Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification

Okfalisa
Department of Informatic Engineering
UIN Sultan Syarif Kasim Riau
Pekanbaru - Indonesia
okfalisa@uin-suska.ac.id

Mustakim
Data Mining Laboratory Department of Information System
UIN Sultan Syarif Kasim Riau
Pekanbaru - Indonesia
mustakim@uin-suska.ac.id

Ikbal Gazalba
Department of Information System
UIN Sultan Syarif Kasim Riau
Pekanbaru - Indonesia
ikbal.gazalba@students.uin-suska.ac.id

Nurul Gayatri Indah Reza
Department of Information System
UIN Sultan Syarif Kasim Riau
Pekanbaru - Indonesia
nurul.gayatri.indah@students.uin-suska.ac.id

***Abstract*— Data mining is the process of handling information from a database which is invisible directly. Data mining is predicted to become a highly revolutionary branch of science over the next decade. One of data mining techniques is classification. The most popular classification technique is K-Nearest Neighbor (KNN). But there is also the Modified K-Nearest Neighbor (MKNN) classification algorithm which is the derived algorithm of KNN. In this paper we will analyze the comparison of KNN and MKNN algorithms to classify the data of Conditional Cash Transfer Implementation Unit (Unit Pelaksana Program Keluarga Harapan) which consist of 7395 records. Comparative analysis is based on the accuracy of both algorithms. Before classification, K-Fold Cross Validation was done to search for the optimal data modeling resulted in data modeling on cross 2 with accuracy of 93.945%. The results of K-Fold Cross Validation modeling will be the model for training data samples and testing data to test KNN and MKNN for classification. Classification result produced accuracy based on the rules of confusion matrix. The test resulted in the highest accuracy of KKN by 94.95% with average accuracy during the test was 93.94% and the highest accuracy of MKNN was 99.51% with the average accuracy during the test was 99.20%, almost all testing from the first test up to the tenth, MKNN algorithm is superior and has better accuracy value than KNN so it can be analyzed that the ability of MKNN algorithm in accuracy is better than KNN. It can be concluded that MKNN algorithm is capable of handling accuracy better for classification than KNN algorithm, by ignoring other aspects such as computerization, time efficiency, and algorithm effectiveness.**

Keywords*— *Classification, Data Mining, K-Nearest Neighbor, Modified K-Nearest Neighbor

I. INTRODUCTION

Data mining is the process of handling information from a database which is invisible directly. Therefore data mining can be used for various purposes in the private sector. Industries such as banking, insurance, and medicine usually use data mining to reduce costs, improving research, and increasing sales. The data analysis techniques is traditionally used for such tasks including regression analysis, cluster analysis, numerical taxonomy, multidimensional analysis, other multivariate statistical methods, stochastic models, time series analysis, nonlinear estimation techniques, and others. Data mining is a science needed to advance science in the field of artificial intelligence and statistics. Data mining is predicted to become a highly revolutionary branch of science over the next decade, the fact is according to MIT Technology Review chooses data mining as one of the ten technologies which is needed to change the world [1].

One of the techniques in data mining is classification, classification is a task to predict category label which is unknown before to distinguish between one object to another based on the attributes or features [1]. Classification is a very important technique in data mining that is used to classify data to be a more specific class. In this technique the data will be divided by training data i.e. data to be associated to know the category label and testing data which will be the test samples to discover the category labels [2].

The most popular technique in classification is K-Nearest Neighbor (KNN), KNN is named as one of the ten most popular and most important algorithms. KNN is known to be very simple and easy [2]. KNN algorithm is an algorithm that is often used in classification. The purpose of this algorithm is to classify new objects based on the attributes and training data. To classify objects based on the training data that has the closest distance to a new object is based on the euclidean equation formula [1]. The KNN algorithm becomes an option

978-1-5386-0958-2/17/$31.00 ©2017 IEEE

because it is good in handling noise, simple, uncomplicated, easy and uses a very large computerized [3].

