Journal of Information Technology and Computer Science
Volume 5, Number 1, April 2020, pp. 57-64
Journal Homepage: www.jitecs.ub.ac.id

Nearest Centroid Classifier with Centroid-Based Outlier Removal for Classification

Aditya Hari Bawono[1], Fitra Abdurrahman Bachtiar[2], Ahmad Afif Supianto[3],

[1,2,3]Fakultas Ilmu Komputer, Universitas Brawijaya
[1]ndaimen@ub.ac.id, [2]fitra@ub.ac.id, [3]afif.supianto@ub.ac.id

Received 28 November 2019; accepted 20 February 2020

Abstract. Classification is a part of data mining which is very useful for many purpose. There are so many method to classify, however a classification process need good source of data to deliver good performance. In the real world, not all data is good. Some of the data contains noise and outlier that can reduce classification performance. However, there are few research of classification with an outlier removal process, especially for Nearest Centroid Classifier Method. Nearest Centroid Classifier is a good yet simple classification method. However, Nearest Centroid Classifier suffers performance loss like the other classification method. The proposed methodology consists of two stages. First, preprocess the data with outlier removal, removes points which are far from the corresponding centroids, the center point of a class. Second, classify the outlier removed data or cleaned data. The experiment covers six data sets which have different characteristic. The results indicate that outlier removal as preprocessing method provide better result by 0.76% to 3.16% for improving Nearest Centroid Classifier performance on most data set.

Keywords : Classification, Outlier, Centroid Based, Preprocessing

1. Introduction

Classification is a part of data processing that has attracted the attention of academics and industry. Classification is an effort to label data based on data that has been previously labeled [1]. In the real world, classification is used to distinguish one object from another. Factors that influence the results of predictions are the data used, preprocessing, and the selection of classification methods [2]. However, there are so many real world data, but they are not always good. Choosing which method to process those data require extra effort, because bad data usually harder to process.

One classification method is the Nearest Centroid Classifier (NCC) [3]. NCC is a centroid-based classification method [4]. The advantage of NCC is that the method does not have parameters, so the results obtained do not depend on parameter configuration, but on data distribution and distance. In the NCC method, the distance between data is used to determine the closest centroid point to the test data point. The advantage of NCC is the simplicity of the process and does not require parameter tuning.

NCC has been used in many field for classification purpose. Several studies using NCC methods are [5], which perform feature selection as preprocessing, [6] used NCC for DNA classification, [7] is about vehicle routing problems, [8] use a centroid-based classifier for text classification, [9] use a centroid-based classifier for intrusion detection, and [4] are used for music genre classification. In related studies [4][5][6][7][8][9], there are no data cleaning process as a preprocessing for the NCC method. Though it could be that the data has missing values and outliers.

In addition, the method does not perform well in certain data sets, especially data sets with outlier [10]. Outlier is an observation result that is different in nature with a majority of other observations [11]. Among comparative studies, it is stated that the existence of outliers can reduce the accuracy of prediction of an algorithm [12][13]. For centroid-based method, outlier can influence how centroid are formed [14], thus non-ideal centroid will be formed and influence the classification process.

This study tries to implement NCC with preprocessing step. The preprocessing step conducted in this study is removing outlier first. The outlier removal method used in this study is centroid based outlier removal. The cleaned data is used for the classification after the outlier observation removed. Centroid-based outlier removal is one of the suitable outlier removals for classification [15]. Further, the experiment uses the Outlier Detection Data sets (ODDS) [16]. The problem is, Outlier Detection Data sets (ODDS) have a list that in UCI data sets there are at least 0.03% to 35% outliers. While the UCI data set itself is a standard data set commonly used in classification. This can be of particular concern regarding the removal of outliers as preprocessing. The focus of this research is to improve the results of the NCC classification by conducting outlier removal as preprocessing. The results of this study are expected to be used as a reference for further research.

The organization of this paper is as follows. First, background study about outlier, outlier effects for classification, how to handle outlier, and classification. Next, proposed methods are structured into preprocessing with outlier removal and classification using NCC. After that, we compare the result between proposed method and standard NCC. The purpose of this study is finding the difference between the result of standard method and proposed method which is using outlier removal as preprocessing method.

2. Centroid-Based Outlier Removal

Centroid-based outlier detection method defines the outlier as a point far from any cluster or data point group. One of the experiment have been done by [15]. Centroid based outlier detection method calculates distance between points and centroid of corresponding class label, then compare the distance of each centroid. The pseudocode of the Centroid Distance Outlier Removal is as follows:

Centroid Distance Outlier Removal

```
Input   : $X_n$ = data matrix, $C_n$ = class labels of $X_n$;
Output  : $K_m$ = cluster of each class labels, $O$ = outlier set;
BEGIN
1.        Calculate each $X_n$ with corresponding class label $C_m$ to form $K_m$
2.        Calculate distance between $X_n$ and $K_m$
3.        If the nearest centroid of ($X_n$, $K_m$) != $C_n$, consider that point as an outlier $O$
4.        Else the point is not an outlier.
END
```

The difference between the proposed method with [15] is the current study do not use any parameter such as *k*-centroid because the data label already known in advance. Thus, using parameter *k*-centroid is not necessary.

