SJIF Impact Factor: 6.093 Volume: 4 | Issue: 7 | July | 2019 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Peer Reviewed Journal

SELECTION FEATURES TO IMPROVE THE ACCURACY OF K-NEAREST NEIGHBOR

Y A Pratama[1]
[1]*Graduate School of Computer Science*

Tulus[2]
[2] *Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia*

S Effendi[3]
[3]*Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia*

ABSTRACT

There have been many developments from the kNN method including Local Mean Based, Distance Weight, and Feature Weight. However, most classification methods skip pre-processing step. Pre-processing has a direct effect on the classification results. One pre-processing method is feature selection. Feature selection aims to reduce features that are not relevant to the classification results. Selections of features that are often used are Principal Component Analysis (PCA) and Gain ratio. Testing of several UCI datasets was conducted to determine whether PCA and Gain Ratio method were able to improve the accuracy of the kNN classification. The test results show that feature selection can improve classification performance, that the highest increase in datasets before and after feature selection is 19.15%.

1. INTRODUCTION

There have been many developments from the kNN method, one of which is [1] the Local Mean Based K-Nearest Neighbor (LMKNN) method, where the class determination with the vote majority system is replaced with the local mean system, so that this method can produce a better classification.

One of kNN method is Distance Weight K-Nearest Neighbor (DWKNN) [2]. This method was developed to cover the weaknesses in the soundest systems on conventional kNN, where the system ignores similarities between data. In this DWKNN method, class determination of the new data is based on the class with the highest distance weighting, this method is able to reduce the influence of outliers so as to provide a better classification.

However, of all the classification methods above there is no single method that passes pre-processing. Even though pre-processing is important in data mining. [3] states that data pre-processing is an important main step in the knowledge discovery process, because the data obtained from the log may be incomplete, have outliers / noise and are not consistent.

One step in pre-processing is feature selection. [4] states that feature selection has a direct effect on the classification results. Feature selection is very important in pattern recognition and data analysis. This process aims to select the best features of the initial features and be able to reduce high data dimensions and be able to get out of the problem of curse of dimensionality, so that the classification becomes more accurate [4-8].

There have been many studies on pre-processing before the classification process, including [9], in which the research used the Principal Component Analysis (PCA) method to perform feature selection steps, and then the data was classified using the C4.5 method. This research shows that feature selection with PCA can increase C4.5 classification by 1.2%.

[10] in his study conducted a comparison of Information Gain, Gain Ratio and Information Value in feature selection. Then the data is classified using

the Predictive Approach method. This study showed a better classification than before feature selection.

[11,12] proposed a Gain Ratio to reduce the influence of irrelevant features, the results showed that Gain ratios that were used as attribute weighting bases were able to reduce the influence of features that were not relevant to the classification results, so that the performance of the kNN was better

Based on several studies above, it shows that the feature selection process can improve the performance of classification methods. In this study, PCA and Gain Ratio were considered capable of reducing features that were less relevant. Then the data that has gone through the feature selection process will be made a classification model using kNN and FW-kNN.

2. FEATURE SELECTION

The feature selection process aims to reduce the dimensions of the features in the data. This process is done by selecting features that are relevant to the classification results. Some of the main reasons why feature selection needs to be done are decreasing the learning cost, increasing the learning performance, reducing irrelevant dimensions, reducing redundant dimensions [12].

The feature selection process is expected to reduce the amount of noise and eliminate features that are less relevant so as to improve the performance of the classification process. One method that can be used to view relevant features is PCA and Gain Ratio [9,11].

2.1. PCA

The PCA method is very useful in data that has many characteristics / attributes, PCA calculations are based on eigenvalues and eigenvectors. According to [13], the PCA process, as follows:

Step 1: The covariance of data can be represented in the following equation:

$$\mathbf{Cov}\,(\mathbf{x},\mathbf{y}) = \frac{\mathbf{1}}{\mathbf{n-1}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i}-\boldsymbol{\mu}_\mathbf{x})(\mathbf{y_i}-\boldsymbol{\mu}_\mathbf{y}) \quad (1)$$

μ_x and μ_y is the sample mean of the variables x and y, where x_i and y_i are the values of the i-observation of the variables x and y. From the data which the value is used then obtained covariance n x n.

