ORIGINAL RESEARCH PAPER
Volume-7 | Issue-11 | November-2018 | PRINT ISSN No 2277 - 8179
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

THE INFLUENCE OF K PROXIMITY SELECTION ON SOME MANIFOLD LEARNING

Computer Science

Han Baojin Tianjin Polytechnic University School of Computer Science & Software Engineering, Tianjin 300387

ABSTRACT

The selection of K-nearest points in manifold learning has always been an important factor affecting the extraction of some manifold learning features, and how to reasonably select K-nearest points has always been obtained in repeated experiments. This paper uses evolutionary algorithm to find out the different popular learning conditions for different K-nearest points. The experimental results show that the optimal K value in the data set improves the classification accuracy and makes the quality of feature vectors after feature extraction better.

KEYWORDS

Manifold learning, Data mine, feature extraction, Immune clonal algorithm

INTRODUCTION

Manifold learning algorithms have been popular since they were put forward in 2000. Part of the reason is that the increase of data volume leads to a greater consumption of computing resources and processing resources. So, the realistic requirement is how to condense data and make the relationship between data clearer after feature extraction. And data dimensionality is decreasing continuously for later processing. Another part of the reason is that in the face of large-scale data sets, people how to effectively use, in the face of data disasters helpless, and popular learning can extract important features, ignore the non-important features, so that we can achieve the goal of data reduction, so that important feature data extraction. In the past popular learning, such as PCA[1], MDS (Multidimensional Scaling Analysis)[2], LDA (Linear Discriminant Analysis)[3], this kind of popular learning does not need to select K-nearest neighbors, because they are equivalent to the global computation in the process of dimensionality reduction, the disadvantage of this method is that the computational complexity is large, the computational resources are more demanding, compared with the local search algorithm. In the global search algorithm, although the nominal effect is better than the local search, in practical application, because of the extraction of data features, this may lead to the important features and noise features are mixed together, making the advantage of global features because of noise. Pollution has a great impact, making the original advantages into disadvantages, and local search is in part of the data is constantly fusing each other, so that the local optimum and local optimum are combined, so that continuous fusion process, you can find the best solution, and the amount of calculation is small. Local search has been widely used for its less computational resource requirements, such as LLE (Local Linear Embedding)[4] and LE (Laplace Feature Mapping)[5]. So, it is very important to select K-nearest neighbor for local search. But how to select K-nearest neighbor is very important. In the past, the method is based on personal experience, or constantly adjusting K-value parameters in experiments. This not only wastes time, but also wastes computing resources, making applications. This paper proposes an immune clonal algorithm to select the K-nearest neighbor parameters, which can not only solve the problem that different data sets show different characteristics under different algorithms. It makes the classification accuracy increase in data dimensionality reduction, and the data quality after dimensionality reduction is better.

specific algorithm description:

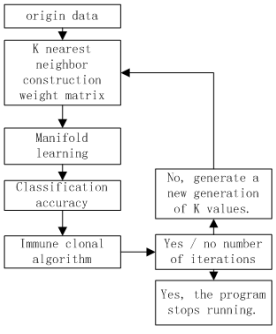

Figure 1 algorithm flow chart

The algorithm shows that: after the original data, K value carries on the weight matrix construction, then uses the popular learning method to carry on the characteristic extraction, after extracting the characteristic carries on the classification operation, with each different K value different classification accuracy, substitutes the immune clone algorithm, causes the classification accuracy rate to be high, the corresponding K value, in the immune gram The algorithm is dominant. After this operation, the optimal K value can be obtained under a certain number of iterations.

Manifold Learning

Building diagram

Use the KNN algorithm to connect the nearest K points of each point.