In addition to the KNN classification technique, there is also a Modified K-Nearest Neighbor (MKNN) classification technique which is derived from classification algorithm of KNN by increasing the calculation of validity and weight voting. The validity is used to test the validity of training data, and the test will be based on the number of neighbors on all of training data samples. And weight voting is used to determine the highest weight of the multiplication calculation between validity with new sample data (data testing) to determine the final prediction class. MKNN was created to improve the accuracy of KNN by adding a step formula that is the step validity of training data and weight voting [3]. Previously in KNN only calculated the neighboring distance between training data and testing data, and determined the neighborhood based on the number of K. However, MKNN will perform the process of validity in training data before calculating training data with testing data, then the highest weight voting from the nearest neighbor based on the number of K or neighbors will be calculated from the multiplication result of training data and testing data

From both KNN and MKNN classification algorithm comparative analysis will be done to classify a sample data. In the previous study by Parvin et al (2008), conducted a comparison between KNN and MKNN with some data monk 1, monk 2, monk 2, monk 3, iso data, and wine. The result of MKNN validity is able to find the validity of stable training data and MKNN has a better accuracy rate than KNN [4]. Then, Parvin et al (2010) did another comparative research of KNN and MKNN with the result that MKNN has a better accuracy than KNN but the data used were data of wine, isodata, iris, bupa, ionosphere, and monk. Then Singh and Patel (2014) conducted a study on threat detection of information system by using KNN and MKNN algorithm, with the result that MKNN managed to find better threat detection than KNN due to the calculation of data validity [5].

In this study before making a comparison between the KNN and MKNN algorithms, K-Fold Cross Validation is done to find and select the best data model. This is done because according to Mutrofin et al (2014) KNN and MKNN algorithms have weakness in determining the parameter of K which is still random and K is determined manually [6]. That is why K-fold cross validation is needed to find the best k parameters. According to Kohavi (1995), to determine the best model is by using k-fold cross validation, starting by dividing data as many as some of K fold, the best fold is 10 fold cross validation and it may be bigger if k-fold is larger depending on requirement [7], however according to Last (2006) there is no exact terms on determining how many k-fold is needed, it may be 33%, 50%, and 60% of the data [8]. But in this study will use 3 fold cross validation and is limited from K = 1 to K = 10.

Comparative analysis of KNN and MKNN algorithm in data classification with K parameter and data model is determined from k-fold cross validation and the accuracy value is resulted from the matrix confusion equation [9]. The data used is taken from Conditional Cash Transfer Implementation Unit (Unit Pelaksana Program Keluarga Harapan) which consist of 7,395 records.

The reason why MKNN and KNN were chosen to be compared was to discuss again and seek justification from previous and related research, however this paper not only compare both algorithm but also form the optimal data modeling using K-Fold Cross Validation formula to determine the structure of training data and testing data to be more stable not randomly before testing the accuracy of both algorithms. Comparative analysis of KNN and MKNN algorithm in data classification with parameter K and data model was determined from the calculation of K-Fold Cross Validation and accuracy value resulting from confusion matrix equation [9].

II. Materials and Method

In general, this research will perform testing of two classification algorithms namely KNN and MKNN for comparison. For more details of Research Methodology in this paper can be seen in the figure below.

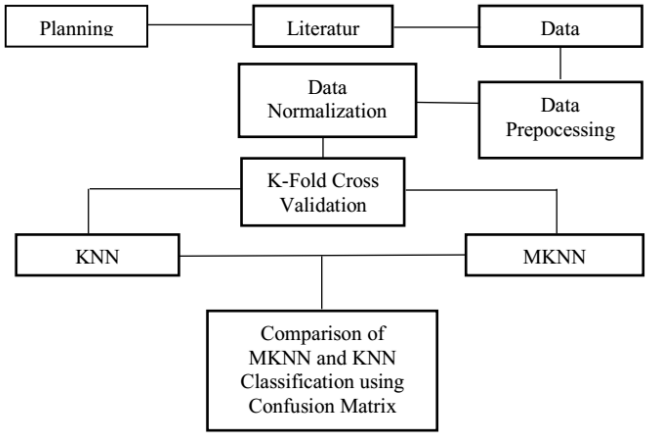

Fig. 1. Research Methodology

A. Data Mining

Data mining is a step in the process of knowledge discovery, it is important to find a hidden pattern for evaluation. However, in industry, in the media, and in the research environment, the term data mining is often used to refer to the entire process of knowledge discovery. Data mining is the process to find the interesting patterns and knowledge from large amounts of data. Data sources include in databases, data warehouses, Web, other information repositories, or dynamically streamed data to the system [9]. According to Gartner Group in Larose (2005) Data mining is the process of finding significant new correlations, patterns and trends by filtering large amounts of data stored in the repository, using pattern recognition technology as well as statistical and mathematical techniques [1].

