Pattern Recognition Letters 88 (2017) 72–80

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Entropy-based matrix learning machine for imbalanced data sets

Changming Zhu [a,*], Zhe Wang [b]

[a] College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, PR China
[b] College of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, PR China

ARTICLE INFO

Article history:
Received 30 November 2015
Available online 26 January 2017

Keywords:
Entropy
Fuzzy membership
Imbalanced data set
Pattern recognition

ABSTRACT

Imbalance problem occurs when negative class contains many more patterns than that of positive class. Since conventional Support Vector Machine (SVM) and Neural Networks (NN) have been proven not to effectively handle imbalanced data, some improved learning machines including Fuzzy SVM (FSVM) have been proposed. FSVM applies a fuzzy membership to each training pattern such that different patterns can give different contributions to the learning machine. However, how to evaluate fuzzy membership becomes the key point to FSVM. Moreover, these learning machines present disadvantages to process matrix patterns. In order to process matrix patterns and to tackle the imbalance problem, this paper proposes an entropy-based matrix learning machine for imbalanced data sets, adopting the Matrix-pattern-oriented Ho–Kashyap learning machine with regularization learning (MatMHKS) as the base classifier. The new leaning machine is named EMatMHKS and its contributions are: (1) proposing a new entropy-based fuzzy membership evaluation approach which enhances the importance of patterns, (2) guaranteeing the importance of positive patterns and get a more flexible decision surface. Experiments on real-world imbalanced data sets validate that EMatMHKS outperforms compared learning machines.

1. Introduction

Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data [1]. For the recognition of patterns, many approaches are proposed and some fields are developed as well. Among them, matrix learning and imbalance problem are two hot spots of present research.

• Matrix learning is developed to process the disadvantages of vector learning. We know that it is necessary to choose an appropriate representation for patterns in terms of pattern classification [1]. In statistical pattern classification, a pattern is generally represented by a point in a d-dimensional space [1,2]. Such a representation is viewed as vector representation and can bring a convenience in mathematics, for example, Ho–Kashyap learning machine (HK) [2] and the modified HK learning machine (MHKS) [3]. Patterns with vector representation are called vector patterns. The learning machine design which is based on vector patterns is called vector-pattern-oriented learning machine (VecC) or vector learning machine. The procedure of designing a VecC is called vector learning. When vector learning machines process patterns with matrix representation, these patterns have to be vectorized. Patterns with matrix representation are called matrix patterns and images are classical matrix patterns. However, such a procedure brings three potential problems [4–6]. First one is the loss of some implicit structural or local contextual information of these matrix patterns. Second one is the requirement of a large memory. Third one is the high risk of overtraining. To solve these problems, matrix-pattern-oriented learning machine (MatC), i.e., matrix learning machine, has been developed. The procedure of designing a MatC is called matrix learning. The matrix learning machine is based on matrix patterns. Moreover, a vector or matrix pattern can be reshaped to a new one by some certain reshaping ways [7]. It has been demonstrated that there are two advantages for the learning machine design based on matrix patterns without a vectorization preprocessing. One is reducing the computational complexity and the other is improving the classification performance [5,6,8,9]. Based on the above advantages, some researchers have developed matrix-pattern-oriented Ho–Kashyap learning machine with regularization learning (MatMHKS) [7], new least squares support vector classification based on matrix patterns (MatLSSVC) [10], and one-class support vector machines based on matrix patterns (OCSVM) [11].

• In many real-world classification problems, such as e-mail foldering [12], fault diagnosis [13], detection of oil spills [14], and medical diagnosis [15], we can always divide a data set into two classes, one is positive class and the other is negative class. When negative class contains many more patterns than that of

* Corresponding author.
E-mail addresses: zcm19880301@163.com, cmzhu@shmtu.edu.cn (C. Zhu), wangzhe@ecust.edu.cn (Z. Wang).

http://dx.doi.org/10.1016/j.patrec.2017.01.014
0167-8655/

the positive class, imbalance problem occurs. Since most standard classification leaning machines including Support Vector Machine (SVM) and Neural Networks (NN) are proposed with the assumption of the balanced class distributions or equal misclassification costs [16], they fail to properly represent the distributive characteristics of patterns and result in the unfavorable performance when they are adopted to process imbalanced data. In order to overcome such a disadvantage, Fuzzy SVM (FSVM) [17], Bilateral-weighted FSVM (B-FSVM) [18], and FSVM-based Class Imbalance Learning (FSVM-CIL) [19] are proposed. FSVM applies a fuzzy membership to each input pattern and reformulates SVM such that different input patterns can give different contributions to the learning of decision surface. B-FSVM treats every pattern as both positive and negative classes, but with different memberships, so we cannot say one pattern belongs to one class absolutely. But for both of them, how to determine the fuzzy membership function is the key point. Then for FSVM-CIL, it adopts more functions $f(x)$ to determine the fuzzy memberships. Besides those learning machines, some learning machines are also proposed to process imbalance problems [20–24].

