IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Determination value k in k-nearest nieghbor with local mean euclidean And weight gini index

To cite this article: M E Saputra *et al* 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* **420** 012098

View the article online for updates and enhancements.

Determination value k in k-nearest nieghbor with local mean euclidean And weight gini index

M E Saputra[1], H Mawengkang[2*], E B Nababan[3]

[1]Student in Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
[2]Faculty of Computer Science and Mathematics, Universitas Sumatera Utara, Medan, Indonesia
[3]Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

hmawengkang@yahoo.com

Abstract. K-*Nearest Neighbor* is the algorithm that included into the category of algorithms supervaised learning is bound process to distinguish the classes that already exists. Nearest neighbor is calculated based on the value of k that determines how the nearest neighbor in consider on the distance class data for *k-nearest neighbor* based on the determination of the value of the *k*. The determination of the class test data *local mean based k-nearest neighbor* using the measurement of the distance closest to each using *eucllidean distance* from each class data. Model based approach the weight of *Gini Index* is in need to give the weight of each attributes to determine the value of k. Research In this time I get the results of k best at the *thyroid data* which is a type of *unbalanced* data to obtain the value of the k highest k=44 until k=46 with accuracy of the closest neighbors of 71,19% and the value of the k lowest is k=50 of 69,30%. Then the results of the value of the k=44 until k=46 become k best on the processing of this time. It can be concluded for the data class is not the same will result in class data become random repeatedly until the limit of the determination of the value of k as well as exceeding the value of the k highest.

1. Introduction

K-*Nearest Neighbor* is the algorithm with the category distance based algorithms [1-6]. Distance based algorithms is the algorithm that determines the similarity of data or object based on the closeness of the distance between the data to a class or a label or other data groups. Equation between the data on the k-*nearest neighbor* is determined by using the measurement of the model of a certain distance.[24] one classifier non-revolutionary Parametric capabilities that has been widely used in various fields, pattern classification learning machine, clustering and various research areas in biometric devices.

[7-10] the purpose of *k-nearest neighbor* where the closest neighbors are calculated very effective against nonparametrik technique in pattern classification, but the performance of the classification depends on the point of the average generally variable that correlates with the point that far *Outlier*. [11-12] generally k used in each class that can cause a high sensitivity with the value of ***k***. If *k* is too small a useful classification information may not be enough, while the value of ***k*** that can be easily cause outlier including in the k closest neighbors class center.

[4]shows that the combination of DWKNN LMKNN and able to improve the accuracy of kNN classification, [14-17] where the accuracy of the average on the test data with enhanced accuracy that occurs on the data sets lower back pain effects.[22-23] In pattern recognition statistics, the classification performance of KNN based revolutionary Parametric capabilities classifiers really influenced by outliers, especially in the case of small training sample size. As we know, LMKNN is for strong outliers using each vector mean local and and then computes the meaning of vectors of the k closest neighbors in each class.

[14-16] Results of the accuracy of the performance of the k-*nearest neighbor* is still lower than the other method. One of the causes of the low accuracy produced, because each of the attribute has the same effect on the process of classification, while some characteristics that are less relevant point on the lack of class classification tasks for the new data. Use the *Gain Ratio* as a parameter to see the correlation between each of the attributes in the data and use the *10-fold Cross-Validation* with abalone data sets.

[2]Algorithm *k-nearest neighbor* samples determined automatically using clustering techniques. After partitioning the train mean data set, labels in the cluster moving centers are determined. [21] in obtaining the value of the k on the mean partitions to a limited number of *k* and homogeneous cluster separately to evaluate the result is to find the galactic again optimal number properties, such as cluster density, size, shape and separability is usually examined by some cluster validation method. [25] an integral part of the value of the action coefficient *k* about advanced violations weighted, by considering the distribution of the characteristics and factors that affect the value of the k and nerve network algorithm is adopted for research to select dynamically the value of k based on the application of techniques.

[14-18] *Gini Index* can be considered as the probability of two randomly selected data that has the *class* different. *Gini Index* is a method that has a good performance to calculate the weight of the attribute because it can reduce the influence of *outlier* from handling ability to measure the data divergensi and measure the impurities data. Select the value of k using the model based approach *Gini Index* is in the need to determine the value of k. In the first category where the entire data classified into data trained and test data. The distance evaluated from all data trained to test data on the lowest distance.

