Journal of Theoretical and Applied Information Technology
15th November 2017. Vol.95. No 21
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFECTIVENESS OF K-MEANS CLUSTERING TO DISTRIBUTE TRAINING DATA AND TESTING DATA ON K-NEAREST NEIGHBOR CLASSIFICATION

MUSTAKIM
Data Mining Laboratory Department of Information System
Faculty of Science and Technology of Universitas Islam Negeri Sultan Syarif Kasim Riau 28293,
Pekanbaru, Riau, Indonesia
E-mail: mustakim@uin-suska.ac.id

ABSTRACT

One of the constraints in classification is how to divide the dataset into two parts, training and testing which can represent every data distribution. The most commonly used technique is K-Fold Cross Validation which divides data into several parts and alternately into training data and testing data. In addition, the commonly used technique is to divide data into percentage form (70% and 30%), also become an option in data mining research. K-Means is a grouping algorithm which able to maximizes the effectiveness of distributing data in classification. The experiments performed using K-Means Clustering against K-Nearest Neighbor (K-NN) which was validated by Confusion Matrix have the highest accuracy of 93.4%, it is higher than the K-Fold Cross Validation data distribution technique for each experiment using data Education Management Information System (EMIS) as well as random data. The concept of distributing data in groups can be a representative to each member and increase the accuracy of classification algorithm, although the experiment only applied 70% of training data and 30% of testing data in each group.

Keywords: *Confusion Matrix, K-Fold Cross Validation, K-Means Clustering, K-Nearest Neighbor.*

1. INTRODUCTION

Data Mining technique has several characteristics in performing a process to achieve a result. In data mining, accuracy becomes a major benchmark in summing up the results obtained [27]. Beside accuracy, several classification algorithms such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) are more likely to use an error approach such as Root Means Square Error (RMSE) as a reference to the success of algorithm [28]. Similarly, other classification algorithms such as K-Nearest Neighbor (K-NN), Decision Tree and Naïve Bayes Classifier (NBC) prioritize the accuracy based on confusion matrix values as the initialization of algorithm accuracy [1].

In the case of algorithm accuracy, the most important thing before classification is the process of determining the training data and testing data. Errors in determining data distribution will be fatal and will affect the results of algorithm accuracy [29]. In several studies, there were many ways to divide the dataset classification such as K-Fold Cross Validation technique to compare ANN algorithm and Support Vector Regression (SVR) for predicting oil palm in Riau Province, with each MSE were 9% and 6% [2]. Another way by Jian Zhou and others was to divide the data into two parts: 70% of training data and 30% of testing data, by applying 10 Cross Validation [30]. However, the division in that way still leaves some unfavorable accuracy in some cases. This is caused by distribution of data which has not been represented by each data component. K-Fold Cross Validation divides data by turning each part of data into training data and the rest as testing data alternately [3][4], this technique is often used either with small data or with data that has more records.

When viewed from the relationship of data distribution, K-Fold Cross Validation technique has not represent the variety of datasets which will be divided into two parts, training data and testing data. One of the techniques which can be a representative to each data is clustering. The concept of clustering can be used as a reference in the distribution of data, that each data which has been grouped with its members will represent the characteristics of the group [5] also no overlap on each data [6]. If we look from the concept of clustering, a dataset will be divided into several groups, each group will take

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

some parts to be used as training data and testing data. Next, from the section of training data will be combined with training data in other groups, as well as on the testing data [31]. From the study the RMSE value of data distribution using K-Means Clustering was lower than the data distribution with 70% and 30% technique, the shortcoming of this study was it did not compare with the K-Fold Cross Validation [31].

