Correlation Coefficient-based K-means Clustering for K-NN

Swe Swe Aung, Itaru Nagayama, Shiro Tamaki
Information Engineering, University of the Ryukyus, Okinawa, Japan
sweswe@ie.u-ryukyu.ac.jp,nagayama@ie.u-ryukyu.ac.jp,shiro@ie.u-ryukyu.ac.jp

Abstract

K-nearest neighbor algorithm is one of the most popular classifications in machine learning zone. However, as k-nearest neighbor is a lazy learning method, when a system bases on huge amount of history data, it faces processing performance degradation. Many researchers usually care about only classification accuracy, but the speed of estimation also play an essential role in real time prediction systems. For this issue, this research proposes correlation coefficient-based k-mean clustering for k-nearest neighbor aiming at upgrading the performance of k-nearest neighbor classification by improving processing time performance. For the experiments, we used the real data sets, Breast Cancer, Breast Tissue and Iris, from UCI machine learning repository. Moreover, the real traffic data collected from Ojana junction, Route 58, Okinawa, Japan, was also utilized to show the efficiency of this method. By using these datasets, we prove the better processing performance and prediction accuracy of the new approach by comparing the classical k-nearest neighbor with the new k-nearest neighbor.

1. Introduction

Nowadays, as there is huge amount of data are able to get straightforwardly in many areas, medical, biology, and transportation, classification of member of predefined class is becoming an important role in different kinds of ways. The k-nearest neighbor is a semi-supervised learning algorithm such that it requires training data and predefined *k* value to find the *k* nearest data based on distance computation [7]. While nearest neighbor can learn in the presence of irrelevant information, it requires more training data to do so and the amount of training data needed to reach or maintain a given accuracy level. Furthermore, nearest neighbor is slow to execute due to the fact example to be classified must be compared to each of the stored training cases in turn [6].

To overcome this issue, this paper introduces a preprocessing stage before coming classification step. In this preprocessing stage, the data which belong to the same characteristic are grouped together by utilizing clustering technique, k-mean clustering. K-means is a typical clustering algorithm. It is attractive in practice, because it is simple and it is generally very fast. It partitions the input dataset into *k* clusters. Each cluster is represented by adaptively-changing centroid, starting from some initial values named seed-points [3]. However, from our experimentation experience, if there is a weak correlation among attributes, it makes wrong decision in grouping the data into the same category. To select the features which have the strong enough correlation each other, this paper applies *Pearson's Correlation* to calculate the relationship among features. Correlation is positive when the values increase together, but correlation is negative as the other increase.

The rest of this paper is organized as follows. Section 2 describes related works. Section 3 presents proposed approach. Section 4 details

with experimentation of new models on four datasets and compares the processing time performance of the new approach with the classical k-nearest neighbor. Finally, section 5 is the conclusion and discusses the future works.

2. Related Works

Machine learning is the study of algorithms that automatically predict the future event based on past experiences as well as improve their prediction performance with those experiences. The more the data they have, the more the accurate estimation they figure out. However, from the point of processing time performance, they need to set much more time on processing large amount of voluminous data.

In reference [1], the author proposed a system which extended nearest neighbor method for pattern recognition considered not only who are the nearest neighbors of the test sample, but also who considered the test sample as their nearest neighbors. By iteratively assuming all the possible class memberships of a test sample, the *ENN* is able to learn from the global distribution, therefore it improved pattern reorganization performance. Reference [4] also designed a system which improved k-nearest neighbor's prediction performance by applying a density peaks clustering and principal component analysis to overcome the generation of wrong number of clusters of real-world data sets. Reference [5] proposed a system which predicted student performance by utilizing correlation based feature selection technique. They did experimentation for their system by using *NBTree*, *Multilayer Perceptron*, *Naïve Bayes* and *Instance Based-K-nearest* neighbor. In reference [7], the authors designed a system which predicted the short-term traffic flow by applying a k-nearest neighbor model and utilizing the Shanghai urban expressway section measured traffic flow data. The reference [2] studied the performance of k-nearest neighbor classification by applying different distance measurements, *Euclidean*, *Standardized Euclidean*, *Mahaloanobis*, *City block*, *Minkowski*, *Chebychev*, *Cosine*, *Correlation*, *Mamming*, *Jaccard*, and *Spearman*.

