International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2015): 3.791

An Efficient Classifying On Several Datasets in Big Data Assistive K-Nearest Neighbor Algorithm

Jaafar Sadiq Qateef

Wasit University/college of education/Dept. of computer science
Wasit, AL-Kut, Iraq
jaafar.sadiq2000[at]gmail.com

Abstract: ***K nearest neighbors (KNN) is very great learning algorithm. Nowadays it`s been upgraded for several real applications. The large scaling of the datasets from the past k-nearest neighbor strategies is very suitable and natural. As we suggest the whole datasets are portioning into many part after selecting the suitable k-mean cluster for the partitioning of that datasets, then conducted a k-nearest neighbor classification for each parts. So, the medical imaging data& the big data can be conducted by set of experiments. Therefore, when we talk about the efficiency and accuracy the proposed K-NN classification is working well according to the experimental result of that algorithm.***

Keyword: big data; K-nearest neighbor; data clustering; classification

1. Introduction

Actually from the scaling within all types of applications, the terms named big data is important and very effectively study, in term of the (clustering &classifying), when we say about the algorithms of classification such as decision trees, SVM, simple network, and (KNN), the evident to scale the past copy of classification algorithms is necessary, so these can be the effective methods to use in the term of big data. Because of the convenience, easy-comprehend efficient performance of KNN, The gradation of K-NN classification into the big data applications will cover in this paper [13].

the method the Past KNN Initially choose the sample of K-nearest training for test sample, and then result the test sample with the class through k nearest training samples. So, KNN will calculate the distance of all training samples for every test sample with the mechanism of selecting neighbors and k-nearest training [25]. In big data, the previous k-NN that has been used is preventing because the high cost (means the sample size got a complexity of the linear time) [21].

Propelled by the late advance on big data, along with a big data, this paper covers new K-NN methodology for treatment. In particular, the suggestion is to select a k means cluster to split the entire Data Set into a many parts [26]. And after that we choose the near cluster for the classifying via KNN as training samples. So, the methods result better outcome than the traditional classification when we talk about the time cost and the performance of classification. [15].

2. Motivation

The analysis of KNN methodology has become an important analysis subject when treating with machine learning and processing of the data since1967 when they were discovered the algorithmic rule. In big data, when the using of traditional K-NN methodology has been applied, it is often classified into 2 components that meaning is to find the nearest samples quickly after that reducing the K-NN calculation by choose the representatives samples or on the other hand remove some samples to [21]. As an example, “in KNN method the certainly factor (CF) is planned and measure the skew class distribution that deal with it as unsuitableness” [14]. “Zhang planned abased density methodology of training data to reduce the quantity” [9]. “Li. got a new algorithmic rule that is perfectly relying on the using the labeled samples then add the conditions of screening method, Within the time complexity it’s creating the new algorithmic rule and will be considerably reduced, furthermore will no result that may be important on a result of algorithmic rule”[16]. According to these behaviors, it is applied for fundamentally in quickly search and for reduction of the dimension and also for enhancing the algorithm’s affectionate [17] [20]. Within the Data Set the gap between every testing sample and the training samples is computed by the algorithmic rule k-nearest neighbor and then result the differences of k nearest samples. The result of the actual k nearest neighbor can be found and by linearly time complexity that is guaranteeing the result of K-NN. Yet, for every test sample, the machine quality of the methodology of the linear search has been commensurate to get up the dataset’s training, wherever the scale of the training dataset is referred by n and the representing of the dimension is referred by d [12]. Within the big data, this complexness that is used is very costly. Due to the methodology of the K-nearest neighbor is not a training method, a brand new training method for KNN is planned here, Along with linear complexity the algorithmic rule clustering banned the training datasets. For every testing sample, during the testing method, we are trust to discover the clustering centers of k-nearest neighbor then selected an appropriate cluster for every one of the test sample. Yet, based on every cluster, it is constructing a new classification model [21].Particularly; the high similarity of the samples inside a cluster will be. The proposed method of k-NN is very well in two things, first reducing the time complexity, second don’t add any important effect on the classification accuracy when compare with traditional k-NN [22].