The most commonly used grouping technique and included in the top ten popular data mining algorithms is K-Means Clustering [7]. In addition, K-Means has a good degree of accuracy [8], low complexity [9] and flexible to other algorithms [10][32]. In the study conducted by Karegowda, K-Means algorithm was able to provide higher values in the process of hybrid algorithm with K-NN to categorize the diabetic patients, with an accuracy of 96.86% [11], in his research K-Means served as one of the data elimination techniques along with Genetic Algorithm (GA) [11]. In terms of cluster validity, in some studies comparing Fuzzy C-Means, it was higher than that of K-Means [12]. However, this study did not discuss the comparison of clustering algorithms and hybrid clustering with other algorithms.

Therefore, this research will be focusing to discuss and compare the distribution of training data and testing data based on K-Fold Cross Validation technique and K-Means Clustering technique which will be tested on K-NN Algorithm. In the classification algorithm, K-NN is an algorithm which works by calculating the closest distance between data attributes [13][14], has an advantage in terms of high-performance computing speed [15], a simple algorithm and resilient to various characteristics of large data [16] also has a good accuracy compared to other algorithms [15]. From those advantages of K-NN, the experiment and the technique of distributing training data and testing data will be applied to data of Pesantren School in Pekanbaru which consist of 4,900 data obtained from Education Management Information System (EMIS). In addition, as a comparison we will also use 150,000 of random data generator, as verification to the accuracy of K-Fold Cross Validation and K-Means as data distribution techniques. The main motivation in this study is to compare the accuracy from distribution of training and testing data for classification algorithm.

2. LITERATUR REVIEW

2.1. K-Fold Cross Validation

K-Fold Cross Validation is done to divide data into training set and testing set. The essence of this validation is to divide the data randomly into the desired subset. K-Fold Cross Validation repeats k-times to divide a set randomly into the most free set of k, Each repetition leaves a set for testing and other set for training [17].

The K subset was selected by choosing one subset into testing data and the rest (k-1) was used as training data. However, in theory there is no definite benchmark for the value of k. The advantage of K-Fold Cross Validation compared with cross validation variations such as repeated random sub-sampling validation is all data was used for testing data and training data [18].

2.2. K-Means Clustering

Cluster analysis is the task of grouping data (objects) based solely on the information found in the data that describes these objects and the relationship between them [4]. Clustering is the process of making a group so that all members of each partition has a similarity based on certain matrix and a number of k in the data [19]. Data objects located in one cluster must have similarities while those who are not in the same cluster have no resemblance. K-means algorithm consists of two separate phases, first is to calculate the k centroid while the second requires the cluster point which has the nearest neighbor to the centroid of each data [7]. There are many ways that can be used to determine the distance from the nearest centroid, one of the most frequently used method is Euclidean Distance [20].

The purpose of clustering is to minimize the objective function that is set in the process of clustering, generally it tries to minimize the variation within a cluster and maximize the inter-cluster variation [6]. The distance between two points of X1 and X2 in manhattan / city block distance space is calculated by using the following formula [21]:

$$D_{L_1}(x_2, x_1) = \|x_2 - x_1\|_1 \quad (1)$$

As for the Euclidean distance space, the distance between two points is calculated by using the following formula [21]:

$$D_{L_2}(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} \left(x_{2_j} - x_{1_j}\right)^2} \quad (2)$$

Journal of Theoretical and Applied Information Technology
15th November 2017. Vol.95. No 21
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3. K-Nearest Neighbor (K-NN)

The algorithm was first introduced by Fix and Hodges in 1951 and 1952 [11]. This algorithm is also one of the lazy learning techniques. KNN is done by searching k-group objects in the closest training data (similar) to object in new data or testing data [22].

K-NN is included as a method of data mining classification based on learning by analogy. The sample of training data has a numerical dimension attribute. Each sample is a point in the n-dimensional space. All training samples are stored in n-dimensional space. When testing the data, it will find the value of k closest to the testing data. The proximity is defined in terms of Euclidean distance between two points X=(x1, x2,…,xn and Y=(y1,y2,…,yn) [23].

2.4. Confusion Matrix

It is a model of classification evaluation based on testing data and all predicted data with appropriate proportion [24].