Weight:

a. LE weight

$$w_{ij} = e^{\frac{||x_i - x_j||^2}{t}}$$

w_{ij} is the weight between x_i and x_j, and x_i can find x_j in the K domain, otherwise w_{ij}=0, T represents the scale factor, which is used to adjust the weight.

b. LEE weight

$$min\ \varepsilon(w) = \sum_{i=1}^{n} | x_i - \sum_{j=1}^{k} w_j^i x_{ij} |^2$$

Among them, x_{ij}(j=1,2,…k) is the k nearest neighbors of x_i,x_i, $w_j^i$is the weight between x_i and x_ij, and to satisfy the condition $\Sigma_{j=1}^{k} w_j^i$=1, $\varepsilon(w)$ is the loss value, and the value of w_j^i is obtained when the loss value is minimal, and x_i can search for x_j in the K domain, otherwise w_j^i=0.

c. ISOMAP[6]

The geodesic distance is then replaced by the MDS algorithm to extract the feature vectors.

d. LTSA

$$\sum_i ||E_i||_2^2 = \sum_i ||T_i\left(I - \frac{1}{k_i} ee^T\right) - L_i\theta_i||_2^2$$

According to n local

projections θ_i=[θ_1^i,θ_2^i,.....,θ_k^i], the global coordinates $\{\tau_i\}_{i=1}^n$, for any $L_i \in R^{d*d}$ and $T = [\tau_1, \tau_2 \ldots . \tau_n]$, $\tau_1 = [\tau_{i1}, \tau_{i2} \ldots . \tau_{in}]$

The global reconstruction error is minimized to get the global coordinates.

Immune clonal algorithm

coding scheme: In solving the problem of K value selection, considering that K is a real value, but for the convenience of crossover and mutation, it is considered to use binary bits, which is convenient for later operation. Considering the part of experimental data set, this paper uses fixed length binary bits to represent K value, so that it can be more convenient in selecting K value. Close to the reality, this not only

simplifies the experiment but also improves the correct rate of classification and achieves the purpose of optimization.

Affinity function: Affinity function guides the evolution process of the whole population. In this paper, the K-means algorithm is used to deal with the classification problem, and the affinity function is set to ACC= $\frac{1}{n}\sum_{i=1}^{n} p(c_i, map(g_i))$, that is, the classification accuracy rate. The purpose of this method is to maximize the classification accuracy, and the parameters will be given in the experimental part.

Among them, c_i and g_i are the class labels which are regenerated after feature selection and those actually labeled in the original data set. Map () is the optimal mapping function, and Hungarian algorithm (Hungarian algorithm) is used to match the class labels generated after feature selection and the actual labels in the original data set. $p(c_i,g_i)$ is the indicator function, when $c_i=g_i$ is 1, when$c_i\neq g_i$ is 0, ACC is the classification accuracy.

The process of immune clone algorithm:
Step1. initialization to generate initialization population.
Step2. calculates initial population affinity.
Step3. determines whether the termination condition is the maximum number of iterations. If so, quit. Otherwise, continue.
Step4. selection, cloning, crossover and mutation operation
Step5. Select the best individuals from Step5 and replace some individuals in the original population to form a new population.
Step6. computing affinity
If Step7. reaches the maximum number of iterations, if it stops, otherwise the number of iterations will be +1 to Step4.

Experiment

We select the most commonly used swiss-roll data set, from three-dimensional to two-dimensional, so it is more intuitive, the method of discrimination is that the data projected after the ghost is small, and the data can reflect the relationship between three-dimensional spatial data.

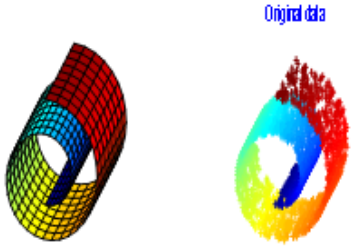

Figure 2 swiss-roll date

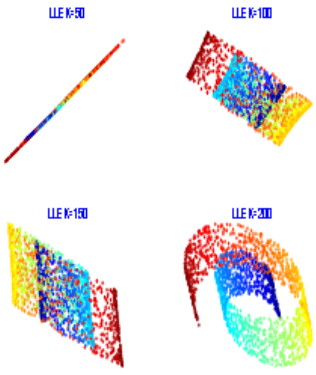

Figure 3 Projection of LEE

Through the above criteria: it can be clearly seen that when k = 200, the relationship between the relative space of the data can be observed more clearly and the location of the projection between the data is also reasonable. This can make the relative position of the data changed very little. K = 50 does not reflect the two-dimensional structure of the data space, on the contrary, it is too strict for data compression, resulting in the relative position of the data space can not be reflected, when k = 100 and K = 150 have a double shadow problem.

