Neurocomputing 143 (2014) 164–169

Contents lists available at ScienceDirect

Neurocomputing

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

A new nearest neighbor classifier via fusing neighborhood information

Yaojin Lin [a,*], Jinjin Li [a,b], Menglei Lin [a], Jinkun Chen [b]

[a] School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China
[b] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, PR China

ARTICLE INFO

Article history:
Received 16 September 2013
Received in revised form
21 April 2014
Accepted 6 June 2014
Available online 17 June 2014

Keywords:
Nearest neighbor classifier
Neighborhood information
Distance metric
Rank aggregation

ABSTRACT

The nearest neighbor (NN) classification is a classical and yet effective technique in machine learning and data mining communities. However, its performance depends crucially on the distance function used to compute distance between samples. In this paper, we first define the concept of sample's neighborhood and present two related criteria according to neighborhood influence. Then, the influence of sample's neighborhood is comprehensively considered when computing the distances between the query and training samples. Finally, we propose an improved nearest neighbor classification algorithm via fusing neighborhood information. The proposed method can precisely characterize the distance among samples as well as enhance the predictive power of classifier to some extent. The experimental results show that the proposed algorithm basically outperforms classical nearest neighbor classifier and some other state-of-the-art classification methods.

1. Introduction

The classical *k*-nearest neighbor (*k* NN) classification algorithm [11] has gained increasing awareness as one of the 10 most important data mining algorithms developed in the 20th century [30]. It is successfully used for both classification problems and function approximation [14,18]. The classification accuracy of *k* NN algorithm significantly depends on the way that distances are computed. Distance functions are used to measure the similarity between the query and training samples for identifying the first *k* nearest neighbors of the query sample.

With increasing awareness on nearest neighbor classification learning in the literature [3,6,8,15,21,33], many distance functions have been proposed to improve the performance of nearest neighbor classifier, which can be further divided into global and local distance metric learning. For the global distance metric learning, many distance metrics are proposed to process heterogeneous feature sets, typical examples of this kind include Heterogeneous Euclidean-Overlap Metric (HEOM) function, Value Difference Metric (VDM), Heterogeneous Value Difference Metric (HVDM) and Interpolated Value Difference Metric (IVDM) [27,29]. In addition, the others are estimated with weighted similarity between samples. Hu et al. [12] proposed an approach to learn sample weights for enlarging margin by using a gradient descent algorithm to minimize margin based classification loss. Goldberger et al. [9] proposed a method via a stochastic variant of the leave-one-out 1-NN score on the training data to learn a Mahalanobis distance measure. Wang et al. [28] developed an extremely simple adaptive distance measure that assigns different weights to each sample. For the local distance metric learning, many methods adopt locally adaptive distances, rather than the Euclidean distance in the traditional *k* NN, to tackle the global optimality problem. Typical examples of this kind include the adaptive distance in ADAMENN, the Mahanalobis distance in LMNN, the weight adjusted metric in WA*k* NN, the discriminant adaptive metric in DANN and the informative metric in I*k* NN [20,25,26]. These local algorithms only consider neighboring pairwise constraints and avoid adopting those conflicting constraints.

All aforementioned methods are all trying to learn an approximate single metric on all data samples. The drawbacks of learning a single metric include (1) a single metric is likely inappropriate for all training samples with different neighborhood information; (2) a single local metric is sensitive to noisy samples; and (3) a single global metric cannot directly deal with the multimodal distribution problem. Therefore, it is reasonable to learn distance metric via incorporating multiple metrics for the query sample.

In fact, neighborhoods and neighborhood relations are a class of important concepts in topology. Lin [16] pointed out that neighborhood spaces are more general topological spaces than equivalence spaces. It is a powerful tool to data mining, pattern classification and reasoning with uncertainty [5,9,13,20,31]. In this paper, we will briefly review some notations of distance metrics and *k* NN algorithm. And then we will propose a method to improve the nearest neighbor classifier by involving sample's neighborhood information. This approach computes the distance between samples via considering not only the individual distance

* Corresponding author.
E-mail addresses: yjlin@mail.hfut.edu.cn (Y. Lin), jinjinli@mnnu.edu.cn (J. Li), menglei36@126.com (M. Lin), cjk99@163.com (J. Chen).

http://dx.doi.org/10.1016/j.neucom.2014.06.009
0925-2312/

between samples but also the neighborhood information of each sample. Furthermore, we obtain the first k nearest neighbors of a query sample by fusing the above two distances. The proposed method can more precisely characterize the distance between samples as well as enhancing the predictive power of the proposed classifier to some extent. Finally, some experimental analyses are extensively conducted on UCI datasets.

The rest of this paper is organized as follows. Section 2 introduces the basic concept of neighborhood and k NN algorithm. In Section 3, we propose our metric to calculate the nearest neighbors of a query sample, and design our related algorithm. Numerical experiments are provided in Section 4. Finally, conclusions are given in Section 5.