Step 2: The eigenvalue of covariance matrix can be represented in the following equation [14]:

$$\text{Determinant}\,(\mathrm{A}-\lambda\mathrm{I}) = 0 \quad (2)$$

Step 3: To calculate the size of proportion of Principal Component using the following equation:

$$\boldsymbol{Proportion\ of\ Principal\ Component}\ (\%) = \frac{\boldsymbol{Eigenvalue}}{\boldsymbol{Covariant}} \times \mathbf{100}\% \quad (3)$$

Step 4: A square matrix A is said to have an eigenvalue $\boldsymbol{\lambda}$, with corresponding eigenvector $x \neq 0$, if

$\boldsymbol{Ax = \lambda x}$ [14] (4)

2.2. Gain Ratio

Gain ratio reduce a bias towards multi-valued attributes by taking intrinsic information into account when choosing an attribute [15]. The steps in determining Gain Ratio are as follows:

Step 1: Calculate *Entropy* by using the following equation:

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * log_2 pi \quad (5)$$

Step 2: Calculate *information gain* of each attributes by using the following equation:

$$Information\ Gain\ (S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (6)$$

Step 3: Calculate *Split Information* each attributes by using the following equation:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{D_j}{D}) \quad (7)$$

Step 4: Calculate *Gain Ratio* each attributes by using the following equation:

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{SplitInfo(A)} \quad (8)$$

Gain Ratio is found in the C4.5 algorithm, where the gain ratio is used to calculate the effect of features on the target of a data [16]. Gain Ratio is the development of information gain, where the gain ratio removes the bias value of each feature.

3. k-NN

The algorithm is simple and easy to implement. There's no need to build a model, tune several parameters, or make additional assumptions to solve problems in the case of text categorization, pattern recognition, classification, etc. [17-21]. The algorithm is versatile. It can be used for classification, regression, and search [22-27].

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors [28, 29], kNN value is the average of the values of k nearest neighbor. If the value of k is too small, the number of neighbors to be obtained is also too small. This will not only reduce the classification

accuracy, but also enlarge the disturbing of noise data. While the k value is too large, increased noise will cause lower classification performance [30].

K-Nearest Neightbor (kNN) classification method is described as follow [11]:

Step 1 : Determine parameter K = number of nearest neighbors..

Step 2 : Calculate the distance between the query example and the current example from the data by using *Euclidean* distance, with the following equation:

$$D(x,y) = ||x-y||_2 = \sqrt{\sum_{j=1}^{N} |x-y|^2} \quad (9)$$

Step 3 : Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.

Step 4 : Pick the first K entries from the sorted collection

Step 5 : The majority class is used as a class for test data.

4. Feature Weight k-Nearest Neighbor (FW-kNN)

K-Nearest Neighbor (KNN) is a nonparametric learning method and sensitive to distance function due to the inherent sensitivity of irrelevant attributes. So to get overcome such issues then FWKNN modeling applied which based on attribute weighting. FWKNN determines the weight of the attribute by identifying the nearest k-neighbor which reduce inherent irrelevant attributes in measuring the distance [31].

By providing weight to its attributes, FWKNN makes a distinction to the features, meaning the more significant attributes have a higher impact on distance determination [32]. This can reduce error in the classification method. The detail of FWKNN algorithm is as follows :

Step 1: Determine the weight of each feature

Step 2: Determine the value parameter k

Step 3: Calculate the distance using equation (1)

$$D(x,y) = ||x-y||_2 = \sqrt{\sum_{j=1}^{N} W_j \times |x-y|^2}$$

Step 4: Sorting of ascending results (sequential order from high to low).

Step 5: Based on the k-nearest neighbor, measure the amount of each class.

Step 6: The majority class is used as a class for test data

5. PROPOSED METHOD

This study will perform the selection feature stages in order to obtain information characteristics suitable to the results, so that the quality of the classification techniques used is expected to be improved. Figure 1 shows the phases in this study.