B. Classification

Classification is one of the important techniques in data mining used to classify data into specified classes. In this technique, a training set (data training) is used in which all criteria or attributes are already associated with known class labels. Classification algorithm learns from data training and model building [2]. In the classification there is a target

variable which can be called as class label. This model will test a large data set containing information with the variable or attribute targets and also a set of input or predictor variables [1]. Classification also serves to predict class label of data input (data testing) where the model built can predict continuously, model like this is called predictor [9].

C. K-Nearest Neighbor (K-NN)

KNN was named as one of the ten most popular and most important algorithms. KNN is known to be very simple and easy. KNN is an example-based learning group. This algorithm is also one of the lazy learning techniques. KNN is done by searching for the group of K objects in the closest training data (similar) to objects in new data or data testing [2]. Generally the Euclidean distance formula is used to define the distance between two training objects and testing [10].

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

D. Modified K-Nearest Neihgbor (MK-NN)

The MKNN algorithm is a development method of KNN, but there are some additional processes or development from KNN [4], as for the stages of MKNN are as follows:

- Do the distance calculation
  KNN method is a classification method based on the closest object from the searched object (Parvin et al, 2008) [4]. The calculation of distance between two points are the point on training data (x) and the point in testing data (y) using Euclidean formula as formula (1) [10].
- Validity of Training Data
  Validity is used to calculate the number of points on all training data, the closest neighbor to each data will affect the validity of data [4]:

$$\text{Validity} = \frac{1}{K}\sum_{i=1}^{K} S\left(\text{lbl}(x), \text{lbl}\left(N_{i(x)}\right)\right) \quad (2)$$

  The S function is used to calculate the similarity between point x and the *i*-th data from the nearest neighbor as follows [4]:

$$S(a,b) = \begin{cases} 1_{a=b} \\ 0_{a\neq b} \end{cases} \quad (3)$$

- Weight Voting
  Determining weight voting is by using the validity of each data in the training data multiplied by the weight based on Euclidean distance. In MKNN method, weight vote calculation of each neighbor is in Equation below [4]:

$$W(x) = \text{Validalitas}(x) x \frac{1}{d_{e+0,5}} \quad (4)$$

E. K-Fold Cross Validation

According to Kohavi (1995), to determine the best model is by using k-fold cross validation, starting by dividing data as many as some of K fold, the best fold is 10 fold cross validation and it may be bigger if k-fold is larger depending on requirement [7], however according to Last (2006) there is no exact terms on determining how many k-fold is needed, it may be 33%, 50%, and 60% of the data [8]. K-Fold Cross validation is used in order to find the best parameters from one model. This is done by testing the amount of error in data testing. In cross validation, the data is divided into K cross samples of equal size. From K subset the data K-1 sample will be used as training data and one remaining sample for testing data. Furthermore, the process of training and testing will calculate the average error (error mean). Every running will found error for data testing, the model which gives the smallest average error is selected to be the best method. In cross validation, it has uncertainty in determining the cross test [8].

F. Confusion Matrix

Confusion matrix is a tool used to evaluate classification models to estimate the true or false objects [9]. A matrix of prediction which will be compared with the original class of inputs or in other words contains information of actual and predicted value on classification [11].

III. Result and Analysis

In accordance with previous research methodology, several important things which will be done in this research is consists of planning, data collection and processing, K-Fold Cross Validation, and Classification of KNN and MKNN, and comparison analysis of KNN and MKNN Algorithm using Confusion Matrix.

The stage of processing data includes in data collection from related institutions, preprocessing the data, and normalizing the data prior to data mining stage. Results of Final data from Preprocessing, Normalization of data can be seen in Table 1 below.