3. NCC with Centroid-Based Outlier Removal

In general the steps taken in this study can be seen in Figure 1.

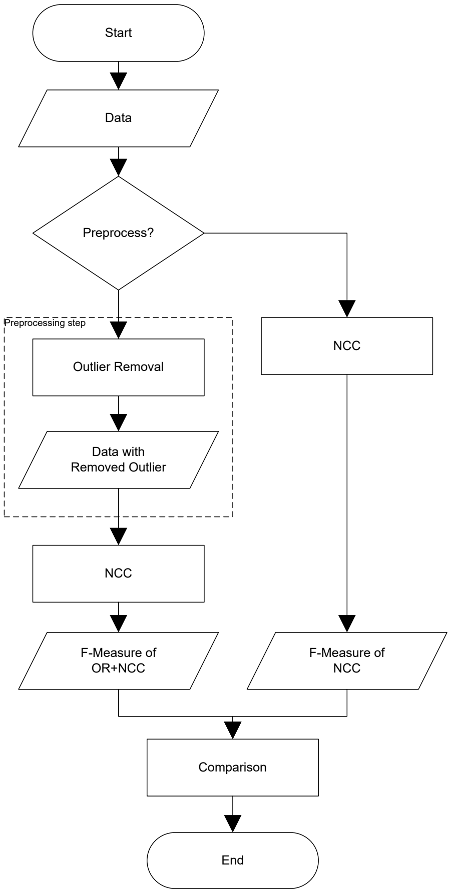

Figure 1. NCC with Outlier Removal

The study begins by specifying data set that are going to be used. Next, we split the process into two ways. The first is classify the data without preprocessing, straight away with NCC. Meanwhile, the other is preprocessing the data with outlier removal first, then classify with NCC. After classification process, evaluation phase with F-Measure is conducted. Lastly, the results are compared with classification without removing the outliers to see which methods perform better.

3.1 Data set

Six UCI standard data sets is used. The properties of the used data sets are also varies. The properties of the data set include the number of data points, the number of features, the number of classes, and the percentage of outliers. Detailed characteristics of the data set will be displayed in the Table 1.

Table 1. Data set List

Data set	**Data Point**	**Feature**	**Class**	**% Outlier**
Breast Cancer	679	9	2	35
Ecoli	336	7	8	2.6
Lymphography	148	18	4	4.1
Vertebral	240	6	3	12.5
Wine	178	13	3	7.7
Yeast	1364	8	10	4.7

In terms of data point properties, Yeast data set has the most data points of 1364 and Lymphography data set has the least data set of 148. However, for feature properties, Lymphography has the most feature of 18, with Vertebral data set has the least. Breast cancer has the least class number, while Yeast has the most classes. However, Breast cancer has the most outlier percentage. It can be seen that the chosen data sets have varying properties.

3.2 Nearest Centroid Classifier (NCC)

NCC is a centroid-based classification method (Praveen et al., 2016). The advantage of NCC is that it does not require parameters, so the results are not determined by parameter configuration, but purely on the distance between data. Euclidean distance is the distance formula used in this study. The role of distance in the NCC method is used to determine the centroid center closest to the data point.

Nearest Centroid Classifier

```
Input   : $X_n$ = training data, $C_n$ = class labels of $X_n$, $Y_m$ = testing data;
Output  : $L_m$ = predicted label of $m$;
BEGIN
Repeat
1. Calculate mean of each data points $X_n$ with corresponding class label $C_n$
2. Take mean of each Class label $C_n$, keep that value to centroid $Z_k$.
Until all centroid is calculated
Repeat
3. Calculate distance of each testing data points $Y_m$ to each centroid $Z_k$, keep the distance to $D(Y_m, Z_k)$
4. Compare distance of data point $Y_m$ and centroid $Z_k$, determine the nearest centroid, then keep the nearest centroid to $L_m$
5. Until all testing data is calculated.
END
```

However, only a few NCC studies have attempted to improve the performance of this method. One of them is feature selection [5] which is proven to improve NCC performance. This study uses preprocessing outlier removal with the aim of improving NCC performance.