Table 1 Table Confusion Matrix 2 Class

Classification	*Prediction Class*	
	Class = Yes	Class = No
Class = Yes	a (*true positive TP)*	b (*false negative FN)*
Class = No	c (*false positve FP)*	d (*true negative TN)*

The calculation of accuracy level on Confusion Matrix 2 classes based on Table 1 above is [24]:

$$Acuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{A+D}{A+C+D} \quad (3)$$

3. RESEARCH METHODOLOGY

Data collection was done by using EMIS data of Ministry of Religious of Pekanbaru City and generated random number. Transformation step is using a numerical scale based on each attribute which then we perform normalization by using Min-max Normalization. The purpose of data normalization is to get the same weight from all of the data attributes and does not have variation or the result from weighting does not consist of more dominant attribute or considered more important than the others [25]. Min-max Normalization performs a linear transformations on the data, by using a minimum value and a maximum value. Min-max normalization maintains the relationship between the values of original data [26].

The data distribution in this study will be done with 4 Fold of 4,900 data before the data is cleaned, each fold will get equal share of the amount of data. It also performed a comparison by generating random numbers of 150,000 data, with 37,500 data testing and 112,500 training data. Data distribution by K-Fold Cross Validation will be compared to data distribution techniques using K-Means Clustering. In K-Means, the data will be divided into 3, 4, 5, 6 and 7 group section then from each group, 30% data will be used as testing data while 70% as training data. From each data group, training data will merge into other training data group and testing data will merge into other testing data group. The two data-distribution models will be implemented using K-NN with Confusion Matrix as its accuracy model.

The three class targets of KNN consist of Economic Life and Level of Family Education Low (1), Medium (2) and High (3). While the 9 (nine) attributes for classification process are (A1) Gender, (A2) Category Students Study, (A3) Level of Education, (A4) Class, (A5) Father's Formal Education, (A6) Father's Job, (A7) Mother's Formal Education (A8) Mother's Job and (A9) Average Parent's Earnings. Similarly, random data also implements 9 random attributes as a comparison of the algorithm result. Generally, the methodology in this study can be shown in Figure 1 below:

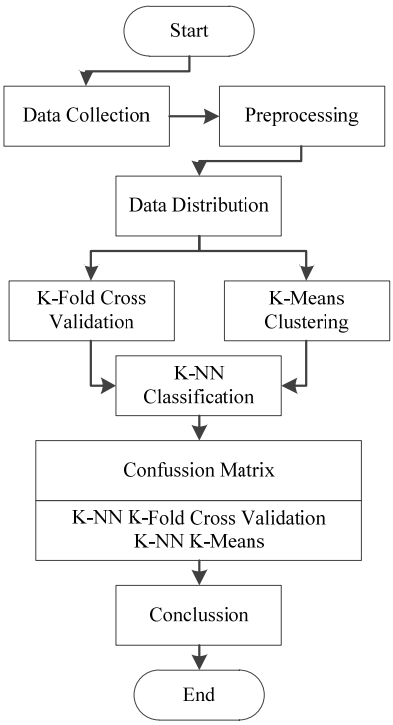

Figure 1. Research Methodology

4. RESULT AND ANALYSIS

The data used in the simulation in this study comes from the EMIS data of Ministry of Religious

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of Pekanbaru City 2016 with the amount of data 4.900 records and 9 attributes that have been validated by relevant agencies. Dataset can be shown in Table 2:

Table 2. Dataset Students of Pondok Pesantren in Pekanbaru Year 2016

No	Name of Pondok Pesantren	NIS Local and National	A1	A2	A3	A4	A5	A6	A7	A8	A9	Class
1	Al-Ikhwan	131214710010130200	M	2	4	10	1	7	2	1	2	2
2	Al-Ikhwan	131214710010130202	F	2	4	10	1	13	2	1	4	3
3	Al-Ikhwan	131214710010130203	M	2	4	10	7	7	2	3	5	3
4	Al-Ikhwan	131214710010130204	M	2	4	10	1	7	1	1	3	2
5	Al-Ikhwan	131214710010130205	M	2	4	10	1	15	1	1	2	2
6	Al-Ikhwan	131214710010130206	M	2	4	10	1	15	1	1	3	3
7	Al-Ikhwan	131214710010130207	F	2	4	10	1	15	1	1	2	2
8	Al-Ikhwan	131214710010130208	M	2	4	10	2	7	2	1	2	2
…	…	…	…	…	…	…	…	…	…	…	…	…
4.576	Umar bin Khattab	510314710002130029	F	1	3	12	2	13	1	12	4	3

Source: Data EMIS Students of Pondok Pesantren in Pekanbaru Year 2016

From table 2 above, some parts of each attribute can be described as follows:

1. A1 = Gender (M: Male and F: Female)
2. A2 = Category of Students Study (1: Just Following the Book Study and 2: Following the Book Study and Other Education Services)
3. A3 = Education Level (1: RA, 2: MI, 3: MTs, 4: MA and 5: Islamic University)
4. A4 = Class (Class 1 to 12)
5. A5 = Father's Formal Education (0: No Education, 1: Elementary and JHS, 2: SHS, 3: D1, 4: D2, 5: D3, 6: D4, 7: Bachelor and 8; Master)
6. A6 = Father's Job (1-15)
7. A7 = Mother's Formal Education (0: No Education, 1: Elementary and JHS, 2: SHS, 3: D1, 4: D2, 5: D3, 6: D4, 7: Bachelor and 8; Master)
8. A8 = Mother's Job (1-16)
9. A9 = Average Parent's Earnings (1: Less than 500.000, 2: 500.001-1.000.000, 3: 1.000.001-2.000.000, 4: 2.000.001-3.000.000, 5: 3.000.001-5.000.000 and 6. More than 5.000.000)

From the data in Table 2 above, normalization can be done with Min-Max Normalization, the result is shown in Table 3 below:

Table 3. Data Normalization

No	NIS Local and National	A1	A2	A3	A4	A5	A6	A7	A8	A9	Class
1	131214710010130200	0.0000	1.0000	1.0000	0.8182	0.1111	0.3529	0.2222	0.0000	0.2000	0.5000
2	131214710010130202	1.0000	1.0000	1.0000	0.8182	0.1111	0.7059	0.2222	0.0000	0.6000	1.0000
3	131214710010130203	0.0000	1.0000	1.0000	0.8182	0.7778	0.3529	0.2222	0.1176	0.8000	1.0000
4	131214710010130204	0.0000	1.0000	1.0000	0.8182	0.1111	0.3529	0.1111	0.0000	0.4000	0.5000
5	131214710010130205	0.0000	1.0000	1.0000	0.8182	0.1111	0.8235	0.1111	0.0000	0.2000	0.5000
6	131214710010130206	0.0000	1.0000	1.0000	0.8182	0.1111	0.8235	0.1111	0.0000	0.4000	1.0000
7	131214710010130207	1.0000	1.0000	1.0000	0.8182	0.1111	0.8235	0.1111	0.0000	0.2000	0.5000
8	131214710010130208	0.0000	1.0000	1.0000	0.8182	0.2222	0.3529	0.2222	0.0000	0.2000	0.5000
…	…	…	…	…	…	…	…	…	…	…	…
4.576	510314710002130029	1.0000	0.0000	0.6667	1.0000	0.2222	0.7059	0.1111	0.6471	1.0000	1.0000

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

As for random data, generated with 150,000 data, and dimensions of 9 x 150,000 and one class was chosen randomly as a representative of the largest data vector. Random data can be shown in Table 4 below:

Table 4. Random Data

Data Record	AR1	AR2	AR3	AR4	AR5	AR6	AR7	AR8	AR9	Class MAX AR
Data 01	0.4467	0.5622	0.7509	0.6921	0.7014	0.1944	0.6650	0.8933	0.9868	0.0000
Data 02	0.0977	0.4136	0.3746	0.5498	0.0390	0.2920	0.2311	0.2319	0.2057	1.0000
Data 03	0.2140	0.6479	0.5541	0.5168	0.1627	0.1241	0.3055	0.9097	0.4552	0.5000
Data 04	0.2807	0.3100	0.4283	0.8103	0.3536	0.0931	0.7350	0.1631	0.0129	0.5000
Data 05	0.4984	0.7033	0.4788	0.4132	0.7848	0.8286	0.4432	0.2507	0.1524	0.5000
Data 06	0.3984	0.5454	0.7180	0.4297	0.2182	0.3775	0.2419	0.8974	0.0232	0.0000
Data 07	0.0241	0.2031	0.0998	0.8371	0.4471	0.6327	0.9407	0.1590	0.7004	0.0000
Data 08	0.7908	0.2412	0.4026	0.4047	0.7780	0.3642	0.3346	0.5327	0.4469	1.0000
…	…	…	…	…	…	…	…	…	…	…
Data 150.000	0.9291	0.3943	0.8623	0.2242	0.2305	0.1772	0.8301	0.2446	0.8975	0.5000

Both EMIS data and generated random data were used as a proof of K-NN accuracy using K-Fold Cross Validation data distribution techniques and K-Means Clustering.

4.1. Distribution of K-Fold Cross Validation Data

The problem in this research is to compare between some parts of data that is separated into several parts based on training data and testing data into k. Of the 4 k to be formed will produce an accuracy based on confusion matrix. 4,576 data will be divided by 1,144 data into each k, as well as 150,000 data divided into 37,500 data into each k.

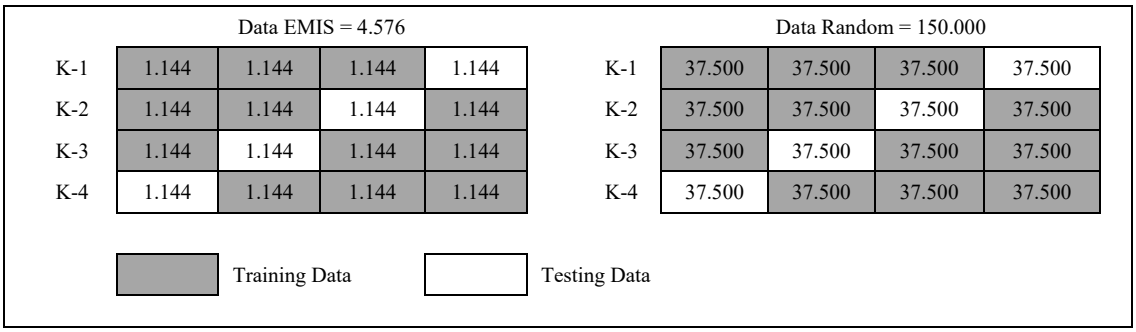

Figure 3. Illustration of Training Data and Testing Data on K-Fold Cross Validation

The experiment using K-Fold Cross Validation resulted in accuracy based on confusion matrix on K-NN with maximum value 77.8% at k = 3 for EMIS data and 71.6% at k = 2 for random data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

