Pattern Recognition Letters 80 (2016) 30–36

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Natural neighbor: A self-adaptive neighborhood method without parameter K☆

Qingsheng Zhu*, Ji Feng, Jinlong Huang

Chongqing Key Lab. of Software Theory and Technology, College of Computer Science, Chongqing University, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 27 November 2015
Available online 25 May 2016

Keywords:
Nearest neighbor
Natural neighbor method
Classification
Outlier detection

ABSTRACT

K-nearest neighbor (KNN) and reverse k-nearest neighbor (RkNN) are two bases of many well-established and high-performance pattern-recognition techniques, but both of them are vulnerable to their parameter choice. Essentially, the challenge is to detect the neighborhood of various data sets, while utterly ignorant of the data characteristic. In this paper, a novel concept in terms of nearest neighbor is proposed and named natural neighbor (NaN). In contrast to KNN and RkNN, it is a scale-free neighbor, and it can reflect a better data characteristics. This article discusses the theoretical model and applications of natural neighbor in a different field, and we demonstrate the improvement of the proposed neighborhood on both synthetic and real-world data sets.

1. Introduction

Over the last decade, nearest neighbor method has received considerable attention from the community of data mining and pattern recognition [1–3]. Traditionally, the definition of neighborhood plays an important role, a reasonable definition of neighborhood can effectually improve the performance. In spite of their simplicity, the k-nearest neighbors (KNN) and reverse k-nearest neighbor (RkNN) are demonstrated themselves to be the most useful and effective algorithms. And then a fundamental task that arises in KNN and RkNN is to determinate the optimum value of the parameter *k*.

The problem of choosing the optimum value of *k* is one of the best studied problem in the area of nearest neighbor method since its birth. The best choice of *k* depends upon the data. Generally, larger values of *k* reduce the effect of noises on the classification, but make boundaries between classes less distinct. If the neighborhood is too large with respect to folds in manifold on which the data points lie, large values of *k* may cause the short-circuit errors. Alternatively, small values of *k* reduce the correlation of neighborhood, or separate the data points from the same class.

In fact, determination of parameter *k* is dependent on knowledge of researchers experience and lots of experiments. In order to solve this problem, a new term called Natural Neighbor (NaN) is presented in this paper. NaN method is inspired by the friendship of human society and could be regarded as belonging to the category of scale free nearest neighbor method. The proposed method makes three key contributions to the current state:

1. Natural neighbor method can create an applicable neighborhood graph based on the local characteristics of various data sets. This neighborhood graph can identify the basic clusters in the data set, especially manifold clusters and noises.
2. This method can provide a numeric result named Natural Neighbor Eigenvalue(NaNE) to replace the parameter *k* in traditional KNN method, and the number of NaNE is dynamically chosen for different data sets.
3. The natural neighbor number of each point is flexible, and this value is a dynamic number ranging from 0 to NaNE. The center point of the cluster has more neighbors, and the neighbor number of noise is equal to 0.

2. Related work

2.1. K-nearest neighbor (KNN) method

The concept of k-nearest neighbor (KNN) is a foundation scientific issue in various fields of application study. Its colorful history begins in 1951 with the pioneering work of Stevens [4], who point out that one point and its nearest neighbor can be considered as a subset, and gave an efficient algorithm for the general version of the problem.

Definition 1 (Nearest Neighbor Search)**.** Give a set *X* of points and a query point *q*, the Nearest Neighbor Search problem is to find a

☆ This paper has been recommended for acceptance by Prof. F. Tortorella.

* Corresponding author. Tel.: +86 023 65105660; fax: +86 023 65104570.
E-mail address: qszhu@cqu.edu.cn (Q. Zhu).

http://dx.doi.org/10.1016/j.patrec.2016.05.007
0167-8655/

subset $NN_s(q)$ of X defined as follows

$$NN_X(q) = \{r \in X \mid \forall p \in X : D(q, r) \leq D(q, p)\} \quad (1)$$

Nowadays the improved algorithms based on the KNN method are widely used in a lot of research fields [5–7]. They increasing the KNN performance is obtained by the estimation of optimal k parameter [8,9] or the forms of distance metrics [10,11].

2.2. Reverse k-nearest neighbor (RkNN) method

One type of neighborhood that received attention is the concept of reverse k-nearest neighbor, and RkNN method appear in many practical situations such as decision support and resource management.

Definition 2 (Reverse Nearest Neighbor Search)**.** Give a set X of points and a query point q, the Reverse Nearest Neighbor Search problem is to find a subset $RNN_s(q)$ of X defined as follows

$$RNN_X(q) = \{r \in S \mid \forall p \in X : D(q, r) \leq D(r, p)\} \quad (2)$$

Korn and Muthukrishnan [12] firstly do the fundamental research in 2000, and then the concept of reverse k-nearest neighbors (RkNN) search is purposed [13]. Recently, research achievements aim to find the reverse k-nearest neighbors quickly and exactly [14].

