Analysis of performances k-nearest neighbor for regulate learning

Rahmad Syuhada[1,*], Herman Mawengkang[2], Maya Silvi Lydia[1]

[1]Department of Computer Sciences, Universitas Sumatera Utara, Medan 20155
[2] Department of Mathematics,Universitas Sumatera Utara, Medan 20155, Indonesia

[1]rahmad.syuhada@yahoo.co.id

Abstract. Self-regulation is a process to turn and arrange thoughts, behavior and deep emotion of achieving goal, predict the outcome of student learning achievements with the criteria for assessing which cognitive and affective. This can be seen in the students who has not reached minimum criteria 100 % of the final test which received by students. Research now applied performance and looking for value *k* best with the nearest distance related the ability self-regulation students to learning achievements. Learning outcomes low is something that should not be left, because bad for the future students. Therefore need to followed up in improving self-regulation learning. So, did analysis by using the method k-nearest neighbor.

1. Introduction

Achievement was the success of a person in a daily activity, In education sector, obtained called achievement students learning achievements. These, to meet the certainly is easily obtained, because needs to students in ability from regulations to the support so that the can soon be realized. Self-regulated learning is the process of constructive of individuals that are active in setting a course of learning who want to be achieved and doing repairement, monitoring regulations and in control of cognition, motivation and behavior and guided by the purpose of who want to be achieved[1].

A function of classification is for the classification of most common as a function of data mining .Of business process produce analysis, risk management, and are aimed at classifies the target classes into selected in the category. Classifications is to the process of finding a model or function that explain and to differentiate a class or the concept. with the aim that a model that is obtained can be used to predict any a class or the object that having label class are not known. A model that is sent down are based on sound analysis. The model has been sent down can be represented in various forms like rule if-then classifications, decision tree, a formula mathematics or neural network[2].

The regulations themselves in study on the in the end will make students active in his course. Zimmerman(1989) expressing that with any regulations selfin learning, students will try to reach a destination learn by activate and maintain the mind, behavior and emotions. In addition, regulations self in to study also pertaining to a change to be better in the mind, feeling as well as actions planned and the reciprocal adjusted to the achievement of a goal personal[3].

In his studies of students performance prediction using knn and naive bayesian, do classifications to help the education ministry in raising prediction performance achievement early students and teachers , so that we can take evaluation in determining levels learning value students. The produce that is better than a naive bayes k-nearest neighbors by receiving

the highest 93,6 % accuracy[6].One example of psychological factors in improve learning outcomes students the ability to students, regulations the same thing told by that, by woolfolk one factor affecting someone achievement sector is making to regulations, the ability produce, mind feelings and the act of, plan and are adapted directly to an end[3].

On this research will predict the outcome students achievement with the assessment criteria cognitive and affective. This is evident in the value of students unable to reach minimum criteria 100 % from final test in the last result which is the value of the report card. received by students learning outcomes that low is one thing that must not go unpunished, because bad for future students. Because it is necessary for follow up in improving student learning regulations.To get the performance of the data precision , then required k-nearest neighbor methods to see the k which one has accuracy better.

2. Research background

2.1 Data mining

Data mining is the search automatically information in a data storage by large size. A term often used are knowledge discovery in databases, knowledge extraction, pattern analysis, data archeology, data dredging, information harvesting, and business intelligence. Data mining involves the search for knowledge of a data supersized through statistical methods, machine learning, and artificial algorithm.The most important of a process with the data mining is feature selection and process selection pattern recognition of a system databases.

2.2 Data mining stage

As a process called a series of, and data mining can be divided into several phase. The phase in which users interactive is directly involved or through knowledge base many people did to data mining as a synonym for other terms popular use knowledge discovery databases or kdd. While others see data mining only as an important step in the process of discovery. knowledge As listed in the book “Data mining concepts and techniques”[2].The process of knowledge discovery consisting of several stages of, a phase among other things as follows.

2.2.1 Data cleaning

Remove noise and data inconsistent, in general the data obtained either from a company database and the, experiment having stuffing imperfectly as, missing data the data invalid or also just one type. Data that is not relevant is kindly disposed because they can reduce quality or accuracy of the data mining.

2.2.2 Data integration

In which some sources of data can be combined, the integration of the data is carried on an attribute that identify entity that unique as the name of, attributes type of product, and the number of customers

2.2.3 Data selection

Where data relevant for analysis is obtained from a database, selection of data and a bunch of the data is done before the operational stage, information the results of the selection process then used for data mining and kept on file separate from operational database.