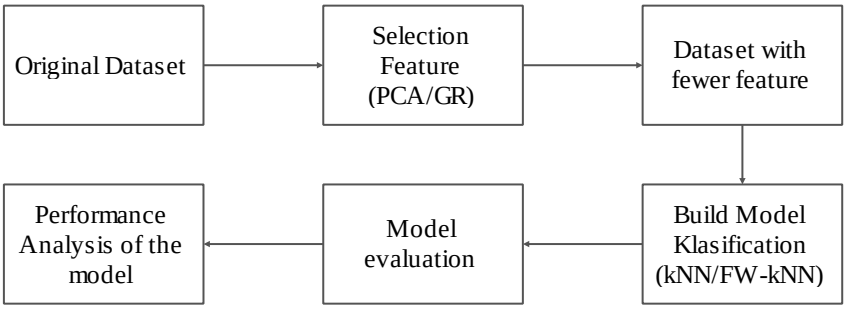

Figure 1. Classification Process

The phases of this study process can be described based on Figure 1:

- Dataset where the information acquired from the UCI Machine Learning Repository will be used in this study.
- Selection feature, this phase will create selection of characteristics using PCA and Gain Ratio, where irrelevant characteristics on regular kNN are eliminated while the irrelevant feature is given a weight of 0 in FW-kNN. The feature is taken of 0.9 (90 percent) proportional information.
- Use kNN and FW-kNN to create a classification
- Use the 10-Fold Cross Validation method at this stage to evaluate the model.
- Analyzing the performance of the model produced, comparing the accuracy of the kNN technique, FW-kNN with the accuracy results after selecting the feature with the parameter k value 1 to 10.

6. RESULT AND DISCUSSION

The research will use multiple datasets from UCI Machine Learning Repositories, including hareman, hayes-roth, glass, and water quality, to get a clear picture of the analysis performed. Table 1 shows the information of pre and post selection of the feature.

Table 1 Details of data used

No	Dataset	Fitur			Type	Classification	Total Data
		Original	PCA	Gain Ratio			
1	*Haberman*	*3*	*2*	*2*	*Integer, Binominal*	2	306
2	*Hayes-roth*	*4*	*3*	*3*	*Real, Integer, Nominal*	3	160
3	*Glass*	*9*	*9*	*9*	*Real, Integer*	6	214
4	*Water Quality Status*	*8*	*7*	*7*	*Real, Integer*	4	210

Table 1 shows that PCA and GR selection of features generates the same results. Where both of these methods generate the same number of new features, but there is no decrease in the feature selection number of new features of the two techniques for the glass data set.