2. Problem

Determine the value of k in *k-nearest neighbor* which is an effective method to be used in the classification [11]. But there are factors that influence to determine the value of k is if the value of k too small can cause the required information is not sufficient and if the value of the k great easy result in outlier[12]. Then, how to get the value of *k* is good between the value of k small with the value of *k*. So the results to get the value of *k* is good on each of the data in detail.

3. Euclidean distance k-nearest neighbor

[19]To search for the closest distance between the data that is evaluated with k in the training data. Using the counting equation to find the distance in *Euclidean* equation.1.

$$\mathrm{d}_1(x,y) = ||x-y||_1 = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2} \tag{1}$$

Sort remote results obtained where:

d: Distance	y_i: test data
x_i: data samples tested	r : the number of features in the vectors of data

3.1. Local mean based k-nearest neighbor

[11-13] *Local Mean K-nearest neighbor* is a simple nonparametrik classification, effective and powerful. Proven to be able to improve the performance of the classification and reduce the influence of outlier and also in the size of the small amount of data. See Figure 1.

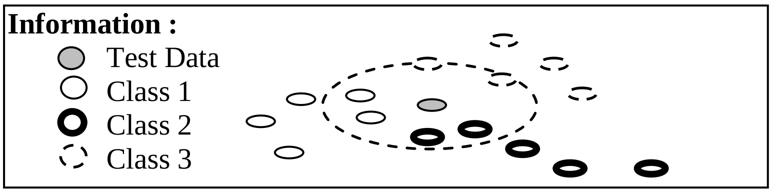

Figure.1 Nearest neighbors from each class

[11-12]this phase is the contribution of LMKNN method. The value of K on LMKNN very far different from the value of k in KNN, where on the KNN the value of k is the number of the nearest neighbor from the whole sample data, [15]while on LMKNN value is the number of k nearest neighbors from each class sample data. [4]Now steps work on *Local Mean Based k-nearest neighbor* in determining the value of *k* will need to calculate the distance test data throughout the data from each class data using *Euclidean distance model* continued to sort the distance between the data from the smallest to the biggest *k* from each class. Count local mean vector from each class with equation.2.

$$m_{w_j}^{k} = \frac{i}{k}\sum_{i=1}^{k} y_{i,j}^{NN} \qquad (2)$$

Specify the class test data with how to calculate the distance closest to the local mean vector from each class data with equation.3.

$$w_c = argmin_{w_j} d\left(x, m_{w_j}^{k}\right), j = 1, 2, \dots, M \qquad (3)$$

[11-13] classification of *local mean based k-nearest neighbor* with 1-NN if the value of the k=1. *Local mean based k-nearest neighbor the* value of k is the number of closest neighbors are selected from each class at trained data.

3.2 Gini Index

[6]*Gini Index* is usually used this approach the algorithm c*lassification and regression trees algorithm* (CARTS) and SPRINT to size of how the object of the choice of random data trained. Size of the ineffectiveness of pure reach 0 when only 1 class that is on a point. But on the contrary will reach a maximum when the size of the class at the point of balance.

[19-20] if S is the association S samples have different classes (C_i, i = 1...n). According to the class differences, we can divide S into n subset (S_i, i = 1...n). Suppose that S_i is a collection of samples including the class C_i, S_i is the number of samples from the set of S_i, then *Gini index* set S is equation.4.

$$Gini\ (S) = 1 - \sum_{i=1}^{n} P_i^2 \qquad (4)$$

Where P_i is the probability of each sample including in C_i and estimated with$\frac{s_i}{s}$. *Gini index* (S) minimum is 0 that all the members in such places including in the same class.

[19]can also for comparison of the advantages and disadvantages of the function of the selection of the weight so it can be used on the equation 5.

$$Gini\ Index(t) = 1 - \sum_{i,j=1} p\ (j|t)p(i|t), i \neq j \qquad (5)$$

Where : - *P(j|t)* = proportion of class j on the symbol of t
- *P(i|t)* = proportion of class i on the symbol of t

3.3 K-Fold cross validation

[19] In this research using the *10-Fold cross validation* is a method using 10 percent of cross *validation* that in each data used in the same amount for training and the right to one time testing. Assume that dataset are broken into two parts of the same size. Part one for data trained and the other for the test data, this approach is called the *two-fold cross validation.* [7]special form of this method when *k* in set *k=N* in the amount of data in a set of data. Can be called also with *leave-one out,* namely test dataset only 1 data only while the training process is done as much as *N* times. The advantages are almost all in the dataset can be in the rice so that got the value k is accurate.