3. The Method

We will name a two type of processes in our rule, particularly a process of training and testing. A new training data set obtained for each test sample by selecting a closest cluster as

a task of the training process, within the K- nearest cluster, the K-NN algorithms is classifying every test samples with the using of testing process.

a. The Processing Of Training

The essential technique that is be used in data mining as a result of it may use for the segmentation of Databases called "clustering", also the compression of Data may also use for data mining. The cluster have been built for classes set of samples. Yet that means the group of testing samples that are be in the same group will have above normal similarity rather than the testing sample in any other groups. So, the high similarity will be in one group at every cluster on other hand the similarity is vanishing or low among clusters.

In fact there is a compound of the strategies of clustering classified into the subsequent categories: "grid-based, density-based, partitioning clustering and class-conscious clustering" [1], severally. even supposing the past clustering strategies showed smart performance, however they're restricted in its pertinence to big data due to theirs high machine complexness. To handle this, we consider using a cluster rule for 2 advantages: "low in complexity; and scales linearly" [5], "severally. The term of LSC means "Landmarks-based-spectral", So, for the linear combination of the landmarks [4], the explanation of this methods is represent the initial sample by picking the sampling P(less than n) as landmark"[24].So, this will be completely different from the technique of the traditional "spectral clustering" that represent every sample by using the whole samples, the complexity of the resemblance matrix is vitally reduced by the "LSC". In the precise time, the linear scales are downed by the complexness of the solving of Eigen value [24].

"The collecting a set of basis vectors and for every sample representation of the bases the method is to compress the original samples that tried by the LSC" [24]. That means, looking for p representative samples. "During this method, we've got two easy and effective strategies to pick landmark sample from original sample, like sampling and k-means-based technique. sampling at random selects samples as landmark samples whereas the k-means-based technique initial conducts clustering on all samples many times (no ought to convergence) then uses the cluster centers as a representative the landmark samples. During this paper, we have a tendency to repeat k-means rule ten times then using the cluster centers as the landmark samples"[25].

First, for compounding the landmark matrix Z, we have a tendency to treat each sample as a "basis vector". Fordoing that every sample$X = [X1, X2, \dots, Xn] \in R^{m \times n}$ is represented by the "landmark P" that has been used by the LSC. So, we want to seek out the matrix W that is projection matrix of X at the landmark matrix Z [10]. The projections perform may be outlined as follows:

$$\mathbf{W_{ji}} = \frac{\mathbf{k_h}(\mathbf{x_i}, \mathbf{z_j})}{\sum_{\mathbf{j'} \in \mathbf{Z_{<i>}}} \mathbf{K_h}(\mathbf{x_i}, \mathbf{z_{j'}})} \; \mathbf{J} \in \mathbf{Z_{<i>}} \qquad \mathbf{1}$$

Where Z_i is j-th column vector of Z, and $Z_{<i>}$is a sub matrix of Z composed of a (r) closest landmarks of x_i Here we need O (pmn) to construct W.K_hIs referred a kernel function along with the bandwidths denoted by h.

The kernel`s Gaussian will be

$$\mathbf{K_h}(\mathbf{x_i}, \mathbf{z_j}) = \mathbf{exp}\ (-|\mathbf{x_i}, \mathbf{z_j}|^2 / 2\mathbf{h^2})$$

It is represented as one amongst of the foremost ordinarily working on it. To calculate the graph of matrix we tend to choose the analysis's spectral on the graph of the base of landmark according to following:

$$\mathbf{G} = \hat{\mathbf{W}}^{\mathbf{T}}\, \hat{\mathbf{W}} \qquad \mathbf{2}$$

During this technique, we consider $\hat{W} = D^{-1/2}W$ which has efficient Eigen-decomposition., we decide wherever D is that the row adds of W. Note that every column of W sums up to one and so the degree matrix G is I.