The systems described above only emphasized on the accuracy in prediction performance of k-nearest neighbor. To get a complete performance, the processing time is also vital in machine learning area. This paper considers not only in improving accuracy, but also upgrading prediction time performance by using correlation coefficient-based clustering for k-nearest neighbor.

3. Proposed Approach

K-nearest neighbor algorithm is a supervised lazy classifier which has local heuristics. For each observation instance, it computes the distance for observed instance to each training data in space *Q* to find the closest distanced. Because of that, k-nearest neighbor comes to be lazy. In general, the time complexity of k-nearest neighbor classifier in *Big Oh* notation is n^2 where n is the number of training example. Therefore, when the amount of data size increases, classical k-nearest neighbor usually becomes slow computation. Finally, this lazy computation significantly kills the performance of k-nearest neighbor. For this issue, our proposed approach aims at improving processing time of k-nearest neighbor by introducing correlation coefficient-based k-means clustering approach as shown in Figure 1.

Figure 1 illustrates the processing flow of proposed approach. In this figure, a preprocessing stage is introduced for the purpose of improving processing time performance. In the preprocessing stage, before clustering, removing noisy features process comes first and then

grouping the same members is carried out by using k-means clustering algorithm.

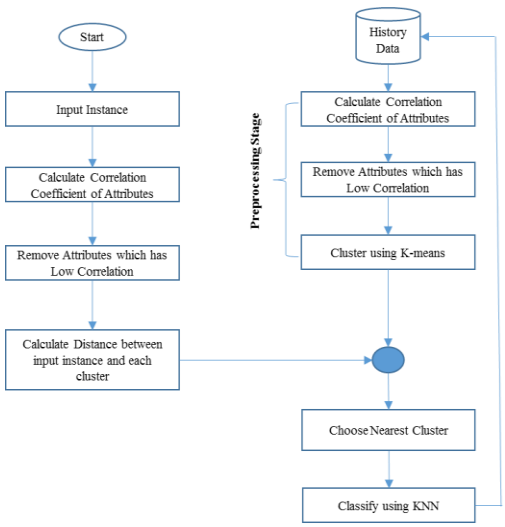

Figure 1. Flow Diagram of Proposed Approach

For filtering noisy features, we measure the strength of relationship among attributes for the reason why the attributes which have a weak relationship makes the algorithm wrong decision. For this issue, this paper utilizes *Pearson's Correlation* for measuring the relationship among attributes described as follow:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

where x and y are the random variable, r_{xy} means the correlation coefficient of variable x and y, and $\bar{x}$ and $\bar{y}$ define the mean values of x and y respectively.

After measuring the strength of relationship among features, the system selects the features which has strong enough correlation coefficient. In this paper, we define 0.30 correlation coefficient as a standard relation based on our experiments. Table 1 shows the correlation coefficient of four data sets, *Breast Cancer*, *Traffic data*, *Breast Tissue* and Iris. *Brest Cancer* has 9 attributes, but all 9 attributes are active because the values of attributes are over 0.30. In Traffic data, it has 7 attributes, but one attribute, *SpecialDay*, is not active because the correlation coefficient is under 0.30. In the case of *Breast Tissue*, there are 9 attributes, but 7 attributes are available. For Iris, only one attribute is not active.