4. Methodology

[4-6]on this research will be a few steps to achieve the goal of research. The first step using the algorithm *Gini Index* for the next weight adjust the data value of ***k*** with four main steps:

a) Gini index
b) Division of data for test data to be 10 percent of fully data
c) Count the distance test data and trained using *Euclidean distance model*
d) Determine the ***k*** with a *Local Mean Based k-nearest neighbor*
e) Value of the k highest k best

Very expected on research design can be in measure the value of each attribute is based on *Gini Index*[5]. *The* determination of the value of *k* based on the distance and data with [19] process *k-fold cross validation* and the *local mean based k-nearest neighbor* in the classification of with k-*nearest neighbor* and obtained the results of ***k*** is good in accordance with the design in research [4]. Steps can be seen in the overall Figure of research through the flow chart in figure 2.

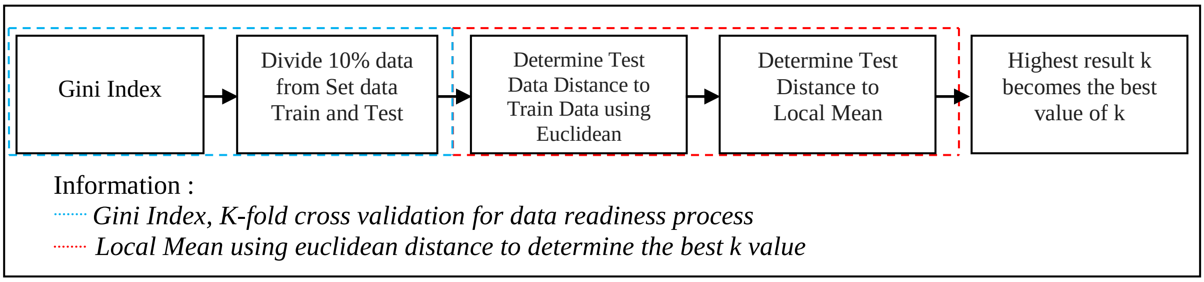
Gini Index
Divide 10% data from Set data Train and Test
Determine Test Data Distance to Train Data using Euclidean
Determine Test Distance to Local Mean
Highest result k becomes the best value of k
Information :
Gini Index, K-fold cross validation for data readiness process
Local Mean using euclidean distance to determine the best k value

Figure 2. Research Design

Research design above can be in describe to measure the value of the weight of each attribute is based on *Gini Index.* Determination of the value of k based on the *euclidean distance* and the sharing of data with the process of *k-fold cross validation* and the *Local Mean Based k-nearest neighbor* in the classification of *k-nearest neighbor* and obtained the results of ***k*** is good for in the set as the results of the rice.

5. Results and discussion

Thyroid is one of *unbalanced* dataset and also popular as a good data. The data set consists of 215 samples with 5 attributes and have 3 class, namely normal (C1), hyperthyroidism (C2), and hipotiroidisme (C3). In this trial, data trained numbered 193 data trained and test data numbered 22 data. With the distribution of the data using the *10-Fold cross validation* on the *Thyroid data*. Then the calculation of the distance between the data trained and test data using *euclidean distance model* using equation.1. Now the distance that is produced can be viewed directly data that is already in the sort by discending, now the order of the closest distance between data can seen table 1.

Table 1. Distance that has to be ordered from the closest on the *Thyroid Dataset*

Data Test	**Sequence of the closest distance**						
	1^{St}	2^{nd}	3^{Rd}	4^{th}	5^{th}	...	193^{th}
T1	2.448383	5.372618	3.086731	3.879169	3.234679	...	6.524390

Results on the process of weigh *Gini Index, data breakdown 10-fold cross validation* and *euclidean distance* on the *local mean* in the classification of *K-Nearest Niegbor* count from the distance as much as k closest neighbors to each class data. After getting each value *k* then continued with the sort the value of *k* in discending in table 2. Below are the results of the sorting of each value *k* in get with a discending sorting.