Let SVD (i.e. Singular value Decomposition) of the $\hat{\mathbf{W}}$ as the following:

$$\hat{W} = \mathbf{U\Sigma V^T} \qquad \mathbf{3}$$

where $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}, \dots, \mathbf{u_k}] \in \mathbf{R^{P \times P}}$is termed as a vectors of left singular of the first k eigenvectors of $\hat{Z}\hat{Z}^T$ $\mathbf{V} = [\mathbf{v_1}, \mathbf{u_2}, \mathbf{v_3}, \dots, \mathbf{v_k}] \in \mathbf{R^{p \times n}}$is termed as the vector of a right singular of the k eigenvectors of $\hat{Z}^T\hat{Z}$ We compute U within $O(p^3)$, linear for the size of sample. So, V can be computed as the following:

$$\mathbf{V^T} = \mathbf{\Sigma^{-1} U^T \hat{W}} \qquad \mathbf{4}$$

The overall time quality of V is $\mathbf{O(p^3+p^2n)}$, that may be a important reduction from $\mathbf{O(n^3)}$ by considering $\mathbf{p<<n}$. every row of V may be a sample and apply k- to induce the clusters. as a result of the time quality from $O(n^3)$ to O(n), the LSC algorithmic program well reduces process time. Thus, the planned algorithmic rule that will be executed within the big data is getting a lower level of complexness.

"LSC can confer bathe following algorithmic program1" [26]

Input: n data points $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \dots, \mathbf{x_n} \in \mathbf{R^m}$; Cluster number k;
Output: k clusters;
1 Produce p landmark points using the k-means method;
2 Constructing a landmarks matrix Z between the data points and the landmark samples, ,
with the affinity computed according to Eq.(1);
3 Compute the first k eigenvectors of WW^T, denoted by $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}, \dots, \mathbf{u_k}]$
4 Compute $\mathbf{V} = [\mathbf{v_1}, \mathbf{u_2}, \mathbf{v_3}, \dots, \mathbf{v_k}]$equation(4)
5.By applying the k-mean we resulting a cluster, and V is being the data point.

b. The Processing Of Testing

First we assuming that the LSC algorithmic rule come to result a clusters center and a k- cluster, after that we are for every test sample discovers the closest cluster center also the freshen training dataset for every test sample the identical cluster used as the corresponding of that training dataset. Furthermore, within the fresh training dataset we tend to apply KNN for classification the test samples because the high similarity for cluster has been selected. Therefore, the

accuracy of the classifying still proves the planning algorithms rule.

The methodology planned in algorithmic rule 2.

algorithms two The pseudo
Input: training dataset, take a look at samples Y;
Output: class label;
1- Turn out m cluster centers exploitation LSC algorithmic rule, indicate by C1, C2, C3, …, Cm
2- Calculate D(y, C) (distance) between the test sample y and every cluster center, indicate by D(y, C_j) where j=1, 2,..., m;
3- Calculate the closest cluster center C_j to (y, C_j)=min, j=1, 2,..., m;
4- The C_j is used as identical cluster ; as a dataset's training, indicate by fresh X_j
5- Apply to k-NN algorithmic rule to portend y within the training dataset.

The new value of Xj that be concluded from the algorithm 2 is very much smaller than the scale of the training datasets. Once a scale of m is massive, LC-k-NN is straightforward to scale back the calculation of k-NN and the quality of the classification will be improved. At an equivalent time, the accuracy of the classification goes to be low because the overhead of cluster will be increased along with the rise of m. the quantity of cluster M must be given a rational vale to avoid this case.

Generally if a value of m is large and the efficiency of classification is high the higher of the accuracy of classification comes up. So, the condition that causes a low accuracy is the training dataset distribution is comparatively focused. Additionally, if the value of m is low, i.e., m=1, that's standard k-NN algorithmic rule, which prohibits the k-NN to be utilized in big data. Therefore we have a tendency to set m in an exceedingly appropriate range. Presumptuous that the planned algorithmic rule would like M memory space, the memory of computer is M1 and therefore the smallest category of training sample is n0. To the present finish, the value range of m is produced according to following;

$$N0 > M > \frac{M}{M1} \qquad 5$$

According to above we have a tendency to conduct a k-means cluster to separate the full dataset into many components, every of that we have a tendency to conduct k-NN classification, the choice of k value is vital within the next process. ", the sample size of datasets that higher than one hundred should be satisfied by the suitable value of k =square root (n) as L all mentioned" [8].

