Knowledge-Based Systems 70 (2014) 361–375

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Improved pseudo nearest neighbor classification

Jianping Gou [a,*], Yongzhao Zhan [a], Yunbo Rao [b], Xiangjun Shen [a], Xiaoming Wang [c], Wu He [d]

[a] School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, PR China
[b] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, PR China
[c] School of Mathematics and Computer Engineering, Xihua University, Chengdu, Sichuan 610039, PR China
[d] Digital Media College, Sichuan Normal University, Chengdu, Sichuan 610068, PR China

A R T I C L E I N F O

Article history:
Received 24 May 2013
Received in revised form 26 June 2014
Accepted 23 July 2014
Available online 31 July 2014

Keywords:
k-Nearest neighbor rule
Pseudo nearest neighbor rule
Local mean vector
Pattern classification
Local mean-based pseudo nearest neighbor rule

A B S T R A C T

k-Nearest neighbor (KNN) rule is a very simple and powerful classification algorithm. In this article, we propose a new KNN-based classifier, called the local mean-based pseudo nearest neighbor (LMPNN) rule. It is motivated by the local mean-based *k*-nearest neighbor (LMKNN) rule and the pseudo nearest neighbor (PNN) rule, with the aim of improving the classification performance. In the proposed LMPNN, the *k* nearest neighbors from each class are searched as the class prototypes, and then the local mean vectors of the neighbors are yielded. Subsequently, we attempt to find the local mean-based pseudo nearest neighbor per class by employing the categorical *k* local mean vectors, and classify the unknown query patten according to the distances between the query and the pseudo nearest neighbors. To assess the classification performance of the proposed LMPNN, it is compared with the competing classifiers, such as LMKNN and PNN, in terms of the classification error on thirty-two real UCI data sets, four artificial data sets and three image data sets. The comprehensively experimental results suggest that the proposed LMPNN classifier is a promising algorithm in pattern recognition.

1. Introduction

k-Nearest neighbor (KNN) rule [1], as one of the top 10 algorithms in data mining [2], has been extensively studied and widely used in the field of pattern classification for many years owing to its superiorities. It has been well known that KNN has the conceptual and implementational simplicity for classification. In addition to the advantage, the most famous characteristic of KNN is that it has the asymptotically optimal performance in the Bayes sense when the number N of training samples and the neighborhood size k both tend to infinity with the constraint $k/N \to 0$ [1,3]. Thus, KNN-based classifiers often achieve good classification performance in many practical applications.

Although the standard KNN classifier has the benefits, there still exists some major drawbacks as follows [2,4]. One prominent issue is the sensitiveness to the choice of the neighborhood size k. In fact, this shortcoming is produced by the existing outliers contained in the k-neighborhood. To address this issue, an adaptive selection of the neighborhood size based on statistical confidence is developed in [5]. The second issue is the approach to combine the class labels in the k-neighborhood for making classification decision. As we know, a simple majority vote is employed in KNN, but this can be a problem when the nearest neighbors vary widely in their distances and the closer neighbors more reliably predict the class of the query point. Aiming at this problem, several weighted voting methods for KNN in [4,6,7] have been developed with more weights for the closer neighbors. The third issue is how to assign weights to nearest neighbors. Generally speaking, more weights can be given to highly reliable nearest neighbors while further reducing the influence of unreliable ones. However, the common ways that more weights are assigned to the closer nearest neighbors according to their distances are not absolutely correct, as some farther neighbors may have more importance for classification. With regard to this problem, adaptive metric nearest neighbor classification is proposed in [8–10]. In practice, these three issues aforementioned together exist in many KNN-based pattern classification algorithms recently.

To deal with the issues in KNN-based classification, a number of variations of the KNN-based approaches have also been developed. Based on the fuzzy sets theory, the fuzzy nearest neighbor classifiers that introduce fuzziness to KNN have been proposed in [11–13]. As an extension of fuzzy sets theory, some evidence-theoretic KNN classifiers have been reported in [12,14–16] from the point of

* Corresponding author.
E-mail address: cherish.gjp@gmail.com (J. Gou).

http://dx.doi.org/10.1016/j.knosys.2014.07.020
0950-7051/

view of Dempster–Shafer theory. In the evidence-theoretic KNN methods, each neighbor of a query point is viewed as an item of evidence that supports certain hypotheses regarding class membership of that pattern. Recently, Derrac et al. fully review the most relevant algorithms to the fuzzy nearest neighbor classification [17]. Instead of using all the training samples by some KNN-based algorithms, there are many prototype-based nearest neighbor classifiers, such as prototype selection [18–21], prototype generation [22–24] and prototype optimization [25–27]. These approaches use a few well represented prototypes to achieve the good classification performance in both speed, storage and accuracy. In addition to the two kinds of nearest neighbor classification, there are some approximate nearest neighbor classifiers [28,29]. Recently, some other latest KNN-based methods are proposed, such as the coarse to fine KNN (CFKNN) rule [30] and linear reconstruction measure steered nearest neighbor classification (LRMNNC) [31].

