Analysis of algorithm support vector machine learning and k-nearest neighbor in data accuracy

[1]Nuranisah, [2]Syahril Efendi, [3]Poltak Sihombing

Graduate Program of Computer Science,Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

nuranisahasriel123@gmail.com, syahnyata1@gmail.com, poltakhombing@yahoo.com.

Abstract. K-Nearest Neighbor is a method of lazy learning method which is a group of instances-based learning. K-NN searches by searching for groups of objects in the training data that are closest to the object on new data or testing data. Support Vector Machine is a learning machine method that works with the aim of finding the best hyperplane that separates two classes in input space. School Achievement is an achievement obtained by serious learning and discipline. The category of outstanding students is to get a good average score and not have an attendance list, especially Absent (A) and a list of late attendance at school can be classified to obtain information on the accuracy of the data being tested. In the testing process both methods obtained good accuracy results between the two methods, namely K-NN obtained an accuracy of 88.52% while SVM is 91.07%.

I. INTRODUCTION

School Achievement is an achievement obtained by serious study and discipline. The category of outstanding students is to get a good average score and not have an absentee list, especially Alpa (A) and the black notes listed in the school supervision supervisor are like fighting or some violations committed during the school study.The average score obtained by students in each semester can be classified to obtain information by adding an absentee list (Alpa) for 3 years to study at the school as an additional assessment of students, so as to get the accuracy of the data tested.

The classification has a function to make class predictions from an object whose classes are unknown. Classification is a pattern that is directed. (Raviya, 2013). The classification methods commonly used are Decision Tree, K-Nearest Neighbor, Naïve Bayes, Neural Network and Support Vector Machine (Sahu, 2011).

According to Nikulin (2008) classification is applied only to stronger training set data which has been explained in the "positive" class represented in the minority without losing its attributes. Whereas research conducted by Mladenic (1999) is research related to selecting features that contribute to classification using certain specifications and the ability to learn from classifiers in text data that is not evenly distributed.

In the classification process, the method used to obtain the best accuracy is the K-Nearest Neighbor and Support Vector method. Both methods will classify datasets on the average value of students with the addition of an assessment of the number of absentee lists (Alpa) and black notes obtained while accumulating knowledge in the school so that they get good accuracy results from both methods.

Ryci Rahmatil (2017) obtained the results of calculating the accuracy of test data manually processed using the Support Vector Machine method with polynomial kernel trick

having an accuracy rate of 43.33%. In Syahfitri's research (2015) Support Vector Machine has the advantage of analyzing data that is linearly distributed. Support Vector Machine has a good average accuracy rate so that many test data have no effect on the results of generalizations. In K-Nearest Neighbor it is necessary to determine the value of the parameter k, so that it has a lot of noise and is effective if the number of test data is large

Oki and Theopilus (2018) in a study of high-interest interest pathway classification using two datasets of ABC majors totaling 288 students and XYZ 280 students. Testing using the Support Vector Machine and Naive Bayes Classifier methods get better performance results using the ABC dataset. Meanwhile, according to Toto (2017) in his research to determine the choice of schools in the acceptance of new students using the K-Nearest Neighbor method produced 159 predictive data that approached the same only differed between the first and second choices and 16 data that had the same prediction between the first and second choices.

Raikwal and Kanak (2012) according to the research results obtained using cross-validation and Euclidean Distance processes to find accuracy and other performance parameters. The lack of the K-NN implementation phase gets good classification results, only not in textual data, the occurrence of performance parameters varies according to the size of the dataset. If the data size increases, the results obtained on K-NN are not good. While SVM values for accuracy and other performance parameters are not too dependent on the training cycle.

From the above explanations, there will be further analysis of the performance of the K-Nearest Neighbor method and Support Vector Machine in terms of classification of datasets. The dataset used in this study is a list of average grades of students from semester one to third grade and absentee list (Alpa) as well as blacklisted records from school guidance supervisors for 3 years to be classified so that they know the accuracy results when testing datasets that are.

II. PROBLEMS IDENTIFICATION

The background of the above research results in a problem statement that needs to be further analyzed about the classification process to predict student achievement because the data storage process uses Microsoft Excel so that the level of accuracy of the data requires a method that has accurate predictions, namely the K-Nearest Neighbor method and Support Vector Machine. In both methods Supervised uses patterns (patterns) from existing data to make predictions on data accuracy. So look for which one has a high ranking in the accuracy of the data between the two methods.