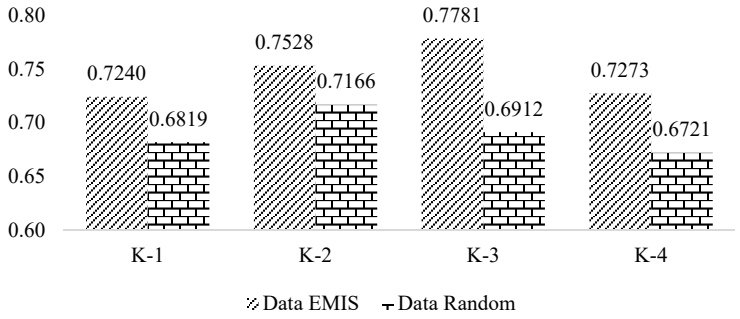

Figure 4. Accuracy Value of Each k on Cross Validation

From Figure 4 above it can be concluded that the higher the value of k then the accuracy will decrease, both for EMIS data and random data. Therefore, in this study the experiments on value of k = 5, k = 6 and k = 7 and so were not performed. From the data used respectively, the highest k value between the EMIS data and the random data has the amount of data corresponding to its class are 3,561 data and 107,490 data.

4.2. Distribution of K-Means Clustering

The cluster formed consists of 4 parts with the K-Means Custering technique as data distribution to verify data accuracy on K-NN. Based on the group produced by K-Means, each group will be divided into 2 parts, 70% for training data and 30% for testing data. Then from the results of the distribution, merging is done to each data. The maximum iteration generated by K-Means with EMIS data is 227 iterations while for random data 612 iterations. The visualization of cluster can be shown in Figure 5 below:

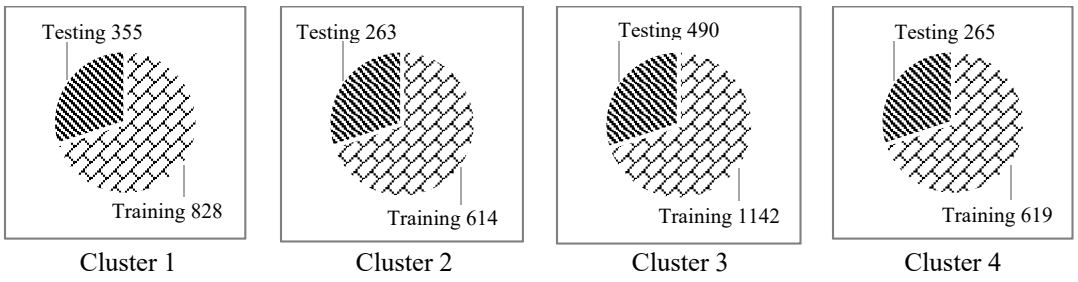

Figure 5. Cluster results for EMIS Training Data and Testing Data

Figure 5 above shows the training and testing sections from each group, so that 3,203 training data and 1,373 testing data were obtained. The classification performed by K-NN from data distribution above shows an accuracy of 91.2% higher than K-Fold Cross Validation data distribution technique which has only 77.8% of accuracy. In this experiment, in addition of using 4 clusters, it also used experimental clusters as many as 3, 5, 6 and 7 clusters. The results show that the number of clusters = 5 has a maximum accuracy of 93.4%.

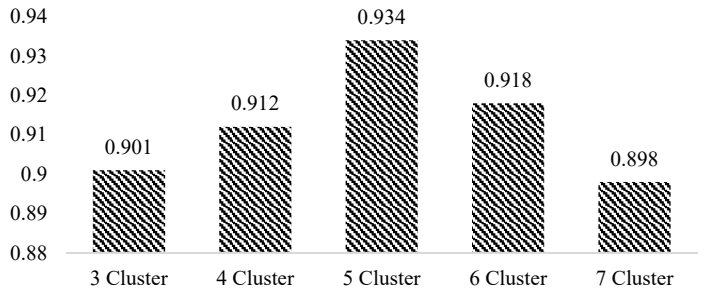

Figure 6. Accuracy of KNN Confusion Matrix In Each Number of Clusters on EMIS Data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

As with EMIS data, random data cluster also has a higher accuracy than random data for K-Fold Cross Validation. The accuracy of random data for K-NN with 3 clusters up to 7 clusters were 84.7%; 84,0%, 84,9%; 85.2% and 85.0%. So it can be concluded that the accuracy generated by random data has a relatively close value to the number of different cluster data.