2.2.4 Data transformation

Where data converted and consolidation into a form in accordance with a summary or aggregation with to carry out. The process is commonly called binning, where 9 first the election of data needed by mining techniques used data. Transformation and selection very determining the quality of data from the mining data.

2.2.5 Data mining

Where vital processes can known with practicing to extract the data. Techniques and methods the mining varied. So the right method or a depend the goal and processes.

2.2.6 Pattern evaluation

Knowledge discovery database is a form of the evaluation to identify by pulling a pattern representing knowledge action on the relatedness measure. This stage includes check whether the or information found counter to the or is hypothesized the previous.

2.2.7 Knowledge presentation

Where is the representation of knowledge and visualization technique used to present the knowledge to a user,mined how to make formulations decision or action of the result analysis obtained.In a presentation, visualization can help to communicate the results of data mining.

3. Proposed method

3.1 K-Nearest Neighbor

K-Nearest Neighbor algorithm is a method of conducting classifications education towards an object based on the data is closest to the object. Data projected into the learning infinite-dimensional, with each dimensions represent features of data, space is divided into parts based on data classifications learning, the purpose of these algorithms to classify new objects is based on attributes and samples from the data training, a point in the is characterized the classes of "c" if the classes of "c" is classifications most common to find *k* on the neighbors. Near or away neighbours euclidean are calculated based on the distance.

Steps K-Nearest Neighbor algorithm indicated as follows:

1. Specify the value of *k* , which is the number of nearest neighbour.
2. Counting the square of euclidean distance any object to sample data.

$$\boldsymbol{D(x,y) = \sqrt{\sum_{k-1}^{n}(X_k - Y_k)}^{\,2}} \qquad \text{.....(1)}$$

Where :

d = distance
x = data training
y = data testing
n = the number of individual attributes between 1-n
f = the function of an attribute i between cases x & y
Wi = a weight that was given to attribute ke-i

3. The distance between object "x"and "y" is defined as "Dxy", where "Xi" is record that will it is predicted and "Yi" is recorded data pattern while value "n" defined as the quantity of attributes and value "I" refer on a record ke-i.
4. Gather category "y"(classifications) nearest neighbor.
5. By the use of category the majority , it can be it is predicted the query instance who has been calculated.

3.2 Research designof the k-nearest neighbor

Shown in Figure 1, Block diagram of the k-nearest neighbour the following six stages：

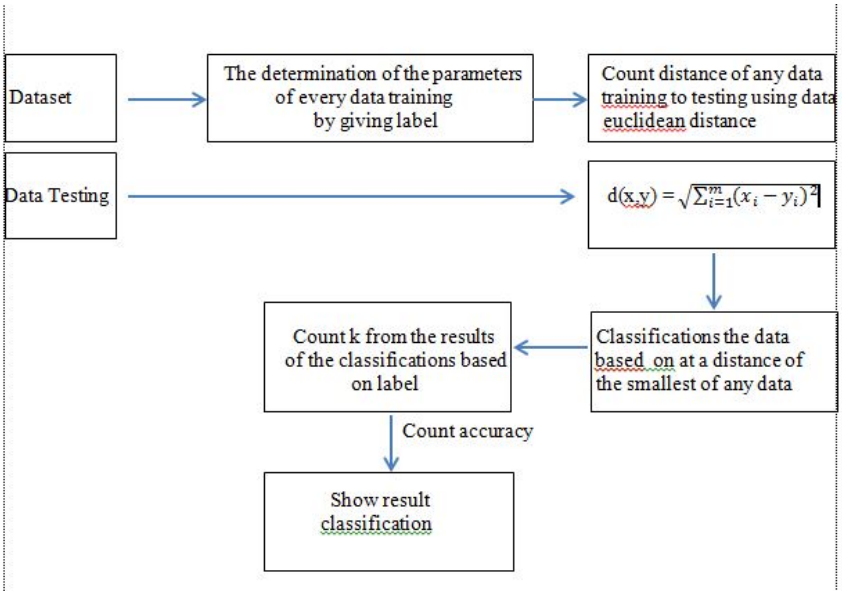

Figure 1. K-Nearest Neighbor phase includes the following six stages.

In testing is provided by way of 590 data , every of testing was definitely data will be tested uses the method k-neirestneighbor by the use of the value of k that an appointed term and , 3,5,19,15,20 .The results of the testing can be seen all said that they had in the following. Based on in figure on 1 according, below can be seen an example dataset, then will be the attributes of parameters in the process will be a distance the modelling on, and cleaning data was undertaken by removing an attribute that is used and reduce the effects a noise during the process.Next , separate data to the set and training is meant to make of testing was definitely model later on obtained has the ability in conducting a generalization that good data classifications.