Additionally, analogous to KNN and RkNN, the mutual k-nearest neighbor (MkNN) capture the inter-connectivity of adjacent regions. Brito et al. firstly use the connectivity properties of mutual nearest neighborhood graphs [15], and recently it is effectively used in classification [16,17] and clustering [18]. MkNN method reduces the computational complexity for large data sets. However, parameter k still exists, a bad k may led to unsatisfactory results.

3. Natural neighbor

Neither KNN nor RkNN, the problem of parameter selection is not in a position to avoid. As several studies have revealed, the solution of this problem often relies on the estimation of datas characteristics, by means of determining local decision boundaries in which the shape of the neighborhood can be modified to be more elongated. In this literature, we present a new term, natural neighbor, not only to estimate the parameter k, but also to explore a new way of nearest neighbor method without the parameter of k.

The natural neighbor method is inspired by the friendship of human society.

3.1. Friendship of human society

Friendship is a relationship of mutual affection between two or more people. It is a stronger form of interpersonal bond than an association. Friendship has been investigated in academic fields such as sociology, social psychology, anthropology, and philosophy.

As we know, friendship is the most elementary relationships in our life. As human, everyone must have one or more friends. If you wish to get along well with others, you are required to be friendly, and it would be used to make more friends.

Now we get two concepts, friendship and friendly. Case 1: considering a sample relationship in three people, named A, B and C. If A is friendly to B, B is friendly to A, C is friendly to A, we can only say that A and B are friends, A is a unilateral friend of C, but not a real friend of C. Thus the first friendship comes into being (A and B), and in contrast, we named the relationship between A and C unilateral friendship.

Case 2: then we can take such problem bigger. There are many people here in the virtual society, and everyone owns some, maybe three, unilateral friends. Everyone is friendly to three people in this case, and two people are friends if and only if each of them is mutually friendly to the other one. Therefore, some people may have friends, and some of them may not. It is dependent on the number of unilateral friends we choose.

Case 3: thinking about a city in the real world, we can ride the problem more complex. In our daily life, it is impossible to calculate the number to distinguish between friendly and unfriendly clearly. Then, someone who is more friendly may has more friends. So in this case, everyone here has a ranking list of the others, in which the higher level means friendlier. With the help of the principles below, we can find the friendship of this city. Firstly, everyone must have one or more friends. Secondly, two people are friends if and only if each of them is mutually friendly to the other one. Thirdly, the friendship is found by searching the lists of all people from beginning to the end. Absolutely, if we search everyones whole lists, everyone will have at least one friend. But it is necessary only if there is one person whose name lies on the end of everyones lists.

Case 4: the stranger. When there is a stranger in the city, his name must lie on the end of everyones lists because no one knows him. Thus we cannot find the real friendship of the city if we use the principles above. So the first principle must be modified. Let the city has the population of n, and then one stranger comes into the city. It is known that without the stranger, the friendship is found when the list number increases to r in case 3. It is clear that no one is friendly to the stranger until the list number increases to n, and we only want the number increases to r and stop, $r = n$ is a wrong result. Thus, the new principle is: anyone who belongs to the city must have one or more friends; the stranger(s) can have no friend only if the others have one or more friends, and when the list number continue increase, the people who have no friends still have no friends.

3.2. Natural neighbor method

As mentioned, the natural neighbor method is inspired by the friendship of human society to find the optimal path to overcome the disadvantages in KNN and RkNN. Particularly, the natural neighbor method can effectively determinate the neighborhood in a data set without the given parameter k and, meanwhile, calculate an approximate k.

The natural spirit of our method mainly manifests in three aspects: the neighborhood, the searching algorithm and the number of neighbors. Firstly, this neighborhood is inspired by the friendship of human society. Secondly, the searching algorithm can independently find the neighbor without human intervention. Thirdly, the process of determining the natural neighbor is a passive process, the number of all points neighbors is mutually independent, and it embodies the thought of nature.

3.2.1. Concepts of natural neighbor method

Given a set of data points $x_1, x_2, x_3, \ldots, x_n$ and some notion of similarity s_{ij} between all pairs of data points x_i and x_j, the goal is to find the natural neighbors of the points in the data set. Particularly, the notion of similarity can be given as prior knowledge, or be calculated and stored in a distance matrix. One of the most popular choices to measure this distance is known as Euclidean.

In what follows, we assume that X is a set of points, s_{ij} is the similarity between two points x_i and x_j. With the help of comparing the similarity, let $findKNN(x_i, r)$ denote the function of KNN searching which return the rth nearest neighbor of point x_i, $KNN_r(x_i)$ is a subset of X, and it is defined as follow

$$KNN_r(x_i) = \bigcup_{n=1}^{r} \{findKNN(x_i, n)\} \quad (3)$$