Table 2. Sort results for each value of k on the *Thyroid Dataset*

value of the *k*									
k=44	k=45	k=46	k=41	k=42	k=43	k=47	k=38	k=39	k=40
0.7119	0.7119	0.7119	0.7073	0.7073	0.7073	0.7073	0.7028	0.7028	0.7028
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k=31	k=32	k=33	k=34	k=35	k=36	k=37	k= 48	k=49	k=50
0.6980	0.6980	0.6980	0.6980	0.6980	0.6980	0.6980	0.6978	0.6978	0.6930

Now the reasons for using the *graph radar* because suitable for read and examine the results in the determination of the value of *k* that is produced from the process of *Gini Index* and *local mean* in the classification of *k-nearest nieghbor* to determine the value of *k*. From process *Gini Index,10-fold cross validation* and *Local Mean* in the classification of *k-nearest nieghbor* in continue to determine the value *k* that both discending with *k* highest that will be in use as the value of *k* best see on figure 3.

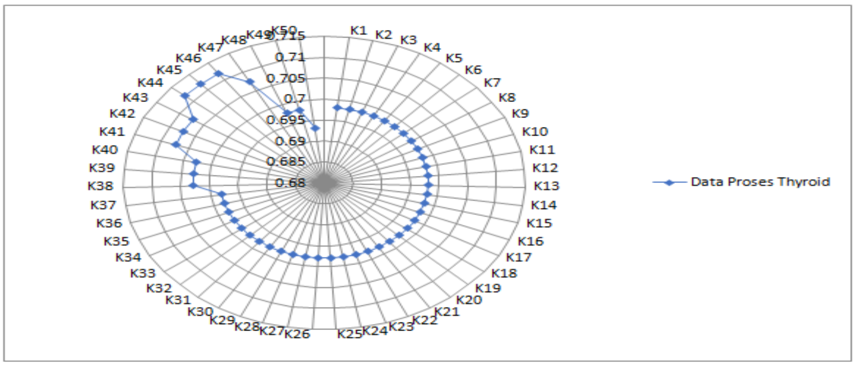

Figure 3. Graph the results of the value of ***k*** is good on dataset *thyroid*

6. Conclusion

From the Results Table.2 with the results of the graph Figure.3 Graphs the results of the value of ***k*** Is Good on the *thyroid dataset* is data that has a number of different classes *Unbalanced* and also data that have different challenges in research. Need for research with the processing of *Gini Index, k-ford cross validation* and l*ocal mean* in the classification of *k-nearest neighbor* Value ***k*** lowest k=50 of 69,30% while the value of the k highest k=44 until k=46 of 71,19%, then the value of the k highest become the value of ***k*** is good for the *Thyroid Data.* Can use this method to determine the best k classification *k-nearest neighbor*, so that further research will no longer need to test to find the value of ***k*** best of basis again. But further research can examine to find the value of ***k*** best on more data and with different data again so also with the methods.

Acknowledgement
High appreciation solution be given to Lembaga Penelitian Universitas Sumatera Utara (LP USU), the Graduate School of Computer Science Fasilkom-TI USU and rector of the Universitas Sumatera Utara has support this research can work with good and the facilities provided.

References

[1] Gou, J., Zhang, Y., Rao, Y., Shen, X., Wang, X. & He, W. 2014. Improved Pseudo Nearest Neighbor Classification. *Knowledge-Based Systems*70: 361-375.

[2] Hosein Alizadeh, Behrouz Minaei-Bidgoli and Saeed K. Amirgholipour A New Method for Improving the Performance of K Nearest Neighbor using Clustering Technique Journal of Convergence Information Technology Volume 4, Number 2, June 2009.

[3] Kataria, A. & Singh, M.D. 2013. A Review Data Classification Using K-Nearest Neighbour Algorithm. *International Journal of Emerging Technology and Advanced Engineering* **3**(6): 354-360.

[4] K U Syaliman, E B Nababan, and O S Sitompul Improving the accuracy of k-nearest neighbor using local mean based and distance weight*International Conference on Computing and Applied Informatics (2017)Journal of Physics: Conf. Series 978 (2018) 012047.*

[5] Tyas Setiyorini, Rizky Tri Asmono Penerapan Gini Index Dan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kognitif Soal Pada Taksonomi Bloom Jurnal Pilar Nusa Mandiri Vol. 13, No. 2 September 2017.