III. RESEARCH METHODS

a. Support Vector Machine

Support Vector Machine (SVM) one of the machine learning methods that works on the principle of Structural Risk Minimization (SRM) to find the best hyperplane and then separate two classes in input space. Basically, SVM is a linear classifier, then developed to be able to work in non-linear cases by incorporating kernel concepts in high-dimensional workspaces (Santoso, 2007).

The characteristics of SVM as described in the previous section are summarized as follows:

1. In principle, SVM is a linear classifier
2. Pattern recognition is done by transforming data in a higher dimensional space input space, and optimization is done in the new vector space. This distinguishes SVM from

the pattern recognition solution in general, which optimizes the parameters in the transformation results space which is lower than the input space dimension.

3. Implementing a Structural Risk Minimization (SRM) strategy
4. The working principle of SVM is basically only able to handle the classification of two classes.

A good method in solving data classification problems is a function of the Support Vector Machine. SVM problems are solved by solving the Lagrangian equation which is a dual form of SVM through quadratic programming (Tiananda, 2009).

Research conducted with computers uses a model of linear support vector machine (SVM) and nonlinear SVM. The set X={x_1, x_2,…x_m}, with X_i € R^n, i=1, ..., m, it is known that X is a certain pattern, that is, if xk is included in a class then labeled (is the target) y_k = -1. Thus the data provided is in the form of pairs (x_1, y_1),..,(x_m, y_m) € X x {+1,-1}. In learning problems, the collection of pairs is SVM learning data. Armed with the learning experience using the learning data, SVM must be able to determine the pattern (generalization) of *x € X*.

$$X_n = \frac{0.8(X-a)}{b-a} + 0.1 \quad \ldots 1$$

Where :

X_n = value - n
A = Low lift value
B = highest number value
0.8 and 0.1 = Determination

Using this equation, we can look for the values of data transformation x_1 (Attendance) and x_2 (Value). The basic problem of SVM is determining a hyperplane <w, x> + b = 0 separating xj data consisting of two classes, namely y_i = {+1, -1}, with maximum margins. Margin here is the distance between the hyperplane to each data class. Next, this hyperplane will be a decision function f (x) for the classification problem of the two classes above.

$$f(\emptyset(x)) = sign \ \ (w\emptyset(x) + b) = sign(\sum_{i=1}^{N} a_i y_i \, \emptyset(Xi)^T . \emptyset(x) + b) \quad \ldots 2$$

Where :

w = weight value
x = value of input variable
b = bias value

b. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) is a group of instance-based learning, K-NN is a lazy learning technique for searching groups of k objects in the most extensive training data on objects in new data or testing data (Chang, 2009).

K-Nearest Neighbor Algorithm is a method that functions to classify objects based on learning data which is the closest distance to the object being tested. Nearest Neighbor is an approach to finding cases by calculating the closeness between new problems and old problems in matching weights to the number of features available. To define the distance between two points, namely the point on the training data (x) and the point on the testing data (y), the Euclidean formula is used, as shown in the following equation:

$$D(x,y) \quad = \sqrt{\sum_{k=1}^{n} (X_k - Y_k)}$$

D is the distance between the points on the training data x and the testing data point y to be classified, where x = x_1, x_2, x_3, ..., x_i and y = y_1, y_2, y_3, ... y_i and represent the value attributes and n are attribute dimensions.

In the training phase, this algorithm only stores feature vectors and classifies sample training data. In this phase, the same features are calculated for testing the data classifications are unknown. The distance from the new vector to which all the training sample vectors are calculated and the closest number of k pieces is taken.

Steps to calculate the K-Nearest Neighbor Algorithm method:

1. Determine Parameter K (number of closest neighbors)
2. Calculates the square of the Euclid distance (query instance) of each object against the sample data provided
3. Sort these objects into groups that have the smallest euclid distance.
4. Collect Y categories (Naerest Neighbor Classification)
5. Using the main Nearest Neighbor category, you can predict the calculated query instance value.

c. Distance Model

Distance model is one way to measure the similarity between data. There are various kinds of distance models, including Chebyshev, Harmonic, Euclidean, Manhattan, Minkowsky, and so on. Following are some of the equations of the distance model:

Manhattan distance measurements using formulas:

$$D(x,y) \quad = \big||x-y|\big|1 = \sum_{j=1}^{N} |x-y|$$

Chebyshev distance measurements using formulas:

$$D(x,y) \quad = \big||x-y|\big|1 = \sum_{j=1}^{N} |x-y|^1$$

Euclidean distance measurements using formulas:

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

Minkowski distance measurements using formulas:

$$D(x,y) \quad = \big||x-y|\big|\lambda = \sum_{j=1}^{N} |x-y|^\lambda$$

Where:

D is the distance between data x and y.

