Expert Systems with Applications 42 (2015) 906–912

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A density and connectivity based decision rule for pattern classification

Tülin İnkaya *

Uludağ University, Industrial Engineering Department, Görükle, 16059 Bursa, Turkey

ARTICLE INFO

Article history:
Available online 29 August 2014

Keywords:
Classification
Nearest neighbor
Gabriel Graph
Density
Connectivity

ABSTRACT

In this paper we propose a novel neighborhood classifier, *Surrounding Influence Region* (SIR) decision rule. Traditional Nearest Neighbor (NN) classifier is a distance-based method, and it classifies a sample using a predefined number of neighbors. In this study neighbors of a sample are determined using not only the distance, but also the connectivity and density information. One of the well-known proximity graphs, Gabriel Graph, is used for this purpose. The neighborhood is unique for each sample. SIR decision rule is a parameter-free approach. Our experiments with artificial and real data sets show that the performance of the SIR decision rule is superior to the *k*-NN and Gabriel Graph neighbor (GGN) classifiers in most of the data sets.

1. Introduction

Classification problem includes the prediction of the class label of a sample. When it is difficult to estimate the probability density distribution of a data set, non-parametric approaches are used for classification. *k*-Nearest neighbor (*k*-NN) decision rule (Cover & Hart, 1967; Duda & Hart, 1973) is one of the most widely used non-parametric classifiers. It is a distance-based method that classifies a sample according to the labels of its *k* closest neighbors. Although it is simple and effective, *k*-NN decision rule has two main drawbacks: (1) neighborhood definition is solely based on the distance information. How the nearest neighbors are distributed around the sample point is not addressed. (2) Classification accuracy is sensitive to parameter *k*, which should be provided a priori.

Several variants of *k*-NN decision rule are proposed in order to overcome these drawbacks. For a comprehensive survey about the methods for improving the accuracy of *k*-NN, one can refer to Jiang, Cai, Wang, and Jiang (2007).

A stream of research aims to improve *k*-NN by selecting a suitable neighborhood size (Guo, Wang, Bell, Bi, & Greer, 2003; Jiang, Zhang, & Cai, 2006; Xie, Hsu, Liu, & Lee, 2002). Xie et al. (2002) propose a selective neighborhood based Naïve Bayes algorithm for lazy classification. In the algorithm, first, multiple classifiers on multiple neighborhoods with different radius are constructed. Then, the classifier with the highest accuracy is selected for the classification of the samples. Jiang et al. (2006) combine eager learning with lazy learning. First, best value of *k* is learned to fit the training data set. Second, classification is performed using a local Naïve Bayes with the best value of *k*. These approaches focus on learning the best value of *k* for the entire data set. However, the local characteristics around a sample, i.e. density and connectivity relations, are not addressed.

Guo et al. (2003) attempt to incorporate the density concept into the neighborhood definition, and propose the *k*-NN model. They extract the largest neighborhoods with the maximum number of points covered with the same class label. If a new sample falls into a neighborhood, it is assigned to the corresponding class. Otherwise, it is assigned to the class having the minimum distance between the sample and the closest boundary. *k*-NN model generates spherical shaped neighborhoods. This gives rise to classification errors, when the data set includes classes with noise, density differences and arbitrary shapes.

Another research stream is motivated by the effective use of proximity graphs (Toussaint, 2002) in pattern recognition. With the help of the proximity graphs such as Gabriel Graph (GG) and Relative Neighborhood Graph (RNG), the connectivity relations are incorporated into the neighborhood definition (Devroye, Gyorfi, & Lugosi, 1996; Jaromczyk & Toussaint, 1992; Sanchez, Pla, & Ferri, 1997a). Devroye et al. (1996) propose a decision rule in which a sample is classified according to the majority of the votes among its Gabriel Graph (GG) neighbors. Sanchez et al. (1997a) extend this idea for other proximity graphs such as Relative Neighborhood Graph. Incorporating connectivity provides some improvement in the data sets with density differences, but a sample may still have neighbor mixes from other classes due to the connected graph property. Proximity graphs are also used for prototype selection, which focuses on the selection of a sufficiently small subset of prototypes and the elimination of erroneous

* Tel.: +90 224 2942605.
E-mail address: tinkaya@uludag.edu.tr

http://dx.doi.org/10.1016/j.eswa.2014.08.027
0957-4174/

prototypes from the original data set (Bhattacharya, Mukherjee, & Toussaint, 2005; Sanchez, Pla, & Ferri, 1997b; Toussaint & Berzan, 2012).

In k-NN it is assumed that k neighbors have equal influence in the voting. In an alternative scheme, different weights are allocated to the k neighbors based on their distances to the new sample to improve the classification accuracy (Dudani, 1976; Hattori & Takahashi, 1999; Parvin, Alizadeh, & Minaei, 2008; Zeng, Yang, & Zhou, 2009; Zhou & Chen, 2006). There are also other variants of k-NN such as weight allocation based on ranking (Bagui, Bagui, & Pal, 2003) and clustering the neighbors (Zhou, Li, & Xia, 2009). A shortcoming of these approaches is the sensitivity of the classification accuracy to the selection of weights.