5. CONCLUSION

Based on the research conducted there are some knowledge found among them such as for the distribution of training data and testing data based on K-Means Clustering has a higher accuracy of confusion matrix compared to K-Fold Cross Validation in all experiments. The highest values of each of these data distribution techniques were 93.4% for K-Means Clustering and 77.8% for K-Fold Cross Validation. Experiments conducted using EMIS data have a higher accuracy tendency than random data using either K-Means Clustering or K-Fold Cross Validation because of distribution range in random data doesn't have specific variation. Unfortunately, this research has not been experimented using large data in number of hundreds of millions data records with many attributes and the distribution on each group only based on 70% of training data and 30% testing data. In addition, the disadvantage of data distribution by clustering leads to the effectiveness of members in each cluster that have many different data between clusters or the possibility of not having members in each group at all. The advantage of using clustering techniques in dividing data is that each training data and test data can be represented by each cluster member so that the proximity concept becomes the best pattern in performing the data sharing in the classification process.

6. ACKNOWLEDGEMENT

A biggest thanks to Faculty of Science and Technology UIN Sultan Syarif Kasim Riau on the financial support for this research, the facilities and mental support from the leaders. And also thanks to Puzzle Reseach Data Technology (Predatech) Team Faculty of Science and Technology UIN Sultan Syarif Kasim Riau for their feedbacks, corrections and their assistance in implementing these activities so that research can be done well.

REFRENCES:

[1]. Navin M and Pankaja R. 2016. “Performance Analysis of Text Classification Algorithms using Confusion Matrix”. 6(4). pp. 75-78.

[2]. Mustakim M, Buono A and Hermadi I. 2015. “Performance Comparison Between Support Vector Regression and Artificial Neural Network for Prediction of Oil Palm Production”. Journal of Computer Science and Information. 9(1). pp. 1-8.

[3]. Vanwinckelen G and Blockeel H. 2012. "On Estimating Model Accuracy with Repeated Cross-Validation". Appearing in Proceedings of BeneLearn and PMLS 2012. pp. 231-239.

[4]. Rao SG. 2015. “Performance Validation of the Modified K-Means Clustering Algorithm Clusters Data”. International Journal of Scientific & Engineering Research. 6(10). pp. 726-730.

[5]. Baarsch J and Celebi ME. 2012. “*Investigation of Internal Validity Measures for K-Means Clustering*”. Proceedings of the International Multi Conference of Engineers and Computer Scientiss 2012. 1 March. pp. 14-16.

[6]. Salman R dan Kecman V. 2011. “Fast K-Means Algorithm Clustering”. International Journal of Computer Networks and Communications (IJCNC). 3(4), pp. 76-85.

[7]. Dhanachandra N, Manglem K and Chanu YJ. 2012. “Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm”. Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015). pp.764 – 771.

[8]. Sakthi M and Thanamani AS. 2011. “An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA”. International Journal of Computer Science and Information Technologies (IJCSIT). 2(3). pp. 955-959.

[9]. Patel VR and Rupa GM. 2011. “Impact of Outlier Removal and Normalization Approach in Modified K-Means Clustering Algorithm”. IJCSI International Journal of Computer Science Issues. 8(5). pp. 111-121.

[10]. Napoleon D and Pavalakod S. 2011. “A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set”. International Journal of Computer Applications. 13(7). pp. 41-46.

[11]. Karegowda AG, Jayaram MA and Manjunath AS. 2012. "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients"

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

International Journal of Engineering and Advanced Technology (IJEAT). 1(3). pp. 147-151.

[12]. Afrin F, Amin M, Tabassum M. 2015. "Comparative Performance Of Using PCA With K-Means and Fuzzy C Means Clustering for Customer Segmentation". International Journal of Scientific & Technology Research. 4(10). pp: 70-74

[13]. Hassanat AB, Abbadi MA and Alhasanat AA. 2014. "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach". International Journal of Computer Science and Information Security (IJCSIS). 12(8). pp. 33-39.