After filtering attributes which have weak relationship, next step is grouping members into the same category. In this paper, we utilize k-means clustering technique. K-means clustering is one of the famous supervised learning algorithm and an easy way to cluster by defining the centric of predefined cluster for a given data set. Firstly, it defines the centric of each cluster by calculating the mean of same objectives. The next step is to yield each point, object, to the nearest cluster according to the distance vector space. For measuring the distance between two points on a straight line, x and y, *Euclidean* distance formula is utilized. The distances for N-dimensional space are as follow:

$$(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Where $d(x,y)$ is the distance of object x and y.

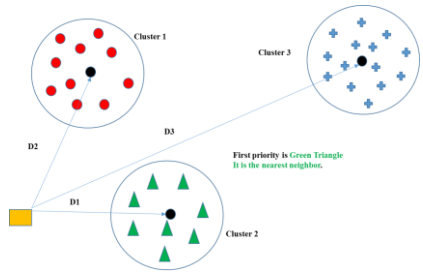

Figure 2. Diagram for Finding the Closest Cluster

After preprocessing stage, the objects are clustered into the same counterpart as shown in Figure 2. Figure 2 represents the condition of after preprocessing stage. In this figure, the observed instance, the yellow rectangle, has to be classified which cluster belongs to. For classical k-nearest neighbor, it has to measure the distance between

Table 1. Correlation Coefficients (CC) of Four Datasets

BreastCancer		*Traffic Data*		*Breast Tissue*		*Iris*	
Attribute Name	*CC of x and y*	*Attribute Name*	*CC of x and y*	*Attribute Name*	*CC of x and y*	*Attribute Name*	*CC of x and y*
Clump	0.71	RushHourTi	0.30	I0	0.77	SepalLeng	0.82
Uniformity ofCell	0.81	Amountof Car	0.99	PA500	-0.64	SepalWidth	-0.49
Uniformit	0.81	CurrentTime	0.30	HFS	-0.15	PetalLeng	0.96
Marginal	0.68	KindofDay	0.30	DA	0.52	PetalWidt	0.96
Single	0.68	SpecialDay	-10.17	Area	0.32		
Bare	0.81	WeatherCond	0.30	A/DA	0.30		
Bland	0.75			Max IP	0.53		
Normal	0.71			DR	0.46		
Mitoses	0.42			P	0.74		

Algorithm 1:

$X = \{x_1, x_2, ..., xi\}$ where x_i are training examples
$Y = \{y_1, y_2, \ldots, y_i\}$ where y_i are input instances

1) ***Preprocessing Stage***
   a) For each data point x_i:
   Calculate correlation coefficient by using equation 1 and remove the attribute which has low correlation
   b)Cluster points in X into the same group by applying k-means algorithm
2) ***Classification Stage***
   For each input instance y_i:
   a)Calculate correlation coefficient of y_i by using equation 1
   b)Remove attributes which has low correlation
   c)calculate distance between input instance y_i and cluster C_i
   d)choose the nearest cluster C according to distance vector $D=\{ d_1, d_2, \ldots, d_i\}$
   e)classify y_i using k-nearest neighbor
3) Store the final result in database for future use and go to step 2.

the observed instance and all training data throughout database for finding the nearest object. However, in the new approach, firstly it finds the closest cluster by measuring the distance between the observed instance and the centric of each cluster. In Figure 2, it is obvious that the observed instance belongs to *Cluster 2* because the distance between the observed instance and *Cluster 2* is the shortest one. After that, it is the turn of k-nearest neighbor and continues to classify only the members of *Cluster 2*, not need to classify all the training examples of dataset. The detail processing steps are shown in ***Algorithm 1***.

4. Analytical Results

This section is going to demonstrate and prove the efficiency of this method, correlation coefficient-based k-means clustering for k-nearest neighbor on four datasets, *TrafficData*, *BreastCancer*, *BreastTissue* and Iris. Besides, we compare the processing performance of the new k-nearest neighbor with the classical one. The prediction accuracy is measured by using k-fold cross validation.