3.3 Evaluation metric

Three evaluation method is used to understand the result of the proposed method that is Precision, Recall, and F-Measure. Precision is symbolized as the ratio of predictions taken that is relevant to the search.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad (1)$$

True Positive (TP) is a case when prediction is correctly predicted by proposed method. False Positive (FP) is a case when proposed method incorrectly predicts positive class. False Negative (FN) is a case when proposed method incorrectly predicts negative class. Recall is also known as the fraction of the relevant example that has been taken from the total number of instances taken.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad (2)$$

F-measure includes Precision and Recall. This can be considered as the harmonic mean of the two values.

$$\boldsymbol{F-Measure} = \frac{\boldsymbol{2*Precision*Recall}}{\boldsymbol{Precision+Recall}} \qquad (3)$$

Not only accuracy, but computational time also will be the focus of this research. Preprocessing will always require additional time, while the whole process will be better if it is fast and accurate. Milliseconds (ms) will be used as unit metric. Further, correlation between number of feature and time needed for each experiment is evaluated to know the influence of number of feature to the running time.

4. Experimental Result

First, the experiment is carried out by performing classification method without removing the outlier on the six data sets. The results of each experiment are recorded. The second step, the classification is repeated but the outlier removal conducted prior to classification step. *F-measure* will be displayed in percentage, while computation time is shown in milliseconds unit. The number of outlier observation removed is also displayed, although the number of outlier will be vary for each data set. The classification results will be displayed at the Table 2.

Table 2. Experimental Result of NCC before and After Preprocessing

Data set	NCC		Number of Outlier	Data Point Removed	+NCC	
	F (%)	Time (ms)			F (%)	Time(ms)
BCW	94.59	1.0441	239	39	95.67	2.3415
Ecoli	61.94	1.1521	9	27	62.70	3.1803
Lymphography	60.40	0.4716	6	4	61.62	2.7644
Vertebral	73.40	0.4456	30	26	73.59	1.4691
Wine	94.15	0.3299	10	6	97.76	1.1425
Yeast	54.27	5.0629	64	114	56.31	15.8266

It can be seen from the Table 2 in the Breast cancer data set, the classification result after preprocessing tend to be better at around 1.16%. The results are also similar with other data sets. Experiments on the Ecoli data set increased the F-measure by 0.76%, an increase in the Lymphography data set by 1.22%, an increase in the vertebral data set by 0.19%, an increase in the Wine data set by 3.16%, and an increase in the Yeast data set by 2.04. It can be concluded that for the F-measure evaluation, preprocessing can help improve performance by an average of 1.48%. The experiment shows that the results yield in good results. The differences between classification with and without outlier removal can be seen in the Figure 2.

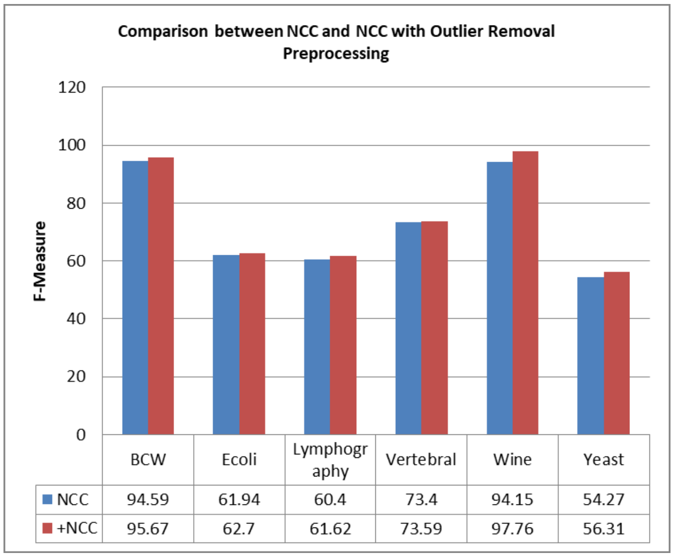

Figure 2. Comparison Graphic of Proposed Method

Preprocessing steps takes time in the classification process. After the time evaluation has been evaluated, the experiment shows that there is an increase, although not significant. Preprocessing in the Breast cancer data set requires an additional time of 1.29 ms, the Ecoli data set requires an additional time of 2,202 ms, the Lymphography data set requires an additional time of 2.76 ms, the Vertebral data set requires an additional time of 1,023 ms, the Wine data set requires additional time of 0.812 ms, and the Yeast data set requires an additional time of 10,763 ms. Yeast data sets require the most time due to larger data sizes than others. The comparison of matrix size to preprocessing time is shown in the Table 3.