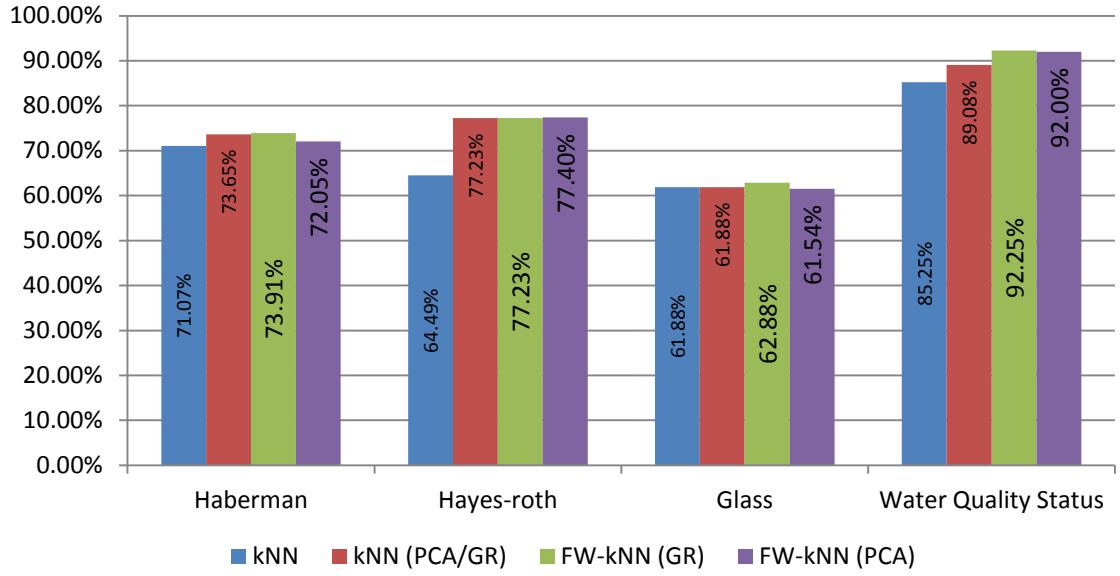

Figure 2. Average Accuracy of All Data

Figure 2 shows that the highest level of accuracy for the original kNN method with kNN that went through the feature selection process (PCA / GR) on the hayes-roth dataset is 12.74 percent, which also occurs in the original kNN and FW-kNN (GR), but FW-kNN (GR) does not experience an increase in kNN (PCA / GR). The maximum increase for FW-kNN (PCA) was 12.91 percent in hayes-roth information.

Overall, the original kNN (PCA / GR) increased by 19.15%, while FW-kNN (GR) increased by 23.58% and kNN (PCA) increased by 20.30%. From all the methods that have passed through the test stage, it can be seen that the selection process can enhance the accuracy of the original kNN method, while the weighting method has proven to be better in the Gain Ratio.

7. CONCLUSION

Based on the findings and discussion in the previous section it can be concluded that:

1. This study shows that the gain ratio in the weighting method is better than the principal component analysis
2. The feature selection process has been shown to be able to enhance the accuracy of the kNN classification method, as shown by the increase in the accuracy of the entire dataset used. The increase in average accuracy for the kNN method with feature selection is 19.15 %, while FW-kNN (GR) is 23.58 %, and FW-kNN (PCA) is 20.30 %

REFERENCES

1. *Mitani, Y. & Hamamoto, Y. 2006. A Local Mean-Based Nonparametric Classifier. Patern Recognition Letter* ***27****(10): 1151-1159.*
2. *Batista, G.E.A.P.A. & Silva, D.F. 2009. How k-Nearest Neighbor Parameters Affect its Performance. 38° JAIIO - Simposio Argentino de Inteligencia Artificial (ASAI 2009), pp. 95-106*
3. *Samsani, S. 2016. An RST based Efficient Preprocessing Technique for Handling Inconsistent Data. 2016 IEEE International Conference on Computational Intelligence and Computing Research, 1-6*
4. *Zhang, X., Shi, Z., Liu, X. & Li, X. 2018. A Hybrid Feature Selection Algorithm For Classification Unbalanced Data Preprocessing. International Conference on Smart Internet of Things, 1-6.*
5. *Mohana, C. P. & Perumal K. 2017. A Comparative Analysis of Feature Selection Stability Measure.*