N is the number of data dimensions.

λ is the Minkowsky distance parameter.

In general, Minkowsky is a generalization of existing distances such as Euclidean and Manhattan (Mergio & Casanovas, 2011). Lamda (λ) is a determining parameter and has a positive number from 1 to infinity (∞), if the value of λ = 1 then the Minkowsky distance space is equal to Manhattan according to Labellapansa (2016), and if λ = 2 the space is equal to Euclidean and Mergio (2008), if λ = ∞ is equal to the Chebyshev distance space (Rao, 2012).

Each distance measurement model has its advantages, Manhattan is very determined to detect outliers in the data, while Euclidean is suitable for determining the closest distance (straight) between two data. However, the Euclidean distance model is considered to be lacking in interpreting similarities between data (Pandit, 2011).

Traditional distance models are very fragile in determining the similarity, moreover in traditional distance models the value of attributes that are too large can mask the influence of other attributes, and most traditional distance models do not portray differences between data, especially in large data samples (Jo. 2017; Loohach & Garg. 2012; Pandit & Gupta. 2011).

Pan (2016) in his study suggested using a Harmonic distance model, where the distance model is considered better in describing the similarities between data. The measurement of harmony distance

$$D(\mathrm{x},\mathrm{y}) = \frac{1}{(\sum_{j=1}^{N} |x - y|)}$$

IV. RESULTS AND DISCUSSION

Testing the classification method is done by two preprocessing processes, namely in the process of Missing value for attributes that have a numerical value to be the average value of the attributes in the same column. If the missing value in a nominal value attribute is replaced by a value that has an equation with other attributes. The next process, which is cleaning, is done by eliminating data duplication.

In the next process by giving a categorical form of each attribute / subset to facilitate the mining process and the accuracy of the classification.

The following are the results of prepocessing data from the dataset of Sekolah Menengah Pertama Negeri 22 Medan.The use of Split data and performance classification to determine the performance of the algorithm that was tested during the classification process and get the results of its accuracy. The working procedure of Support Vector Machine on the rapid miner for the first step is to enter a dataset that has been formatted into Excel (Read Excel), then do Split Data by separating the dataset into two parts, 80% training data and 20% test data randomly. Enter the SVM model after the Split Data results are done for training data and apply the model as a Performance operator. The following are the results of data distribution on the SVM method for the dataset.

Table 1. Results of SVM Method Data Distribution of Value Dataset

No	Predicate	Confidence (Good)	Confidence (Quite)	Prediction (Predicate)	Sem. 2	Sem. 4	Sem. 6	Absent	overdue
1	Quite	0.448	0.552	Quite	80.0	80.1	83.8	6.0	2.0
2	Quite	0.420	0.580	Quite	79.5	79.3	83.3	2.0	6.0
3	Quite	0.410	0.590	Quite	79.2	79.4	82.5	2.0	9.0
4	Quite	0.474	0.526	Quite	80.8	80.1	82.4	2.0	6.0
5	Good	0.529	0.471	Good	79.8	81.0	85.7	1.0	7.0
6	Good	0.514	0.486	Good	80.9	80.7	83.4	1.0	8.0
7	Good	0.679	0.303	Good	81.9	82.3	86.3	3.0	1.0
8	Good	0.373	0.627	Quite	77.6	79.7	80.2	2.0	8.0
9	Good	0.305	0.695	Quite	76.9	77.5	79.4	10.0	9.0
10	Quite	0.440	0.560	Quite	79.7	77.5	81.2	1.0	3.0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
61	Quite	0.345	0.655	Quite	78.8	79.2	79.1	3.0	3.0

a. Confusion Matrix SVM

Confusion Matrix is one method to measure the performance of a classification method that contains information to compare the results of classification carried out by the system with the results of classification.