[14]. Seetha M, Sunitha KVN, and Devi GM. 2012. "Performance Assessment of Neural Network and K-Nearest Neighbour Classification with Random Subwindows". International Journal of Machine Learning and Computing. 2(6). pp. 844-847.

[15]. Khamis HS, Cheruiyot KW and Kimani S. 2014. "Application of K- Nearest Neighbour Classification in Medical Data Mining". International Journal of Information and Communication Technology Research. 4(4). pp. 121-128.

[16]. Imandoust SB and Bolandraftar M. 2013. "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background". 3(5). pp.605-610.

[17]. Rao SS. 2009. Engineering Optimization: Theory and Practice. John Wiley and Sons, Newyork.

[18]. Adhani G, Buono A, Faqih A. 2013. Support Vector Regression modelling for rainfall prediction in dry season based on Southern Oscillation Index and NINO3.4. International Conference on Advanced Computer Science and Information Systems. pp. 2013: 315-320.

[19]. Khan SS and Ahmad A, 2004, "Cluster Centre Initialization Algorithm for K-means Cluster", International Conferences In Pattern Recognition Letters, pp. 1293–1302.

[20]. Yedla M, Pathakota SR and Srinivasa TM, 2010, "Enhanced K-means Clustering Algorithm with Improved Initial Center", International Journal of Science and Information Technologies, 1(2), pp. 121–125.

[21]. Celebi and Emre M, 2012, "Deterministic Initialization of The K-Means Algorithm Using Hierarchical Clustering", International Journal of Pattern Recognition and Artifcial Intelligence. 26(7), pp. 55-61.

[22]. Parvin H, Alizadeh H and Minati B. 2010. "A Modification on K-Nearest Neighbor Classifier". Global Journal of Computer Science and Technology. 10(14). pp. 37-41.

[23]. Verma V, Bhardwaj S and Singh H. 2016. "A Hybrid K-Mean Clustering Algorithm for Prediction Analysis". Indian Journal of Science and Technology. 9(28). pp. 1-5.

[24]. Santra AK, Christy CJ. 2012. "Genetic Algorithm and Confusion Matrix for Document Clustering". IJCSI International Journal of Computer Science Issues. 9(1). pp. 322-328.

[25]. Xu Q, Ding C, Liu J and Luo B, 2015, "PCA-guided search for K-means", Pattern Recognition Letters 54, pp.50–55.

[26]. Jain YK and Bhandare SK, 2011, "Min Max Normalization Based Data Perturbation Method for Privacy Protection", International Journal of Computer and Communication Technology, 2(8).

[27]. Gupta N, Rawal A, Narasimhan VL, Shiwani S. 2013. "Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data". IOSR Journal of Computer Engineering (IOSR-JCE). 11(5). pp: 70-73.

[28]. Behzad M, Asghari K, Eazi M, Palhang M. 2009. "Generalization Performance of Support Vector Machines and Neural Networks in Runoff Modeling". Elsevier Expert Systems with Applications. 36(4). pp: 7624-7629.

[29]. Wei Q and Dunbrack RL. 2013. "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics". PLOS-ONE Journal. 8(7). pp: 1-12.

[30]. Zhou J, Li X; and Mitri HS. 2016. "Classification of Rockburst in Underground Projects: Comparison of Ten Supervised Learning Methods". Journal of Computing in Civil Engineering. 30 (5).

[31]. Saputra R. 2016. "Penerapan Algoritma Backpropagation untuk Memprediksi Bobot Hidup Kambing Berdasarkan Ciri Morfometrik pada Aplikasi Berbasis Mobile". Master Thesis UIN Sultan Syarif Kasim Riau. Pekanbaru – Indonesia.

[32]. Mustakim M. 2017. "Centroid K-Means Clustering Optimization Using Eigenvector Principal Component Analysis". Journal of Theoretical and Applied Information Technology. 95(15). pp: 3534-3542.