Internationa Conference on Trends in Electronics and Informatics, 1-6

6. *Pokhriyal, N. & Verma, S.K. 2016. Statistical Feature Extraction/Selection for Small Infrared Target. International Conference on Advances in Computing, Communications and Informatics, 1-6*
7. *Gupta S, et al. 2016. Class Wise Optimal Feature Selection For Land Cover Classification Using Sar Data. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 68-71*
8. *Peker. M, et al. 2015. A Novel Hybrid Method for Determining the Depth of Anesthesia Level: Combining ReliefF Feature Selection and Random Forest Algorithm (ReliefF+RF). 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), 1-6*
9. *Nasution M. Z. N., Sitompul O. S, & Ramli, M. 2018. Pca Based Feature Reduction To Improve The Accuracy Of Decision Tree C4.5 Classification. 2nd International Conference on Computing and Applied Informatics 2017, 1-6*
10. *Duneja, A., Puyalnithi, T. 2017. Enhancing Classification Accuracy of K-Nearest Neighbours Algorithm Using Gain Ratio. International Research Journal of Engineering and Technology (IRJET) **4**(9): 1385-1388*
11. *Nababan A. A., Sitompul O. S, & Tulus. 2018. Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio. MECnIT. 1-6*
12. *Mainmon, O. & Rokachi, L. 2010. Data Mining And Knowledge Discovery Handbook. Springer : New York*
13. *Jolliffe I T 2002 Principal Component Analysis 2nd Ed. (New York: Springer-Verlag)*
14. *Johnson R A and Wichern D W 2007 Applied Multivariate Statistical Analysis 6th Ed. (New Jersey: Pearson Prentice Hall)*
15. *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*
16. *Mitchell, T.M. 1997. Machine Learning. McGraw-Hill Science/Engineering/Math; (March 1, 1997).ISBN: 0070428077*
17. *Bhatia, N. & Vandana., 2010. Survey of Nearest Neighbor Techniques. International Jurnal of Computer Science and Information Security (IJCSIS) **8**(2): 302-305.*
18. *Jabbar, M.A., Deekshatulu, B.L. & Chandra. P. 2013. Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), pp. 85-94.*
19. *Rui-Jia, W. & Xing, W., 2014. Radar Emitter Recognition in Airborne RWR/ESM Based on Improved K Nearest Neighbor Algorithm. 2014 IEEE International Conference on Computer and Information Technology (CIT), pp. 148-151.*
20. *Sánchez, A.S., Iglesias-Rodríguez, F.J., Fernándes, P.R. & Juez, F.J.de.C. 2015. Applying The K-Nearest Neighbor Technique To The Classification Of Workers According To Their Risk Of Suffering Musculoskeletal Disorders. International Journal of Indsutrial Ergonomics **52**: 92-99.*
21. *Zheng, K. Si, G. Diao, L. Zhou, Z. Chen, J. & Yue W., 2017. Applications Of Support Vector Machine And Improvedk-Nearest Neighbor Algorithm In Fault Diagnosis And fault Degree Evaluation Of Gas Insulated Switchgear. 1st International Conference on Electrical Materials and Power Equipment - Xi'an – China, pp. 364-368.*
22. *Wang. J., Neskovic . P. & Cooper L.N., 2007. Improving Nearest Neighbor Rule With A Simple Adaptive Distance Measure. Pattern Recognition Letter **28**: 207-213.*
23. *García-Pedrajas, N. & Ortiz-Boyer, D. 2009. Boosting K-Nearest Neighbor Classifier By Means Of Input Space Projection. Expert System With Application **37**(7): pp.10570-10582.*
24. *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.*
25. *Ougiaroglou, S. & Evangelidis, G. 2012. Fast and Accuratek-Nearest Neighbor Classification using Prototype Selection by Clustering. Panhellenic Conference on Informatics, pp. 168-173.*
26. *Song, Y., Liang, J., Lu, J. & Zhao, X. 2017. An Efficient Instance Selection Algorithm For K Nearest Neighbor Regression. Neurocomputing **251**: 26-34.*
27. *Feng, Y., Jian-Chang, L. & Dong-ming L. 2016. An Approach for Fault Diagnosis Based on an Improved k-Nearest Neighbor Algorithm. Control Conference (CCC), 2016 35th Chinese, pp. 6521-6525.*
28. *Kalaivani, P. & Shunmuganathan, K.L. 2014. An Improved K-Nearest-Neighbor Algorithm Usinggenetic Algorithm For Sentiment Classification. 2014 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1647-1651.*
29. *Iswarya, P. & Radha, V. 2015. Ensemble learning approach in Improved K Nearest Neighbor algorithm for Text Categorization. 2015 International Conference on*
30. *Gou, J., Zhang, Y., Rao, Y., Shen, X., Wang, X. & He, W. 2014. Improved Pseudo Nearest Neighbor Classification. Knowledge-Based Systems **70**: 361-375.*
31. *Chen et al. 2018. A Feature Selection Approach for NetworkIntrusion Detection Based on Tree-SeedAlgorithm and K-Nearest Neighbor. The 4th IEEE International Symposium on Wireless Systems within the In ternational Conferences on Intelligent Data Acquisition and Advanced Computing Systems, pp. 68-72.*
32. *Kuhkan, M. 2016. A Method to Improve the Accuracy of K - Nearest Neighbor Algorithm. Internatonal Journal of Computer Engineering and Information Technology **8**(6): 90-95.*