To address the limitations of the aforementioned methods, we aim to develop a parameter-free neighborhood classifier. For this purpose, first, we extract a sufficiently large neighborhood around each sample point using the distance, density and connectivity relations. We conceive of two important properties for the neighborhood concept: (1) the neighbors should be close to the sample. (2) The neighbors should lie homogeneously around the sample (Sanchez et al., 1997a). We use GG to define the density-based connectivity, and construct a unique neighborhood for each sample. This neighborhood is called *Surrounding Influence Region* (SIR). Second, we propose a novel neighborhood classifier based on SIR. We demonstrate the performance of the proposed approach in the artificial and real data sets.

A density-based connectivity scheme is also adopted in DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), which is a well-known density-based clustering algorithm. In DBSCAN the neighborhood of a point is defined as the hypersphere with a given radius, and the density is defined as the number of points falling in this neighborhood. Clusters are generated using the points with the neighborhoods that satisfy a given density threshold. Hence, the radius of the neighborhood and the density threshold are two important parameters that affect the density-based connectivity relations among the points. These two parameters define a single density region, so this approach gives rise to neighborhood mixes in the data sets with varying densities. Different from DBSCAN, SIR is a parameter-free approach, and it avoids generalizations about the data set. Instead, it extracts the density and connectivity relations in an adaptive manner. Moreover, the neighborhood of a point is not restricted to a hypersphere. This helps determine a unique neighborhood for each point. In addition, it can handle the data sets with varying density.

To sum up, the contributions of our study are fourfold: (1) the neighbors of a sample are determined using not only the distance, but also the connectivity and density relations. (2) The neighborhood of a sample is uniquely determined. (3) There is no generalization about the local characteristics and neighborhoods in a data set. (4) The proposed neighborhood classifier is a parameter-free approach.

The rest of the paper is organized as follows. We provide the related work and the shortcomings of k-NN and previous approaches in Section 2. We introduce the novel neighborhood decision rule in Section 3. We compare the proposed decision rule with the competing decision rules, and provide the experimental results in Section 4. Finally, we conclude in Section 5.

2. Related work

2.1. The k-NN decision rule and Gabriel Graph neighbor decision rule

Given a set of prototypes D, a sample x and parameter k, the k-NN decision rule (Cover & Hart, 1967; Duda & Hart, 1973) is as follows:

- Determine k nearest neighbors of sample x, $\mathrm{KNN}_x = \{a_1, \ldots, a_k\}$.
- Assign sample x to the class with the majority of votes in KNN_x (resolve ties randomly).

The k-NN decision rule is sensitive to the value of k. Moreover, the size of the training set affects the performance of the k-NN.

Gabriel Graph neighbor (GGN) decision rule (Devroye et al., 1996; Sanchez et al., 1997a) is derived from GG to overcome some of the limitations of the k-NN. It considers the prototypes that are around and relatively close to the sample.

Let D be the set of prototypes, p be a sample and d_{xp} be the Euclidean distance between sample x and prototype p in R^d. Prototype p is a *GG neighbor* of sample x if $d_{xp} \leqslant min_z\{\sqrt{d_{xz}^2 + d_{zp}^2} : z \in D\}$. In other words, prototype p is a *GG neighbor* of sample x if no other point lies inside the hypersphere centered at their middle point and whose diameter is the distance between them. Then, GGN decision rule becomes as follows:

- Determine GG neighbors of sample x, $\mathrm{GGN}_x = \{a_1, \ldots, a_m\}$ where $m \leqslant |D|$.
- Assign sample x to the class with the majority of votes in GGN_x (resolve ties randomly).

2.2. Shortcomings of k-NN and GGN decision rules

We consider a 2-dimensional example data set. The data set includes 82 prototypes and four classes as shown in Fig. 1(a).

The correct class for samples A and B is Class 1. Fig. 1(b) and (c) present the neighborhoods generated by 1-NN and 3-NN, respectively. Both decision rules misclassify sample A whereas they assign sample B to the correct class. In Fig. 1(d) GG neighbors result in a tie between the Classes 1 and 4 for sample A. If the tie is broken randomly, the probability of classifying sample A correctly is 0.50. GGN decision rule misclassifies sample B in Fig. 1(d). Both samples are located in the boundary of Classes 1 and 4. Sample A is close to both classes. Taking into account the distance relations is not sufficient. The only way to classify sample A correctly is by exploration of the connectivity of neighbors and the density change between two classes. Sample B is a misclassification example when only proximity relations are considered. Sample B can also be classified correctly by embedding the density information.

3. The SIR decision rule

3.1. Definitions

In this section we provide the definitions used in the SIR decision rule. Let D be the set of data points, and d_{pq} be the Euclidean distance between points p and q in R^d.

Definition 1. Points p and q are *directly connected* by an edge of the GG if and only if $d_{pq} \leqslant \min_z\{\sqrt{d_{pz}^2 + d_{zq}^2} : z \in D\}$, or equivalently, $B(s, d_{pq}/2) \cap D = \varnothing$; where s is the midpoint on the line connecting points p and q, and $B(s, r)$ denotes the set of points included in an open hyperball centered at point s with radius r, i.e. $B(s, r) = \{z: d_{sz} < r, s \neq z\}$.

Definition 2. Points p and q are *indirectly connected* if the hyperball centered at their midpoint with diameter d_{pq}, $B(s, d_{pq}/2)$, contains at least one other point of D in its interior.

Indirect connection implies that there exists at least one path between the two points whose maximum edge length is shorter than d_{pq}.