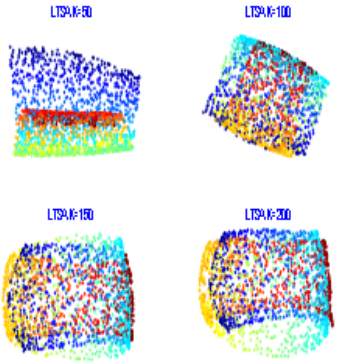

Figure 4 Projection of LTSA

When k = 100, it can be clearly observed that the structure of the data is similar to the structure of the three-dimensional space. When k = 150 and K = 200, it can be clearly observed that the data structure space is obviously deformed and does not conform to the relative relationship of the original data space. When k = 50, it is obvious that the data do not meet the observation requirements, most of the points do not reflect the real space, and focus on the tail of the projection.

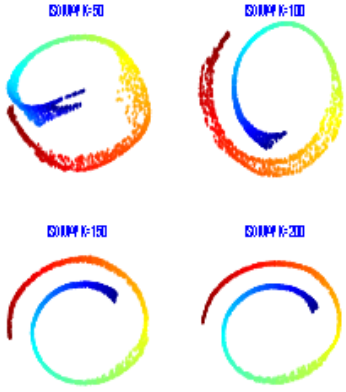

Figure 5 Projection of Isomap

When k = 100, the relationship between the data is distinct, and the ghost rate between the data is very small, the relative relationship is clear. When k = 50 has partial ghost, when k = 150 and K = 200, the data compression is too strict, forming a kind of original data space relative relationship can't be expanded.

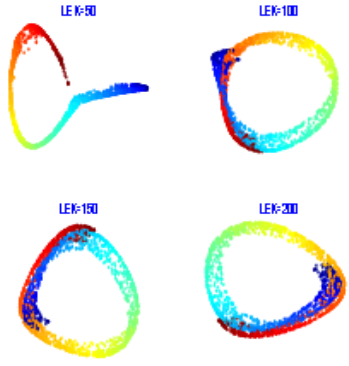

Figure 6 Projection of LE

When k = 50, relatively loose space can be formed between the data, and the data can reflect the three-dimensional data relationship, when k equals to the other three values, it is obvious that data duplication occurs.

CONCLUSION

The role of K-nearest neighbor in data feature extraction can be clearly seen from the above experiments, but with the increase of the amount of data, it is difficult to solve the problem of how these data will change

for the selection of K value, and the reasonable selection of K can really improve the quality of the data, the selection is not good will be the opposite, but whether these data and the number According to the noise, or is it related to the limitations of the algorithm itself, this point has yet to be clarified, with the development of science and technology, data in daily life is becoming more and more important, how to rationally use these data is still some problems to explore.

REFERENCES:

[1] Zarmehi N, Marvasti F. Sparse and low-rank recovery using adaptive thresholding[J]. Digital Signal Processing, 2018: 145-152.
[2] Ye K, Lim L. Every Matrix is a Product of Toeplitz Matrices[J]. Foundations of Computational Mathematics, 2016, 16(3): 577-598.
[3] Phinyomark A, Quaine F, Charbonnier S, et al. EMG feature evaluation for improving myoelectric pattern recognition robustness[J]. Expert Systems With Applications, 2013, 40(12): 4832-4840.
[4] Zhu Y, Zhu C, Li X, et al. Improved Principal Component Analysis and Linear regression classification for face recognition[J]. Signal Processing, 2018: 175-182.
[5] Bonvicini S, Mazzuoccolo G. Edge-Colorings of 4-Regular Graphs with the Minimum Number of Palettes[J]. Graphs and Combinatorics, 2016, 32(4): 1293-1311.
[6] Lee C, Elgammal A M, Torki M, et al. Learning representations from multiple manifolds[J]. Pattern Recognition, 2016: 74-87.