In this paper, we propose a learning machine which can process matrix patterns and imbalance problem. First, in order to process matrix patterns, we adopt MatMHKS as a basic. Then in order to process imbalance problem, we adopt the notion of FSVM-CIL, namely, adopts a function which generates a value between 0 and 1 so as to reflect the importance of a pattern in its own class and applies a fuzzy membership to each input pattern. Furthermore, for the fuzzy membership, we propose a new fuzzy membership evaluation approach which assigns the fuzzy membership of each pattern based on its class certainty that denotes the probability of a pattern belonging a certain class. Due to the entropy is an effective measure of certainty, we adopt the entropy to evaluate the class certainty of each pattern. In doing so, the entropy-based fuzzy membership evaluation approach is proposed. This approach determines the fuzzy membership of training patterns based on their corresponding entropies. With such an entropy-based fuzzy membership evaluation and MatMHKS, this paper proposes a new learning machine named EMatMHKS.

The rest of this paper is organized as below. In Section 2, we want to give a brief review about the classification of imbalanced data. In Section 3, we introduce the proposed entropy-based fuzzy membership evaluation approach, then give the details of EMatMHKS. In Section 4, several experiments on real-world imbalanced data sets are conducted to validate the effectiveness of EMatMHKS. Following that, the conclusions and future work are given in Section 5.

2. Review about the classification of imbalanced data

There are many reviews for the classification of imbalanced data [25–27]. According to these reviews, if a data set contains many more patterns from one class than from the rest of classes, we call it an imbalanced data set and the corresponding problem is named imbalance problem. Moreover, for a data set, at least one class which is called the minority class possesses only a small number of training patterns while other classes make up the majority. In the imbalance problems, the classifiers always have high performances on the majority class while low performances on the minority classes. Because compared with the minority classes, the influence of majority class on traditional training criteria is larger and most classifiers are designed for the objective of minimizing the error rate which represents the percentage of the incorrect prediction of class labels. Indeed, these classifiers ignore the difference between types of misclassification errors. In particular, they implicitly assume that all misclassification errors cost equally [25].

As these reviews said that in order to tackle with the imbalance problems, a number of solutions are proposed both at the data and algorithmic levels. For the data level, one widely used solution is resampling the original data set and this solution also has two ways, one is oversampling the minority class and the other is undersampling the majority class. These two ways are always used in the preprocessing phase. In terms of oversampling the minority class, Chawla has proposed Synthetic Minority Over-sampling Technique (SMOTE) [28] and some modifications including SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) and SMOTE-N (Synthetic Minority Over-sampling Technique Nominal). Then some scholars have proposed a multiple resampling method [29], a cluster-based over-sampling method [30], borderline-SMOTE1, borderline-SMOTE2 [31], Adaptive Synthetic Sampling Technique (ADASYN) [32], and Cluster Based Synthetic Oversampling (CBSO) [33]. All of these methods can increase the size of the minority class. In terms of undersampling the majority class, it uses a subset of the majority class to train the classifier. Since many majority class patterns are ignored, the training set becomes more balanced and the training process becomes faster. There are some widely used methods for undersampling the majority class, for example, hybrid sampling technique [34]. However, there is a main disadvantage of undersampling that some potentially usefully information which are contained in the ignored patterns is neglected. For the algorithmic level, solutions include adjusting the costs of the various classes, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning [25]. Among these solutions, cost-sensitive learning is a widely used approach. Some methods are proposed to realize the cost-sensitive learning [35–38]. Besides the cost-sensitive learning, some modifications of SVM can also be used to process imbalance problems, for example, FSVM, B-FSVM, FSVM-CIL, and Modified Proximal SVM (MPSVM) [39].

In order to evaluate the performances of the methods which are used to process imbalance problems, some metrics of quality are used [40,41]. Some scholars have concluded three kinds of evaluation methods, one is numerical value performance measure, another is graphical performance analysis with probabilistic classifiers, and the third is complex numerical evaluation measures. In terms of first kind, four terms are used, i.e., Accuracy, True Positive Rate(acc^+), True Negative Rate(acc^-), and Positive Predictive Value (PPV). These four terms are defined as below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$acc^+ = \frac{TP}{TP + FN} = Recall^+ = Sensitivity \tag{2}$$

$$acc^- = \frac{TN}{TN + FP} = Recall^- = Specificity \tag{3}$$

$$PPV = \frac{TP}{TP + FP} = Precision \tag{4}$$

where TP, FP, FN, and TN represent True Positive, False Positive, False Negative, and True Negative respectively [40]. In terms of second kind, lift chart, Receiver Operating Characteristic (ROC) curve, recall-precision curve, and cost curve are widely used. In terms of third kind, five terms are used, i.e., F-Measure [42], G-Mean [43], Youden's Index [44], Likehoods [45], Discriminatory Power [46]. Especially, F-Measure and G-Mean are widely used and they are defined as below.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

$$G - Mean = \sqrt{acc^+ \times acc^-} \tag{6}$$