Yet, these are proved it is not appropriate to all cases of data sets. In sub-cluster that the k value not be setting vary large, the samples have strong correlation. So, the choice of k value ought to be set as little as possible within the case of the precision of the high classification.

4. Experiments

To see the efficiency of LC-k-NN rule, we tend to took the k-NN because the baseline and created a differentiation between k-NN, (LC, RC (random cluster)) k-NN, and some other real datasets.

a. Results

Due to the value of m directly affect the real applications final performance; it is extremely vital to the LC-kNN algorithms. To bringing a great efficiency for LC-KNN, we are selecting a different value of M for the LC and RC (KNN) algorithms and a set of experiment are trying on datasets that led to select a suitable value of m to apply on that datasets. For applying that, the different value of M for example M= {5, 10, 15, 20, 25, 30} the algorithms of KNN (LC&RC) are carrying out on the datasets. We showed in this table the difference of time cost and the classifying performance of the algorithms KNN (RC&LC) according to previous algorithms 1 & algorithms 2 respectively.

Table 1: The accuracy of classification and the time-cost on USPS dataset

M-value	Criteria	LC-K-NN	RC-K-NN
30	Time-cost	1.9249±0.0023	1.6441±0.0002
	Accuracy	0.9275±1.1596e-005	0.8607±4.6629e-005
25	Time-cost	1.9971±0.0042	1.8586±0.0008
	Accuracy	0.9279±1.0632e-005	0.8788±4.9912e-005
20	Time-cost	2.5152±0.01923	2.3197±0.0005
	Accuracy	0.9295±4.9233e-006	0.8765±7.4884e-005
15	Time-cost	2.7255±0.0072	2.4852±0.0027
	Accuracy	0.9333±4.1620e-006	0.8959±5.1798e-005
10	Time-cost	3.7599±0.0237	3.5584±0.0102
	Accuracy	0.9321±7.1298e-006	0.9019±1.6460e-005

Table 2: The accuracy of classification and time-cost on MINST datasets

M-value	Criteria	LC-K-NN	RC-K-NN
30	Time-cost	1.7274±0.0011	1.5457±0.0002
	Accuracy	0.8313±3.8678e-005	0.6396±6.9156e-005
25	Time-cost	2.1148±0.0094	1.8240±0.0020
	Accuracy	0.8338±8.7844e-005	0.6478±2.2689e-004
20	Time-cost	2.1490±0.0065	2.0564±0.0011
	Accuracy	0.8353±3.3233e-005	0.6657±2.4739e-004
15	Time-cost	3.1222±0.01397	2.8905±0.0456
	Accuracy	0.8364±2.3136e-005	0.6840±2.3333e-004
10	Time-cost	3.5504±0.0927	2.9369±0.0508
	Accuracy	0.8389±3.1656e-005	0.7221±4.8878e-005

Table 3: The accuracy of classification and time-cost on GISETTE dataset

M-value	Criteria	LC-K-NN	RC-K-NN
30	Time-cost	11.3922±0.0658	8.4064±0.0784
	Accuracy	0.9192±5.3796e-004	0.9079±1.0366e-004
25	Time-cost	13.8645±1.5093	9.9201±0.3696
	Accuracy	0.9321±6.4810e-004	0.9150±7.0000e-005
20	Time-cost	16.2759±0.8880	12.7685±0.0966
	Accuracy	0.9411±5.4699e-004	0.9166±2.8267e-005
15	Time-cost	23.1904±1.0894	18.0106±0.2434
	Accuracy	0.9494±1.3378e-005	0.9252±1.0573e-004
10	Time-cost	28.5940±3.2405	23.3933±0.9677
	Accuracy	0.9526±1.4511e-005	0.9311±5.0989e-005

