaches have been proposed, including the implementation of direct adaptations of accuracy-based methods and the use of genetic algorithms, anytime methods, and boosting and bagging techniques [4].

Chan proposed a cost-based sampling method [5]. Specifically, the proportions accounted for by diverse classifications in the classifier are changed so that the classifier can incorporate cost sensitivity. Chan successfully applied this method in a bank credit setting, thus providing a good example of the application of a cost-sensitive method to practical problems.

Elkan argued that the essence of cost-sensitive problems is classification expectation minimization [6] and proposed a training set in which the proportions observed in real-life examples are adjusted to acquire a cost-sensitive algorithm. In addition, Elkan also presented the Elkan theorem to ensure the validity of the corresponding sampling method. However, repeated training samples can lead to an increased risk of overfitting in the sampling process. In response, Zadrozny offered an alternative costing approach [7]. In this approach, instead of directly repeating samples in the training dataset, samples are initially collected from the training dataset and reserved according to a certain proportion.

MetaCost is a representative boosting algorithm. The basic principle underlying this method is the adoption of a Bayesian risk minimization theory to construct multiple classifiers for every template in the training dataset [8]. The class with the minimum cost at the time of classification is defined as an optimal tag, which is then used to re-label the corresponding real-life example. Finally, the re-labeled training set is learned and a minimum cost-oriented cost-sensitive algorithm is obtained.

Ting introduced a boosting technique-based decision tree algorithm C4.5cs [9] in which different samples are endowed with different weights in direct proportions to their costs. Subsequently, these data with weights are used to train a C4.5 classifier. A similar cost-sensitive Naive Bayesian algorithm based on a boosting technique is also presented in [10].

2.2. Methods to relax the independence assumption

A Bayesian algorithm is a simple and effective classification method. Nevertheless, it is only effective when the Naive Bayesian model assumption can be established. In most practical applications, this assumption cannot be justified. Accordingly, researchers have proposed various methods to relax the Naive Bayesian model assumption. Despite low efficiency, LBR [11] and SP-TAN [12] both exhibit favorable classification results. Although the former requires a long classification period, the latter requires a long training time.

In [13], Webb proposed an adjusted probability Naive Bayesian induction, which adds a simple extension to the Naive Bayesian classifier. To relax the Naive Bayesian model assumption, Webb [14] proposed the design of an AODE with a given class and a special super parent attribute, and it included the assumption that other attributes are independent. By the selecting different super parent nodes, this algorithm can be adapted to determine their mean value to reduce the variance. On this basis, Webb also proposed the ANDE algorithm in 2012 [15]. For averaged n dependence estimators, each model