TABLE I. Final Result of Data

No	Name	V-1	V-2	V-3	V4	Class
1	A Sihaloho	2	1	0	0	4
2	Aberta Purba	2	1	0	2	5
3	Addes Lani	1	0	0	0	1
4	Adek Risna	4	0	0	0	4
5	Aderlina	2	0	0	1	4
...	...	...	...	...	...	...
7.395	Zurmailis	2	1	0	1	5

Explanation:

V-1 : Number of elementary students
V-2 : Number of junior high student
V-3 : Number of pregnant woman
V-4 : Number of toddler

Due to the amount of 7,395 data records then in cross validation they will be divided equally as many as 3 models, then the division is:

$$\text{ratio} = \frac{7.395}{3} = 2.465$$

In the Cross validation data of Conditional Cash Transfer will be tested in 3 samples. In 3 to 1 samples as training data, and 1 sample as data testing. Cross validation takes 10 tests to determine the value of "K" parameter and a good model to use in algorithm test. Cross validation is tested using Weka Machine Learning tool or application. Validation test results can be seen in Figure 2 below.

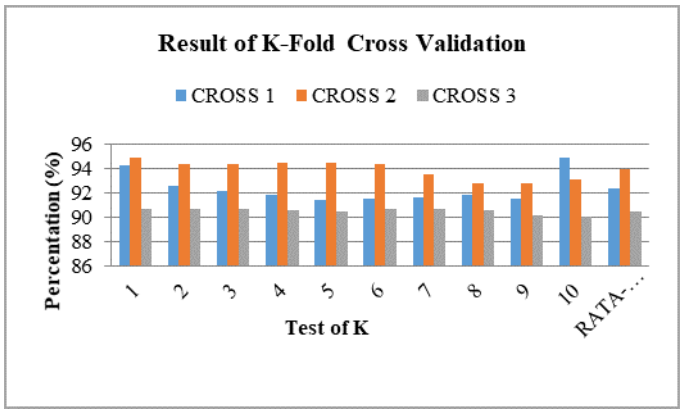

Fig. 2. Accuracy Measurement Results

Detailed results of cross validation calculations can be seen in Table 3 Results Measurement Accuracy.

TABLE II. ACCURACY MEASUREMENT

K	*CROSS* 1	*CROSS* 2	*CROSS* 3	Rata K
1	94,323	94,950	90,664	93,312
2	92,635	94,405	90,664	92,568
3	92,212	94,409	90,643	92,421
4	91,837	94,484	90,617	92,313
5	91,476	94,495	90,484	92,157
6	91,519	94,364	90,709	92,197
7	91,595	93,583	90,690	91,956
8	91,903	92,831	90,553	91,762
9	91,562	92,831	90,161	91,518
10	94,950	93,097	90,037	92,695
Average of *Cross*	92,401	93,945	90,522	91,566

Based on the results of Cross validation it can be seen that the modeling of the 3 data models divided into 3 cross were tested 10 times. The best "K" parameter was found in the "K" = 1 with the average accuracy of 93.312% and a good Cross Model was in Cross 2 with an average accuracy of 93.945%. And so, the modeling on Cross 2 will be used as sample for training data and testing data on comparative analysis of MKNN and KNN.

This stage will test every capability of KNN and MKNN algorithm, the data used is CCT data from social office of Pekanbaru city which consist of 7,395 data records and data modeling used is Cross 2 based on K-Fold Cross Validation modeling. It consists of 4 attribute variables and 1 class label. The first test is KNN algorithm will classify the data as much as 10 times, start from K = 1 to K = 10, by producing the test value with the highest accuracy is in the test K = 1 by 94.95% and with the average accuracy during the test is 93.94%. The next test is MKNN algorithm to classify the data as much as 10 times, start from K = 1 to K = 10, by producing the test value with the highest accuracy on K = 1 by 99.51% and with the average accuracy during the test is 99.20%. For more details see Table 4 below.

TABLE III. COMPARISON OF KNN AND MKNN ACCURACY

Value K	KNN	MKNN
K=1	94.95	99.51
K=2	94.41	99.11
K=3	94.41	99.23
K=4	94.48	99.39
K=5	94.50	99.23
K=6	94.36	99.27
K=7	93.58	99.11
K=8	92.83	99.11
K=9	92.83	99.06
K=10	93.10	98.98
Average	93.94	99.20

From Table 3 we will visualize the comparison of KNN and MKNN algorithms in data classification shown using line graph which can be seen in Figure 3 below.