Figure 3,4,5, and 6 show the dataset of *TrafficData*, *Breast Cancer*, *Breast Tissue* and *Iris* respectively before applying correlation coefficient. This paper also uses *Principle Component Analysis* for analyzing the relationship among objects and visualization purpose. In Figure 3, it has three states, *Green*, *Yellow* and *Red*. That shows the conditions of traffic. *Green* means no traffic, *Yellows* is awareness situation and *Red* shows that there is a

heavy traffic jam. In Figure 4, it can be seen that there are two states, *Absent* and *Present*. Red is no breast cancer and blue means a person has a cancer. Figure 5 is about breast tissue. It has 6 conditions. *Red* is car, pink means *Fad*, yellow defines *Mas*, blue is *Gla*, green is *Con* and purple presents Adi. Figure 6 talks about *Iris* Dataset which has three conditions. Those are Iris-setosa, Iris-versicolor, and Iris-virginica.

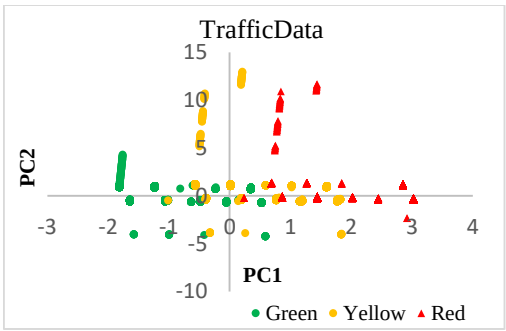

Figure 3. Traffic Dataset before applying Correlation Coefficient

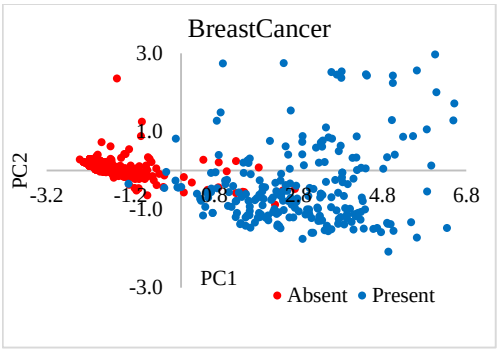

Figure 4. Breast Cancer Dataset before applying Correlation Coefficient

Figure 7,8,9 and 10 illustrate the four datasets on which correlation coefficient is applied. The most obvious dataset is *TrafficData*. It is obvious that the new dataset is clear than the old one. The second obvious one is Iris data. Some noisy features are removed because of that filtering process it makes clustering right decision in the same counterparts. Figure 11, 12, 13 and 14