Table 2. Confusion Matrix The SVM Method Uses The Student Dataset

Performance Classification	**Predicted Class**	
Actual Class	Predicted.Good	Predicted. Quite
Actual. Good	20 (True Positive)	3 (False Negative)
Actual. Quite	2 (False Positive)	31 (True Negative)

Based on table 2, it is continued by calculating the Accuracy value of the classification of the SVM classification model using the Student Value Dataset.
The following results are calculated:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{20+31}{20+31+2+3} = \frac{51}{56} = 0.9107*100\% = \mathbf{91.07\%}$$

It is known that the highest result of the closeness between the predictive value of the class and the actual value of the correct class from the SVM classification model of the dataset is 91.07%.

b. Confusion Matrix K-NN

Confusion Matrix is one method to measure the performance of a classification method that contains information to compare the results of classification carried out by the system with the results of classification.

Table 3. Confusion Matrix The K-Nearest Neighbor Method Uses The Student Dataset

Performance Classification	**Predicted Class**				**Class precision**
Actual Class	Predicted.Good	Predicted. Quite	Predicted. Better	Predicted. Low	
Actual. Good	18	1	0	0	94.74%
Actual. Quite	4	33	0	2	84.62%
Actual. Better	0	0	3	0	100.00%
Actual. Low	0	0	0	0	0.00%
Class recall	81.82%	97.06%	100%	0.00%	

Based on table 3, explain using the K-NN algorithm. The calculation of the proximity of the old case on the dataset has an accuracy rate of 88.52%.

V. CONCLUSION

The SVM concept explains in a simple way to look for the best hyperplane that functions as a separator of two classes in input space. In the dimensional space, input data x (i = 1 ... k) which belongs to class 1 or class 2 and the corresponding label becomes -1 for class 1 and +1 for class 2. While K-NN is a nonparametric model which means do not assume something about the distribution of instances in the dataset. Usually this model is difficult if interpreted,

only has excess as the class's decision line becomes flexible and linear. The following will explain the results of the analysis of the two methods.

REFERENCES

[1]. Chang, C., Wu, Y., Hou, S.2009. *Preparation and Characterization of Super paramagnetic Nanocomposites of Alumino silicate/Silica/Magnetite,* Coll. Surf. A336: 159,166.

[2]. Hand, David., Mannila, Heikki., Smyth, Padhraic. 2001. *Principles of Data Mining,* MITPress, Cambridge, MA. ISBN 0-262-08290-X.

[3]. Kirdat, Tejshree., Patil, V.V. 2016. *Apllication Chebyshev Distance amd Minkowski Distance to CIBR Using Color Histogram. International Journal of Innovative Research in Technology (IJIRT).* Volume.2, Issue.9. ISSN:2349-6002.

[4]. Larose, D.T. 2005.*Discovering Knowledge in Data: An Introduction to Data Mining, John Willey & Sons*. Inc. pp. 129-240

[5]. Loohach, R. & Garg, K. July 2012. Effect of Distance Functions on K-Means Clustering Algorithm. *International Journal of Computer Applications*.Volume.49, No.6.

[6]. Mladenic, D &Grobelnik, M. 1999. *Feature Selection for Unbalanced Class Distribution and Naïve Bayes.* Department of Intelligent Systems, J.StefanInstitue, Jamova 39,1000 Ljubljana, Slovenia.

[7]. Nikulin, V. 2008. *Classification of Imbalanced Data with Random Sets and Mean-Variance Fltering*. International Journal of Data Warehousing and Mining, 4(2):63–78.

[8]. Nur Asiyah, Siti., Kartika, Fithriasari. 2016. *Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K-Nearest Neighbor.* Jurnal Sains dan Seni ITS. Volume.5, No.2.

[9]. Pandit, S. & Gupta, S. A. December 2011. Comparative Study on Distance Measuring Approaches for Clustering. International *Journal of Research in Computer Science* **2**(1): pp. 29-31.

[10]. Raikwal.J.S&Saxena, Kanak. 2012. *Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set.* International Journal of Computer Applications (0975-8887). Volume.50. No. 14.

[11]. Raviya.Kaushik H &Gajjar, Biren. 2013. *Performance Evaluation of Different Data Mining Classification Algoritma Using WEKA. Indian Journal of Research*. Volume. 2. Issue.1. ISSN: 2250-1991.

[12]. Sahu, Mridu.,Nagwani. N.K., VermaShrish., Shirke. Saransh. 2015. *Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal.International Journal of Knowledge Engineering*, Volume.1, No.2.

[13]. Turban, E. et al. 2005. *Decision Support and Intelligent Systems*.Upper Saddle River, NJ:Prentice Hall. ISBN: 978-81-203-2961-4, pp. 263-264.

[14]. Vapnik, V. 1995. The Nature of Statistical Learning Theory, Springer-Verlag.