Knowledge-Based Systems 115 (2017) 87–99

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Entropy-based fuzzy support vector machine for imbalanced datasets

Qi Fan[a,c], Zhe Wang[a,c,*], Dongdong Li[a], Daqi Gao[a,*], Hongyuan Zha[b]

[a] Department of Computer Science & Engineering, East China University of Science & Technology, Shanghai, 200237, P.R. China
[b] School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, Atlanta, Georgia
[c] Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, PR China

ARTICLE INFO

Article history:
Received 19 April 2016
Revised 20 September 2016
Accepted 21 September 2016
Available online 15 October 2016

Keywords:
Information entropy
Fuzzy support vector machine
Imbalanced dataset
Pattern recognition

ABSTRACT

Imbalanced problem occurs when the size of the positive class is much smaller than that of the negative one. Positive class usually refers to the main interest of the classification task. Although conventional Support Vector Machine (SVM) results in relatively robust classification performance on imbalanced datasets, it treats all samples with the same importance leading to the decision surface biasing toward the negative class. To overcome this inherent drawback, Fuzzy SVM (FSVM) is proposed by applying fuzzy membership to training samples such that different samples provide different contributions to the classifier. However, how to evaluate an appropriate fuzzy membership is the main issue to FSVM. In this paper, we propose a novel fuzzy membership evaluation which determines the fuzzy membership based on the class certainty of samples. That is, the samples with higher class certainty are assigned to larger fuzzy memberships. As the entropy is utilized to measure the class certainty, the fuzzy membership evaluation is named as entropy-based fuzzy membership evaluation. Therefore, the Entropy-based FSVM (EFSVM) is proposed by using the entropy-based fuzzy membership. EFSVM can pay more attention to the samples with higher class certainty, i.e. enhancing the importance of samples with high class certainty. Meanwhile, EFSVM guarantees the importance of the positive class by assigning positive samples to relatively large fuzzy memberships. The contributions of this work are: (1) proposing a novel entropy-based fuzzy membership evaluation method which enhances the importance of certainty samples, (2) guaranteeing the importance of the positive samples to result in a more flexible decision surface. Experiments on imbalanced datasets validate that EFSV outperforms the compared algorithms.

1. Introduction

In recent years, imbalanced problem has become one of the challenges in data mining community [21,32,49]. In imbalanced problem, the negative samples outnumber the amount of positive class ones [17]. However, most standard classification algorithms are proposed based on the balanced class distributions or equal misclassification costs [8]. When facing to the imbalanced problem, these algorithms fail to properly represent the data distributive characteristics, and result in the unfavorable performance [24]. Support Vector Machine (SVM) known as the representative kernel-based learning paradigm, can provide relatively robust classification performance on imbalanced datasets [27]. By utilizing support vectors, SVM maximizes the separation margin between the support vectors and the decision boundary, meanwhile minimizes the total classification error [46]. The effects of imbalanced datasets on SVM exploit inadequacies of the margin maximization paradigm [1,37]. Since SVM treats all samples with the same importance, it ignores the differences between the positive and negative classes, which results in the learned decision surface biasing toward the negative class. To overcome the inherent drawback of SVM, Lin et al. [29] propose the Fuzzy SVM (FSVM) which applies fuzzy membership to each input sample and reformulate SVM such that different input samples have different contributions to the learning of decision surface.

To determine the fuzzy membership function is the key point in FSVM. Lin et al. [29] propose a method that based on class centers. Although it is easy to use, it assigns smaller fuzzy memberships to support vectors, which might decrease the effects of support vectors on the construction of classification surface. Meanwhile, it has some disadvantages in dealing with imbalanced datasets, e.g. the misclassification accuracy of positive class is higher than that of the negative class. In order to address the disadvantages of FSVM, Tian et al. [43] introduce a novel fuzzy membership determining function and propose a new FSVM based on non-equilibrium data which effectively reduces the misclassification accuracy of the

* Corresponding authors.
E-mail addresses: wangzhe@ecust.edu.cn (Z. Wang), gaodaqi@ecust.edu.cn (D. Gao).

http://dx.doi.org/10.1016/j.knosys.2016.09.032
0950-7051/

positive class. In literature [12], Chen et al. raise a fuzzy rough SVM, which improves the traditional hard margin SVM by considering the membership of every training sample in constraints. Zhang et al. [50] introduce a rough margin based SVM to overcome the over-fitting problem of traditional SVM. Moreover, Wang et al. [48] propose a Bilateral-weighted FSVM (B-FSVM) in credit scoring area. In credit scoring areas, financial intermediaries usually cannot label one customer as absolutely good who is sure to repay in time, or absolutely bad who will default certainly, so B-FSVM treats every sample as both positive and negative classes, but with different memberships. Then, Chaudhuri et al. [10] apply FSVM to solve bankruptcy prediction problem. Batuwita et al. [4] propose a method to improve FSVM for class imbalance learning, which is used to handle the class imbalanced problem in the presence of outliers and noise.