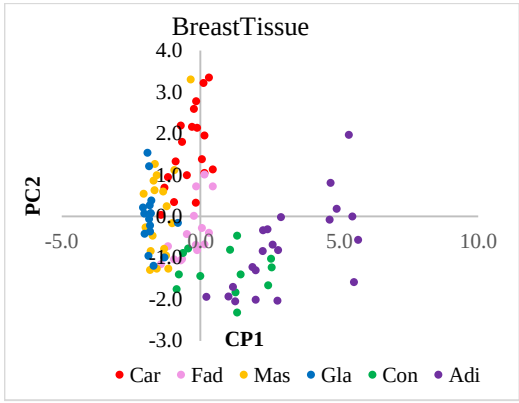

Figure 5. Breast Tissue Dataset before applying Correlation Coefficient

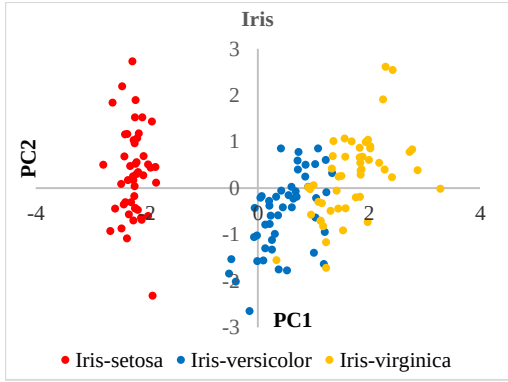

Figure 6. Iris Dataset before applying Correlation Coefficient

represent a scatterplot for visualizing the classification of observed instances to its right class. In these figure, there are two kinds of circles, filled circle and no filled circle. Filled circle denotes history data and no filled circle means an observed or predicted instance. Different color is different class type. For example, in Figure 11, where green filled circle represents traffic in *Green* state, yellow filled color is traffic in *Yellow* condition and red filled color means traffic in *Red* condition. In the lower left corner of Figure 11, no filled circle which has green border color is mostly close to green filled circle. It means that

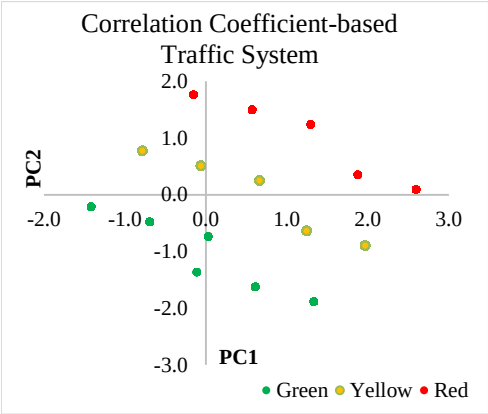

Figure 7. Traffic Dataset After applying Correlation Coefficient

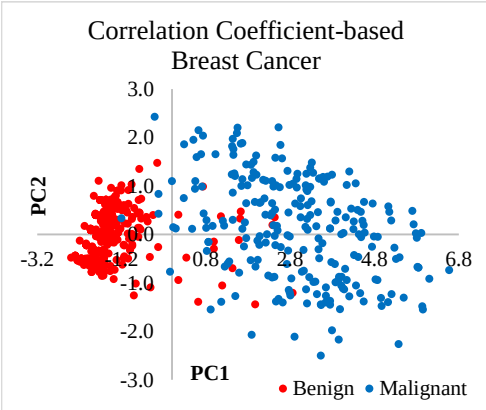

Figure 8. Breast Cancer after applying Correlation

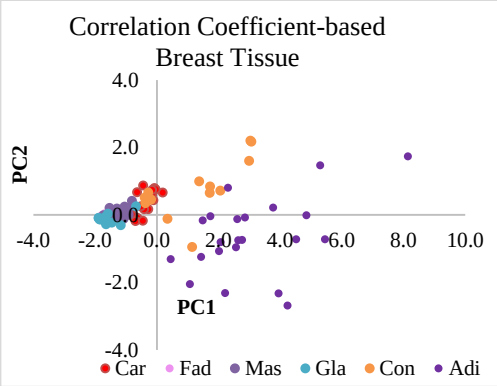

Figure 9. Breast Tissue after applying Correlation Coefficient

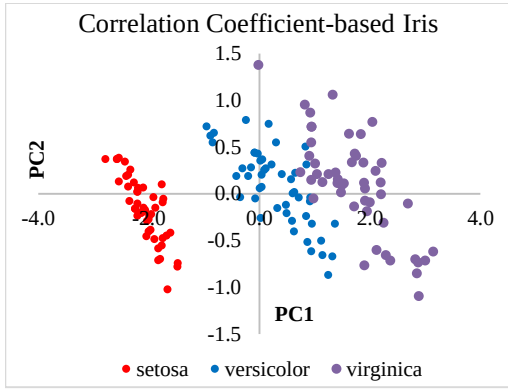

Figure 10. Iris after applying Correlation Coefficient

the observed instance which belongs to the class type *Green* was correctly classified into traffic in *Green* condition. Moreover, in the right most part, the observed instance which has *Red* class type is correctly classified into traffic in *Red* condition. However, there is no point which was wrongly classified. In the right most part, the observed instance which has *Red* class type, the classifier correctly estimated it into traffic in *Red* state. It is obvious that the estimation skill of the new k-nearest neighbor is still good enough. Figure 12 shows the correctly classification of the new k-nearest neighbor for *Breast Cancer* dataset. Figure 13 and 14 gives the visualization of the correctly classification of the new method for *Breast Tissue* and *Iris* datasets respectively.