Table 4: The accuracy of classification and time-cost on PENDIGITS dataset

M-value	Criteria	LC-K-NN	RC-K-NN
30	Time-cost	2.2022±8.9611e-005	2.1805±7.4785e-005
	Accuracy	0.9683±1.5809e-006	0.9088±1.8409e-004
25	Time-cost	2.5468±0.0083	2.5270±0.0056
	Accuracy	0.9687±3.5642e-006	0.9216±1.5677e-004
20	Time-cost	2.2554±2.1569e-004	2.2233±6.4795e-005
	Accuracy	0.9700±2.5390e-006	0.9163±1.5515e-004
15	Time-cost	2.5709±0.0089	2.5451±0.0011
	Accuracy	0.9711±6.0196e-006	0.9316±1.0341e-004
10	Time-cost	2.4056±0.0101	2.3380±0.0041
	Accuracy	0.9721±4.7991e-006	0.9452±3.5382e-005

From Table 1~4, we have a tendency to found that the planned 2 algorithms required less time with the larger range of clusters m, whereas high time cost for the smaller number of m. for instance, once m value is ten, i.e., we have a tendency to separate the complete dataset into ten components. We have a tendency to conducted kNN classification for every part, and in the whole data set the classification obtained the cost time from 1 to 10 forth method conducting.

Do not forget that the KNN is the parent of these 2 algorithms and can be consider them as extended algorithms of the k-NN, and the classification accuracy of these algorithms should be very close to the possible K-NN. So, these are very sensitive for the large number of cluster m so when the value of m is high the high classification accuracy resulted, otherwise when m is low the low classification accuracy resulted. Have a tendency to found that the lager range of clusters m have high classification accuracy; on other hand a small number of cluster m give low classification accuracy of these two algorithms. Additionally, from Table five, we have a tendency to found that 2 algorithms were nearer to K-NN classification accuracy with m =10.

b. Performance comparison of k-NN

This section and due to the evaluations for the classification task, we have a tendency to set m=10 and k=1. Then, we have a tendency to use period of time and the classification accuracy [23]. Experimental results area unit are shown the best time (shorter), best performance and the best accuracy of classification in Table 5.

Table 5: The affection of three algorithms on classification accuracy, Time-Cost of five datasets

Types of Datasets	K-NN		LC-K-NN		RC-K-NN	
PENDIGITS-dataset	0.9780	0.9721	2.4056	7.2982	0.9452	2.3380
GISETTE-dataset	0.9660	0.9526	28.594	217.3327	0.9311	23.3933
USPS-dataset	0.9482	0.9355	3.7605	32.8764	0.9019	3.5584
MNIST-dataset	0.8635	0.8389	3.5504	24.1575	0.7221	2.9369

The time cost that is resulted in table 5 is telling us that the projected LC-K-NN&RC-K-NN enhanced approximately seven to nine times than standard K-NN. The percentage (3.4 to 14) of RC-K-NN and (1 to 2.6) of LC-K-NN is lower than standard K-NN in terms of the evaluation of classification accuracy, So, the conclusion according to time cost and the accuracy of classification is tend to result that the LC-k-NN works well than all other.

5. Conclusions

The main idea of this work is to separate the whole data set for several parts by selecting the appropriate k-means clustering according to the new proposed efficient K-NN classification. Then for every part we conducted K-NN classification. For doing that, the process of testing and the process of training are parted from the conventional K-NN method. Furthermore, the parameters K&M should we analyzed by an appropriate value. Moreover, we took a set of experimental comparison among the RC-K-NN and LC-K-NN and K-NN with noted that the K-NN is the baseline for them. As a result of our work in term of efficiency and accuracy the suggested K-NN classification working well, and in big data it is suitable to dealing with.