2. Related work

k NN is one of the effective and famous algorithms for function approximation and data classification [30], many methods have been proposed to improve the predictive power of k NN classifiers. In general, the performance of k NN classifiers depends on three factors: the sample size, the selection of distance metric and the value of k. Notably, the sample size is a restrictive factor for k NN classifiers. As we know, if the available samples are of high-volume, k NN classifiers require computing all the distances between the query sample and the training sample, and it is time and memory consuming. Therefore, many algorithms are developed to reduce the size of dataset, such as condensed nearest neighbor classifier [1], orthogonal search tree nearest neighbor classifier [19], and outlier elimination based nearest neighbor classifier [24]. Appropriate measures of distance very strongly affect the accuracy of k NN classifiers. Therefore, many distance functions, such as HEOM, VDM, IVDM and HVDM, are developed to improve the predictive power of k NN classifiers. However, it is considered as a challenging problem to define an accurate distance measure for a given k NN classification task. Besides the above two factors, the final factor which should be taken of is the value of k (the number of nearest neighbors), on the basis of which category of the query sample is determined. We could set the value of k as a predefined value or select it automatically. For example, Hu et al. [12] introduced a neighborhood rough set model as a uniform framework to understand and implement neighborhood classifiers. Bhattacharya et al. [4] used an affinity function for distance measure on the basis of $k=\sqrt{N}$, where N is the number of data used for training purpose. Guo et al. [10] constructed a k NN model which replaces the data to serve as the basis of classification, and the value of k is determined in terms of classification accuracy.

3. Preliminaries

In this section, we will briefly review some kinds of distance metrics and the classical k NN learning algorithm.

3.1. Distance metric

A distance metric is a distance function on a set of points, mapping pairs of points into the non-negative real number. Generally, there are three metric functions widely used. Considering two samples x and y in an N-dimensional space $A=\{a_1,a_2,\ldots,a_N\}$, $f(x,a_i)$ is the value of x in the ith feature a_i, then a general metric, named the Minkowsky distance, is defined as

$$\Delta_p(x,y)=\left(\sum_{i=1}^{N}|f(x,a_i)-f(y,a_i)|^p\right)^{1/p}, \tag{1}$$

where (1) it is called Manhattan distance Δ_1, if $p=1$; (2) Euclidean distance Δ_2, if $p=2$; and (3) Chebychev distance Δ_∞, if $p=\infty$. A detailed survey about distance functions can be seen in [27,29].

In order to deal with heterogeneous features, there are a number of distance functions for mixed numerical and nominal data [27,29], such as Heterogeneous Euclidean-Overlap Metric (HEOM) function, Value Difference Metric (VDM), Heterogeneous Value Difference Metric (HVDM) and Interpolated Value Difference Metric (IVDM). HEOM is defined as

$$\mathrm{HEOM}(x,y)=\sqrt{\sum_{i=1}^{N} w_{a_i}\times d_{a_i}^2(x_{a_i},y_{a_i})}, \tag{2}$$

where N is the number of features, w_{a_i} is the weight of feature a_i, $d_{a_i}(x,y)$ is the distance between samples x and y with respect to feature a_i, defined as

$$d_{a_i}(x,y)=\begin{cases}1 & \text{if the feature value of } x \text{ or } y \text{ is unknown,}\\ overlap_{a_i}(x,y) & \text{if } a_i \text{ is a nominal feature,}\\ rn\_diff_{a_i}(x,y) & \text{if } a_i \text{ is a numerical feature,}\end{cases}$$

where $overlap_{a_i}(x,y)=0$, if $x=y$, otherwise $overlap_{a_i}(x,y)=1$, and $rn\_diff_{a_i}(x,y)=|x-y|/(\max_{a_i}-\min_{a_i})$.

Likewise, a simplified version (without the weighting schemes) of the VDM is defined as

$$\mathrm{VDM}(x,y)=\sum_{i=1}^{N} d_{a_i}(x_{a_i},y_{a_i}). \tag{3}$$

For $\forall a_i\in A$, $d_{a_i}(x_{a_i},y_{a_i})$ is defined as

$$d_{a_i}(x_{a_i},y_{a_i})=(P(x_{a_i})-P(y_{a_i}))^2, \tag{4}$$

where $P(x_{a_i})$ is the probability of object x on feature a_i and $P(y_{a_i})$ is the probability of object y on feature a_i.

3.2. k NN

k NN is one of the famous algorithms for function approximation and multivariate data classification [11,30]. Given a dataset D consisting of n pairs of sample and label, i.e., $D=\{(x_1,c_1),(x_2,c_2),\ldots,(x_n,c_n)\}$. The classification task is to determine the mapping function f, on which the class labels of unclassified samples can be predicted precisely. The nearest neighbor (1-NN) rule, first proposed by Fix and Hodges, is one of the simple and yet effective pattern classification algorithms [11].

Let $N_k(x)$ be a set of k nearest neighbors of x, the label c of x is determined by its k nearest neighbors with a majority of voting strategies, that is

$$c=\underset{c_i}{\operatorname{argmax}}\sum_{c_i\in C}\sum_{x_j\in N_k(x)} I(c_j=c_i), \tag{5}$$

where c_j is the class label of x_j and $I(\cdot)$ is an indicate function. If c_j is the same as c_i, $I(c_j=c_i)=1$; otherwise, $I(c_j=c_i)=0$.

The performance of k NN classification heavily depends on the way that distances are computed, and its decision boundary is very sharp and it is sensitive to noise. To alleviate these problems, many effective solutions are to extend the nearest neighbor to k nearest neighbors (k NNs) [4,17,32].

4. Fusing neighborhood information to the nearest neighbor classifier

4.1. Neighborhood and its influence

In essence, the k nearest neighbor classification algorithm identifies the category of an unknown sample by computing the distance among samples, and not including sample's neighborhood information. Fig. 1 shows an example of sample distance in