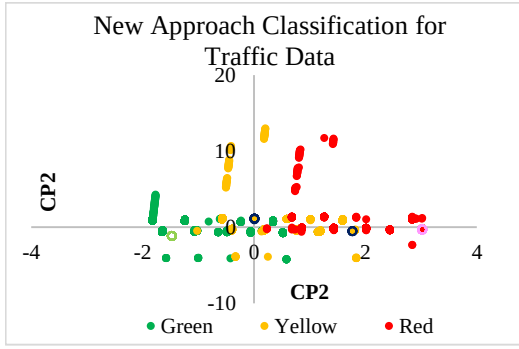

Figure 11. Illustration of Correctly Classification for Traffic Data

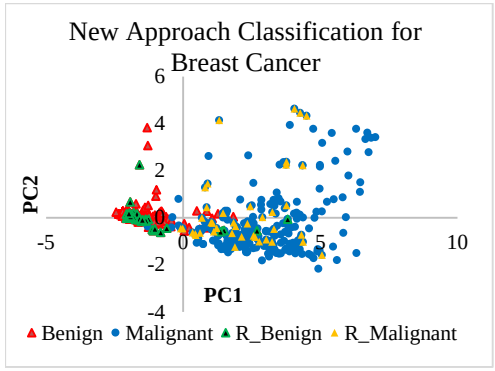

Figure 12. Illustration of Correctly Classification for Breast Cancer

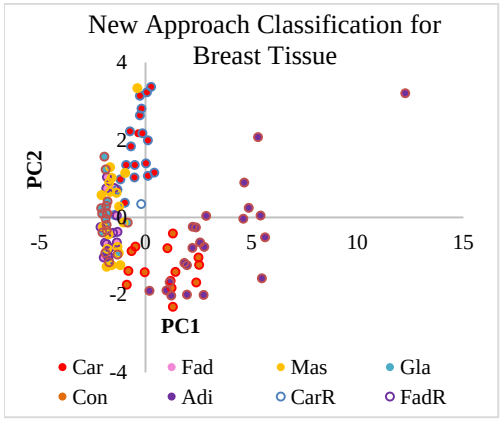

Figure 13. Illustration of Correctly Classification for Breast Tissue

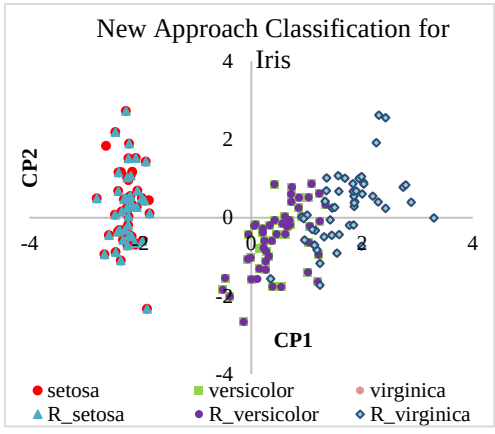

Figure 14. Illustration of Correctly Classification for Iris Data

Table 2 illustrates K-fold cross validation of correlation coefficient-based k-means clustering for k-nearest neighbor on four datasets, *Brest Cancer*, *Breast Tissue*, *TrafficData* and Iris by applying **Algorithm 1**. It can be clearly seen that the new k-nearest neighbor could unexpectedly did better in each prediction of four datasets. According to the table, the average k-fold cross validations of *Traffic Data* and *Breast Tissue* are 100%, whereas the validations of *Breast Cancer* and *Iris* are 95% and 82% respectively.