References

[1] A. Andrew, M. Jordan, Y. Weiss. On spectral clustering: Analysis and algorithm. In Advance in Neural Information Processing Systems 14.MIT Press.
[2] K. Bache and M. Lichman. UCI machine learning repository, 2013.
[3] C.-C. Chang and C.-J Lin. LIBSVM: A library for support vector machines. ACM Trans-on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[4] X. Chen, D. Cai. Large Scale Spectral Clustering with Landmark-Based Representation. AAAI, 313-318, 2011.
[5] M. Filippone, F. Camestra, F. Masulli, et al. A survey of kernel and spectral methods for clustering. Pattern Recognition, 41:176-190, 2007.
[6] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, X. Wu. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search, IEEE Transactions on Image Processing, 22(1):363 376, 2013.
[7] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai. "3D Object Retrieval and Recognition with Hypergraph Analysis", IEEE Transactions on Image Processing, 21(9): 4290-4303, 2012.

[8] U. Lall and A. Sharma. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research, 32(3):679–693, 1996.
[9] R. Li, Y. Hu. A density-based method for reducing the amount of training data in kNN text classification. Journal of Computer Research and Development. 41(4):539-545, 2004
[10] W. Liu, J. He, S. Chang.Large graph construction for scalable semi-supervised learning. In Proceedings of the 27 International Conference on Machine Learning, 2010. 11. X. Wu, C. Zhang, S. Zhang. Database classification for multi-database mining.
[11] Information System, 30(1):71-88, 2005. 12. X. Wu, S. Zhang. Synthesizing High-Frequency Rules from Different Data Sources.
[12] IEEE Transactions on Knowledge and Data Engineering, 15(2):353-367, 2003.
[13] Wu, C. Zhang, S. Zhang. Efficient mining of both positive and negative association rules. ACM Transactions on Information Systems, 22(3):381-405, 2004.
[14] S. Zhang. KNN-CF approach: Incorporating certainty factor to knn classification. IEEE Intelligent Informatics Bulletin, 11(1):24–33, 2010.
[15] Y. Zhao, S. Zhang. Generalized dimension-reduction framework for recent-biased time series analysis. IEEE Transactions on Knowledge and Data Engineering, 18(2):231-244, 2006.
[16] D. Zhao, Wei. Zou, G. Sun. A Fast image Classification Algorithm using Support vector Machine. In ICCTD, 2010.
[17] Zhu, Z. Huang, H. Cheng, J. Cui, H. Shen. Sparse hashing for fast multimedia search. ACM Transaction on Information Systems, 31(2):9, 2013.
[18] Zhu, Z. Huang, H. Shen, X. Zhao. Linear cross-modal hashing for efficient multimedia search. In ACM Multimedia, pages 143-152, 2013.
[19] Zhu, Z. Huang, H. Shen, J. Cheng, C. Xu. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognition, 45(8):3003-3016, 2012.
[20] Zhu, Z. Huang, Y. Yang, H. Shen, C. Xu, J. Luo. Self-taught dimensionality reduction on the high-dimensional small sized data. Pattern Recognition, 46(1):215-229, 2013.
[21] Zhu, L. Zhang and Z. Huang, A Sparse Embedding and Least Variance Encoding Approach to Hashing, IEEE Transactions on Image Processing, 2014.
[22] Zhu, S. Zhang, Z. Jin, Z. Zhang, Z. Xu: Missing Value Estimation for Mixed-Attribute Datasets, IEEE Transactions on Knowledge and Data Engineering, 23(1):110-121, 2011.
[23] Zhu, Z. Huang, J. Cui, H. T. Shen. "Video-to-Shot Tag Propagation by Graph Sparse Group Lasso". IEEE Transactions on Multimedia, 15(3): 633-646, 2013.
[24] Zhu, H.-I Suk, D. Shen. A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis. NeuroImage, 2014.
[25] Zhu, H.-I Suk, D. Shen. Matrix-similarity based loss functions and feature selection for alzheimer's disease diagnosis. In CVPR, pages 3089-3096, 2014.
[26] Zhu, X. Li, and S. Zhang, Block-Row Sparse Multiview Multilabel Learning for Image Classification, IEEE Transactions on Cybernetics, accepted, 2015