Table 3. Data Matrix and Time Difference

Data set	Data Point	Feature	Data Matrix	Time Difference
Vertebral	240	6	1440	1.0235
Wine	178	13	2314	0.8126
Ecoli	336	7	2352	2.0282
Lymphography	148	18	2664	2.2928
Breast Cancer	679	9	6111	1.2974
Yeast	1364	8	10912	10.7637
Correlation			0.884	

The table shows that the greater the data matrix, the greater the time it takes. if a correlation is calculated between the size of the data matrix and time, a positive correlation of 0.884 will be found. This shows that the efficiency of the method should be done especially on data that is getting bigger.

5. Conclusion

In this study an implementation of NCC with outlier removal as preprocessing method is conducted successfully. First, we calculate centroid of each class. After that, for every training data points, we compare the distance of each centroid. Any data points that are nearer with other class are considered as outlier. After that we calculate the centroids of each class once again to form classifier. The classification process of NCC are comparing test data points with each centroid. Then the nearest centroid will considered as related class. After the classification process, evaluate the result with F-measure and measure the time consumed. Finally, the result are obtained and analysis can be conducted.

After conducting the experiment, it shows that preprocessing with outlier removal can increase NCC classification performance. Even with simple preprocessing method, the proposed method can increase up to 3.16% F-Measure. However, it is true that two stages classification process is taking time considerably. The larger the data size, the longer the processing time. This is evidenced by positive correlation between the size of the data matrix with a time difference between methods of 0.884.

There are several directions that could expand this study. Future works is expected to cut computational time while increasing the performance of the proposed method. Wrapper method by combining the preprocessing step and classification might decrease the time needed for the computation.

References

1. S. Ougiaroglou and G. Evangelidis, "Dealing with noisy data in the context of k-NN Classification," *Proc. 7th Balk. Conf. Informatics Conf. - BCI '15*, pp. 1–4, (2015).
2. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. (2012).
3. J. C. Bezdek and L. I. Kuncheva, "Nearest prototype classifier designs: An experimental study," vol. 16, no. 12, pp. 1445–1473, *Int. J. Intell. Syst.*, (2001).
4. E. N. Tamatjita and A. W. Mahastama, "Comparison of music genre classification using Nearest Centroid Classifier and k-Nearest Neighbours," *Proc. 2016 Int. Conf. Inf. Manag. Technol.* no. November, pp. 118–123, *ICIMTech 2016*, (2017).
5. R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays," *Stat. Sci.*, vol. 18, no. 1, pp. 104–117, (2003).
6. A. R. Dabney, "Gene expression Classification of microarrays to nearest centroids," vol. 21, no. 22, pp. 4148–4154, (2005).
7. V. Praveen, K. Kousalya, and K. R. P. Kumar, "A nearest centroid classifier based clustering algorithm for solving vehicle routing problem," *Proceeding IEEE - 2nd Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics,* pp. 414–419, *IEEE - AEEICB 2016*, (2016).
8. C. Liu, W. Wang, G. Tu, Y. Xiang, S. Wang, and F. Lv, "A new Centroid-Based Classification model for text categorization," , vol. 136, pp. 15–26, *Knowledge-Based Syst.*, (2017).
9. B. Setiawan, S. Djanali, and T. Ahmad, "A Study on Intrusion Detection Using Centroid-Based Classification," vol. 124, pp. 672–681, *Procedia Comput. Sci.*, , (2017).

10. S. Mehta, X. Shen, J. Gou, and D. Niu, “A New Nearest Centroid Neighbor Classifier Based on K Local Means Using Harmonic Mean Distance,”(2018).
11. D. Hawkins, *Identification of Outliers*. (1980).
12. G. M. Foody, “The Effect Of Mis-Labeled Training Data On The Accuracy Of Supervised Image Classification By Svm Giles M . Foody,” *2015 ,* pp. 4987–4990, *IEEE Int. Geosci. Remote Sens. Symp.*, (2015).
13. C. Pelletier, S. Valero, J. Inglada, and G. Dedieu, “New Iterative Learning Strategy To Improve Classification Systems By Using Outlier Detection Techniques C . Pelletier , S . Valero , J . Inglada , G . Dedieu CESBIO - UMR 5126 18 avenue Edouard Belin 31401 Toulouse CEDEX 9 - FRANCE IGN Espace - MATIS / Un,” *IGARSS*, p. 3676, (2017).
14. V. R. Patel and R. G. Mehta, “Impact of outlier removal and normalization approach in modified k-means clustering algorithm,” vol. 8, no. 5, pp. 331–336, *IJCSI Int. J. Comput. Sci. Issues*, (2011).
15. X. Wang, Y. Chen, and X. L. Wang, “A Centroid-Based Outlier Detection Method,” *Proc. - 2017 Int. Conf. Comput. Sci. Comput. Intell.* pp. 1411–1416, *CSCI 2017*, (2018).
16. S. Rayana, “Outlier Detection DataSets,” *ODDS Library*, 2016. [Online]. Available: http://odds.cs.stonybrook.edu.