Table 2. Accuracy of New Approach using K-Fold Cross Validation

	Exp 1	Exp 2	Exp 3	Exp 4	Total Accuracy
BreastCancer	89%	96%	98%	97%	95%
BreastTissue	100%	100%	100%	100%	100%
Iris	81%	78%	81%	87%	82%
Traffic Data	100%	100%	100%	100%	100%

Table 3. Accuracy of Classical One using K-Fold Cross Validation

	Exp 1	Exp 2	Exp 3	Exp 4	Total Accuracy
BreastCancer	90%	95%	98%	99%	96%
BreastTissue	50%	100%	100%	52%	76%
Iris	100%	100%	100%	100%	100%
Traffic Data	100%	100%	100%	100%	100%

Table 3 gives the accuracy of classical k-nearest neighbor too. The new approach could predict the best estimation accuracy for all datasets, whereas the old one failed in prediction accuracy for *Breast Tissue* with 76%. By comparing the new approach with the classical

one, the new one did better estimation accuracy than the old one.

Table 4 illustrates and compares the processing time performance of the classical k-nearest neighbor with the new k-nearest neighbor by applying on four datasets. It can be clearly seen that the processing time performance of proposed approach for BreastCancer did 5.6 times (560%) faster than the old one. For *Traffic Data*, the proposed approach did about 1.58 times (158%) faster than the old one. In this case, there is a reason why the system did *Breast Cancer*, *Breast Tissue* and Iris faster than *Traffic Data* is that the different datasets have different shape and form. In other word, these datasets have different attribute and different value. Based on these facts, the processing time of each dataset are different each other. In *Breast Cancer*, it has nine attributes, but Iris has only four attributes. Additionally, the proposed approach could estimate faster than the classical k-nearest neighbor.

Table 4. Comparison of Processing Time Performance of Classical and New K-NN

Data Set	Amount of Data (KB)	Processing Time (Milliseconds)		Performance of New KNN over Classical KNN
		Classical KNN	Proposed KNN	
Breast Cancer	92	1250	222	563%
Traffic Data	81	1160	734	158%
Breast Tissue	58	91	32	284%
Iris	221	83	31	268%

5. Conclusion

In this study, we propose the new classification approach, correlation coefficient-based k-nearest neighbor, and compare the processing time performance of the new approach with the old own. In this work, the new approach did better performance in processing time than the classical one. Besides, the new approach could predict with the higher accuracy than the classical k-nearest neighbor. For future work, we will focus on improving processing time of each processing step and estimation skill more accurately.

References

[1]B.Tang and H.He, Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA, "Enn: Extended Nearest Neighbor Method for pattern Recognition", IEEE Computational Intelligence Magazine, 1556-603X, August 2015.

[2]K.Chomboon, P.Chujai, P.Teerarassamee, K.Kerdprasop and N.Kerdprasop, "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm", Proceedings of the 3rd International Conference on Industrial Application Engineering, 2015.

[3]K.R.Zalik, University of Maribor, Faculty of Natural Sciences and Mathematics, Department of Mathematics and Computer Science, "An efficient k-means clustering algorithm", Pattern Recognition Letters Journal, 29 (2008) 1385-1391.

[4]M.Du,S.Ding and H.Jia, School of Computer Science and Technology, China University of Mining and Technology, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis", Knowledge-Based Systems Journal, Knowledge-Bawd System 99 (2016) 135-145.

[5]M.Doshi and S.K.Chaturvedi, "Correlation Based Feature Selection Technique to Predict Student Performance", International Journal of Computer Networks & Communications (IJCNC) Vol.6, No.3, May 2014.

[6]M.Lucinska and S.Wierzchon, "Spectral Clustering Based on k-Nearest Neighbor Graph", CISIM 2012, pp 254-265.

[7] L.Zhang, Q.Liu, W.Yang, N.Wei and D.Dong, "An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction", 13th COTA International Conference of Transportation Professionals (CICTP 2013), pp 653-662, 2013.