In information theory, the entropy is an effective measure of certainty. Shannon et al. [39] defined the entropy as the negative logarithmic function of probability of the occurrence of an event. Several researchers [40] have employed Shannon entropy for image processing and pattern recognition. Moreover, Renyi [38] generalizes Shannon entropy by introducing a parameter called the power of probability which controls the shape of the probability distribution. Recently, to address the issues of certainty of a neighborhood system and extend the traditional accuracy and roughness measures to deal with neighborhood systems, Chen et al. [13] introduce the concept of neighborhood entropy to evaluate the certainty of a neighborhood information system. In literature [31], Long et al. introduce the entropy concept to the allocation of an unknown pixel to a certain predefined class to develop a robust effective multi-spectral image classification algorithm. Meanwhile, Hanmandlu [23] designs a new entropy function to improve the results on the infra-red thermal face recognition.

In this paper, we propose a novel fuzzy membership evaluation which assigns the fuzzy membership of each sample based on its class certainty. The class certainty, in this paper, demonstrates the certainty of the sample classified to a specific class. Due to the entropy is an effective certainty-measuring approach, we adopt it to evaluate the class certainty of each sample. In doing so, the entropy-based fuzzy membership evaluation is proposed by determining the fuzzy membership of the training samples based on their corresponding entropy. With the entropy-based fuzzy membership evaluation, the Entropy-based FSVM (EFSVM) has the robust capacity of dealing with the imbalanced datasets. In practice, as the importance of the positive class is higher than that of the negative one in imbalanced phenomenon, the classifier should pay more attention to the positive samples than the negative ones. Thus, the positive samples are assigned to the relatively large fuzzy memberships to guarantee their importance. While, the fuzzy memberships of negative samples are determined by the entropy-based fuzzy membership evaluation method, which is based on the criterion that the samples with lower class certainty are more insensitive to the noise, and easily mislead the decision surface, thus their importance should be weakened in the learning process. After evaluating the fuzzy membership of all training samples, EFSVM is adopted to classify the imbalanced datasets.

The contributions of this paper can be highlighted as follows:

- This paper proposes a novel entropy-based fuzzy membership evaluation, which adopts the entropy to evaluate the class certainty of a sample, and determines the corresponding fuzzy membership based on the class certainty. In doing so, the learning algorithm can pay more attention to the samples with higher class certainty to result in more robust decision surface.
- This paper adopts a simple criterion to guarantee the importance of positive class. The criterion assigns the relatively large fuzzy memberships to positive sample, which results in the decision surface paying more attention to the positive class so as to increase the generalization of the learned classifier.

The remainder of this paper is structured as follows. Section 2 gives a brief introduction on the related work of imbalanced problems. In Section 3, we firstly introduce the developed entropy-based fuzzy membership evaluation approach, then give the detailed demonstration on our proposed entropy-based fuzzy support vector machine. In Section 4, several experiments on both synthetic and real-world imbalanced datasets are conducted to validate the effectiveness of our proposal. Following that, the concluding remarks are presented in Section 5.

2. Related work

In this section, we briefly introduce the related work on the problem of classification with imbalance datasets. A dataset is said to be imbalance when there is a significant difference between the number of samples belonging to different classes [41]. Generally, the class with an abundant number of samples is marked as negative class. While, the class with few number of samples is named as positive class. In recent years, the imbalanced problems have achieved much attention because that it presents in lots of area such as e-mail foldering [14,51], fault diagnosis [16,47], detection of oil spills [22], medical diagnosis [34], and sensor data analysing [36]. In these problems, the positive samples are usually more important than the negative ones for learning robust classifiers. Moreover, the positive samples entail high costs when their identification are not properly performed [19]. The difference in the class distribution of imbalanced problem poses a major challenge to traditional algorithms because most of learning algorithms are optimized to achieve the minimization of the objective functions which do not consider the differences between different classes. Therefore, the positive samples are usually neglected during the classifier learning process. In the paper, the Imbalance Ratio (IR) [33] is the main measure of evaluating the difficult level of a specific imbalance dataset. IR is defined as the ratio of the number of negative samples and the positive ones, i.e. $IR = n_{neg}/n_{pos}$ where n_{neg} and n_{pos} are the number of negative samples and positive samples, respectively.

In order to deal with the imbalance problems, numerous techniques have been proposed, which can be separated into tree groups: data level approaches [18], algorithm level approaches [32] and the combination data and algorithm level approaches which is also known as cost-sensitive approaches [11]. Data level approaches modify the original training set to obtain a relatively balanced dataset which can be used to standard learning algorithms. Generally, the data level approaches are divided into three groups, i.e. over-sampling methods, under-sampling methods and the hybrid methods. Over-sampling methods [9] balance the datasets by adding new positive samples. The easiest over-sampling method is Random Over-Sampling (ROS) [3] which randomly replicates positive samples from the original dataset until its IR is near to one. One of the most popular over-sampling method is Synthetic Minority Over-sampling TEchnique (SMOTE) [5] which creates synthetic positive samples by a pre-processing before the classifier learning process. In SMOTE, for each positive sample, a new synthetic sample is generated on the line joining it to one of the k nearest positive samples. Under-sampling methods [20] try to balance the datasets by deleting several samples from the negative class. The simplest under-sampling method is Random Under-Sampling (RUS) [3] which randomly removes the negative samples from the original dataset to achieve a balanced datasets. One of the popular over-sampling method is One-Sided Selection (OSS) [28]. In OSS, all positive samples are preserved, and the negative samples are under-sampled based on the nearest neighbor