Up to now, the classification performance of many KNN-based methods are still impacted by the problems above, particulary in the small sample size cases with the existing outliers [32] and the curse of dimensionality [26]. To overcome the negative effect of the outliers, a simple nonparametric classifier named the local mean-based k-nearest neighbor (LMKNN) rule has been developed in [33,34]. In the LMKNN rule, the local mean vector of k nearest neighbors from each class is employed to classify the query pattern. Since the first introduction of the LMKNN rule, its basic idea has already been successfully borrowed to many new methods [35–41]. Among these algorithms, pseudo nearest neighbor (PNN) rule [35] is another promising KNN-based classifier, which is based on the distance weighted k-nearest neighbor (WKNN) rule [7] and the LMKNN rule. Instead of the original nearest neighbor, PNN first attempts to seek the pseudo nearest neighbor in each class, and then classifies the class of the closest pseudo nearest neighbor to the query point. From the extensively experimental results reported in [33,35], it has been proven that the LMKNN and PNN classifiers are more robust to the outliers than the classical KNN classifier with better classification performance.

Motivated by the ideas of the LMKNN and PNN rules, we propose a local mean-based pseudo nearest neighbor (LMPNN) rule in this work. As an extension of the PNN rule [35], the purpose of our LMPNN is to further overcome the existing problems in KNN, so as to well improve the classification performance. In the LMPNN rule, unlike the PNN, we make use of the k local mean vectors of k nearest neighbors in each class instead of themselves to find the pseudo nearest neighbor, which is named as local mean-based pseudo nearest neighbor here. In contrast with k nearest neighbors, the corresponding local mean vectors can represent more class information for classification. Then, the class label of the closest local mean-based pseudo nearest neighbor is allocated to the query pattern. To verify the classification performance of the proposed LMPNN, we conduct extensive experiments on thirty-two real UCI data sets, four artificial data sets and three image data sets, compared to the competing approaches: LMKNN, PNN, CFKNN, WKNN and KNN. The experimental results demonstrate that our proposed LMPNN is an effective and robust classifier with satisfactory classification performance.

The main contributions of this paper can be summarized as follows:

- A new scheme of designing pseudo nearest neighbor of a query pattern is introduced by using k local mean vectors corresponding to k nearest neighbors from each class. The newly designed pseudo nearest neighbor can capture more class information for classification.
- A local mean-based pseudo nearest neighbor rule is proposed by the newly designed pseudo nearest neighbor for each class. It can well integrate the superiorities of LMKNN and PNN.
- The effectiveness and robustness of LMPNN is explored by extensive experiments in terms of the classification error and the good classification of LMPNN is credibly demonstrated by the nonparametric statical tests.

The organization of the rest of this article is as follows. In Section 2, we briefly give the outline of the related classifiers. In Section 3, we introduce the proposed LMPNN method. In Section 4, we conduct experiments for all the competing methods on the real UCI data sets, artificial data sets and image databases. In Section 5, we discuss the computational complexities of the competing classifiers. Finally, the conclusions are made in Section 6.

2. Related classifiers

In pattern classification, the k-nearest neighbor (KNN) rule has been one of the most widely used and extensively studied nonparametric classifiers. In recent years, there exists many variants of KNN, such as the local mean-based KNN (LMKNN) rule [33] and the pseudo nearest neighbor (PNN) rule [35]. In this section, we briefly review both LMKNN and PNN classifiers which motivate our work.

2.1. The LMKNN rule

As an extension of KNN, the local mean-based k-nearest neighbor (LMKNN) rule is a simple and robust nonparametric algorithm in pattern classification. Mitani and Hamamoto have empirically proven that LMKNN can well overcome the existing outliers, particularly in the small sample size cases [33,34]. The basic idea of LMKNN is that the local mean vector of k nearest neighbors of a test point from each class in training set is adopted to determine its class label when making classification decision.

In the general classification problem, suppose a training set $T = \{x_n \in \mathbb{R}^d\}_{n=1}^N$ with M classes consists of N training samples in d-dimensional feature space, and the class label of one sample x_n is c_n, where $c_n \in \{\omega_1, \omega_2, \ldots, \omega_M\}$. Let $T_{\omega_i} = \left\{x_j^i \in \mathbb{R}^d\right\}_{j=1}^{N_i}$ denote a class subset of T from the class ω_i, with the number of the training samples N_i. Given a query point x, the LMKNN rule is carried out as follows:

(1) Find k nearest neighbors from the set T_{ω_i} of each class ω_i for the unknown query point x, and let $T_{\omega_i}^k(x) = \left\{x_j^i \in \mathbb{R}^d\right\}_{j=1}^k$ indicate the set of k nearest neighbors for x from the class ω_i. The distance between x and the neighbor x_j^i is measured by the Euclidean distance metric, i.e., Eq. (1)

$$d\left(x, x_j^i\right) = \sqrt{\left(x - x_j^i\right)^T \left(x - x_j^i\right)}. \tag{1}$$

(2) Compute the local mean vector $u_{\omega_i}^k$ for x from the class ω_i, using the set $T_{\omega_i}^k(x)$

$$u_{\omega_i}^k = \frac{1}{k}\sum_{j=1}^k x_j^i. \tag{2}$$

(3) Calculate the distance $d\left(x, u_{\omega_i}^k\right)$ between the test point x and the local mean vector $u_{\omega_i}^k$ for the class ω_i with Eq. (1).

(4) Assign x into the class c if the distance between the local mean vector for c and the test point x is minimum among M classes.