[6] Wenqian Shang, Youli Qu, Haibin Zhu, Houkuan Huang, Yongmin Lin and Hongbin Dong An Adaptive Fuzzy kNN Text Classifier Based on Gini Index Weight Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC'06) 0-7695-2588-1/06 $20.00 © 2006 IEEE.

[7] Bhatia, N. & Vandana., 2010. Survey of Nearest Neighbor Techniques. *International Jurnal of Computer Science and Information Security (IJCSIS)* 8(2): 302-305.

[8] Buana, P.W., Jannet. S.D.R.M., & Putra, I.K.G.D. 2012. Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News. *International Journal of Computer Applications*11(11):37-42.

[9] Chen, Y., Hao, Y. 2017. A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction. *Expert Systems With Applications (2017)* 80: 340-355.

[10] Gou, J. & Xiong, T. 2011. A Novel Weighted Voting for K-Nearest Neighbor Rule. *Journal of Computer*6(5): 833-840.

[11] Pan, Z., Wang, Y. & Ku, W. 2016. A New K-Harmonic Nearest Neighbor Classifer Based On The Multi-Local Means. *Expert Systems With Applications*67: 115-125.

[12] Pan, Z., Wang, Y. & Ku, W. 2017. A New General Nearest Neighbor Classification Based On The Mutual Neighborhood Information. *Knowledge-Based Systems*121: 142-152.

[13] Y Mitani and Y Hamamoto 2006 A local mean-based nonparametric classifier Patern Recognition Letter pp 1151-115.

[14] A A Nababan, O S Sitompul, and Tulus Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio IOP Conf. Series: Journal of Physics: Conf. Series 1007 (2018) 012007.

[15] Danades, A., Pratama, D., Anggraini, D., Anggriani, D. 2016. Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status.*International Conference on System Engineering and Technology*, pp. 137-141.

[16] Zhibin Pan, Yidi Wang, Weiping Ku A new k-harmonic nearest neighbor classifier based on the multi-local means. *Cina Expert Systems With Applications (2016).*

[17] Priyadarsini, R.P., Valarmathi, M.L., Sivakumari, S. 2011. Gain Ratio Based Feature Selection Method For Privacy Preservation.*ICTACT Journal on Soft Computing*1(4): 201-205.

[18] Zuguang HuˈBing XuLingling ˈ Pan, Shangfeng ZhangˈJuying Zeng The Dynamic KNN Clustering of Undergraduate Consumption With Gini Coefficient: A Case of Zhejiang 0-7695-2882-1/07 $25.00 ©2007 IEEE.

[19] Eko Prasetyo.2014, Data Mining-Mengelola Data Mining menjadi Informasi menggunakan Matlab. Penerbit Andi. Ed.I, ISBN : 978-979-29-4351-1.

[20] Wang. J., Neskovic . P. & Cooper L.N., 2007. Improving Nearest Neighbor Rule With A Simple Adaptive Distance Measure. *Pattern Recognition Letter* **28**: 207-213.

[21] Arshad Muhammad Mehar, Kenan Matawie, Anthony Maeder 2013 IEEE. Determining an Optimal Value of K in K-means Clustering 978-1-4799-1310-7/13/$31.00 ©2013 IEEE.

[22] Jianping Gou[1,*], Zhang Yi[2], Lan Du[3] and Taisong Xiong[1] The Computer Journal., Vol.55 No.9, 2012 A Local Mean-Based k-Nearest Centroid Neighbor Classifier doi:10.1093/comjnl/bxr131

[23] Liangping Tu, Huiming Wei and Liya Ai 2015 IEEE. Galaxy and Quasar Classification Based on Local Mean-based k-Nearest Neighbor Method 978-1-4799-7284-5/15/$31.00 ©2015 IEEE.

[24] Nordiana Mukahar, Bakhtiar Affendi Rosdi, ICoAIMS 2017. Interval valued fuzzy sets k-nearest neighbors classifier for finger vein recognition IOP Conf. Series: Journal of Physics: Conf. Series 890 (2017) 012069 . doi :10.1088/1742-6596/890/1/012069.

[25] FAN Xing-ming, HE Jia-min, ZHANG Xin, LIANG Cong, HUANG Zhi-chao, SHI Wei-jian, IEEE 2012. The K Value Determination Research of Advanced Breaking Current Weighted Cumulative Method for VCB Electrical Endurance Detection 978-1-4673-1266-0/12/$31.00 ©2012 IEEE.