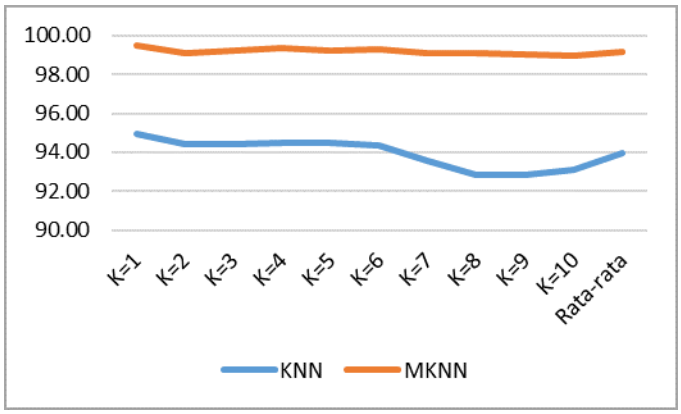

Fig. 3. Accuracy Measurement Results

IV. CONCLUSION

Comparative analysis of both MKNN and KNN was done with the aim of knowing the accuracy capability for classification from the two algorithms. And also to know the optimal data patterns obtained from k-fold cross validation into ideal data training and data testing. As was found in Table 3 and Figure 2, good data modeling will affect data accuracy. The calculation of accuracy uses the rules of Confusion Matrix. A good data modelling in this paper was found in cross 2 with accuracy of 93.945%. This modeling will be taken and used as training data and testing data to be tested in KNN and MKNN to analyze the accuracy ratio with result that the highest accuracy of KNN was 94.95% with average accuracy during the test was 93.94% while MKNN's highest accuracy was 99.51% and average accuracy during the test was 99.20%. so it can be said the ability of MKNN algorithm is better in terms of accuracy with the difference in accuracy by 5-7%.

ACKNOWLEDGMENT

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 the research can be done well.

REFERENCES

[1]. D.T. Larose. "Discovering Knowledge in Data An Introduction to Data Mining". Wiley Interscience, pp. 90-106. 2005

[2]. R. Agrawal. "K-Nearest Neighborn for Uncertain Data". International Journal of Computer Applications (0975-8887). 2014. Vol. 105 No. 11 pp 13-16.

[3]. Parvin, Hamid, Alizadeth, Hoseinali, and M. Behrouz, "A Modification on K-Nearest Neighborn Classifier". Global Journal of Computer Science and Technolgy. 2010. Vol. 10 No. 14, pp. 37-41.

[4]. Parvin, Hamid, Alizadeth, Hoseinali, Minati, M. Behrouz, and Bidgoli. "MKNN: Modified K-Nearest Neighborn". Proceedings of the World Congress on Engineering and Computer Science WCECS. 2008. ISBN: 978-988-98671-0-2 pp. 1-4.

[5]. Singh, Aruma, and Patel, Smita, Shukla,. "Applying Modified K-Nearest Neighborn to Detect Threat in Collaborative Information Information Systems". International Journal of Innovative Research in Science, Engineering and Technology. 2014. Vol. 3 No. 6, pp. 14141-14151.

[6]. S. Mutrofin, A. Kurniawandhani, A. Izzah, and M. Masrur, "Optimasi Teknik Klasifikasi Modified K-Nearest Neighborn Menggunakan Algoritma Genetika". Jurnal GAMMA. 2016. Vol. 10 No. 1, pp. 1-5.

[7]. R. Kohavi, "A Study of Cross Validation and Boostrap for Accuracy Estimation and Model Selection". Appears in the International Joint Conference on Artificial Intelligence (IJCAI). 1995. Pp. 1-7.

[8]. M. Last, "The Uncertainty Principle of Cross Validation" IEEE. 2006. Pp. 275-280.

[9]. J.K. Han, Micheline, J. Pei,. "Data Mining: Concepts and Techniques", Third Edition. 2011. British Library Cataloguing-in-Publication. Morgan Kaufmann. Pp. 235-236.

[10]. T.A. Singh, S. Namrata, "Speech Recognition Using Eulidean Distance", International Journal of Emerging Technology and Advanced Engineering. 2013. Vol. 3 No. 3, pp. 587-590.

[11]. Patil, R. Tina, and S. Sherekar, "Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification". International Journal of Computer Science and Applications. 2012. Vol. 6 No 2, pp. 256-261