In k-nearest neighbor, when distance have been found through a system of accounts with distance euclidean. Then the distance the best claim to be employed in sort , based on k appointed of testing was definitely is compared with the data testing . In figure 2, flowchart k-nearest neighbour.

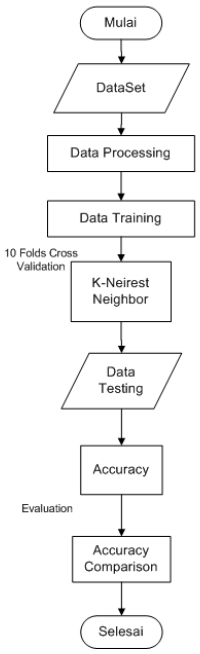

Figure 2. Flowchart of k-nearest neighbor

4. Result and analysis

In this section grouping step will be conducted k-nearest algorithm neighbor by the use of the value of the nearest distance .In this study , we will assume the nearest neighbor with a value of k .In this research discuss how to find the value of k with the kind of accuracy that good .In the matter of validating classifications , this study using formulas distance euclidean. Then the value of distance is in the range of the smallest sorted , distance whose value the least , then arranged based on the results of choice.

Next, separate data to testing sets and training is meant to make model later on obtained have the ability generalization that both in do classifications data.

Table 1. Data Training

NIS	Metakognitif	Motivation	Behavior	Achieved	Not Achieved	Study Choices
20173496	90	93	93	90	1	Kognitif
20173585	90	92	93	80	1	Kognitif
20173459	80	92	90	87.5	1	Kognitif
20173456	80	92	90	86	1	Kognitif
20173428	80	90	90	87.5	1	Kognitif
20173490	80	58	90	90	0	Afektif
20173489	80	92	89	86	1	Kognitif
20173526	100	92	93	89	1	Kognitif
20173362	80	56	90	91.5	0	Afektif

Next, data testing to the test results with the best model from k-nearest neighbor. Then from this data will be obtained a conclusion an data will enter into a category four, third, both, or first based on grades k that has been in the process of. The process of reckoning k-nearest neighbor using equation eulidean distance of testing was definitely the following data can be seen in table2 in the following :

Table 2. Data Testing

NIS	Metakognitif	Motivation	Behavior	Achieved	Not Achieved	Study Choices
20173482	80	58	95	88	1	?
20173483	90	80	90	78	0	?
20173498	100	58	92	85	0	?
20173499	80	92	89	86	1	?

In the study used taken from students self-regulation and data normalizing. this here the testing k-nearest algorithms based on the k value and accuracy. After doing preprocessing data with how to separate the data do not need, then process data ready training to classify based on the nearest distance with the data testing. Using k = 3, k = 5, k = 9, k = 15 and k = 20 to find value in optimal. k of the value compared k which that produces the percentage of higher.

Table 3. The number of class category by determining label regulation data modelling.

Classification Model	
Category_Cognitive	390*instances*
Category_Affective	200*instances*
Sum of weights	590

Table 4. k-nearestneighbor test results with the process of the amount k-fold cross validation

	k-fold cross validation	*Accuracy*	*RMSE*	*RRSE*
Algorithm K-Nearest Neighbor	*k=3*	100%	0,0083 %	1,7445 %
	k=5	100%	0,0041 %	0,8701 %
	k=9	100%	0,0021 %	0,4445 %
	k=15	100%	0,0348 %	7,3212 %
	k=20	100%	0,0348 %	7,3520 %

5. Conclution

From our analysis testing with a performance method k-nearest neighbor using data self-regulation, learning in this research model produced tried to get a RMSE. Accuracy and value the classification methods this data obtained by means of clustering.In the results of the determination of the nearest distance method k-nearest neighbor obtained using value by the number of k=3 Accuracy is 100 % and value RMSE is 0,0083, with clusters of data k=5 Accuracy is 100 % and value RMSE is 0,0041, with clusters of data k=9 Accuracy is 100 % and value RMSE is 0,0021, with clusters of data k=15 Accuracy is 100 % and value RMSE is 0,0348, with clusters of data k=20 Accuracy is 100 % and value RMSE is 0,0348.

References

[1] Pintrich, P. R., Roeser, R.W., & De Groot, E. V.1994. 'Classroom and individual differences in early adolescents' motivational and self-regulated learning. Journal of Early Adolescence, 14(2), 139–161.

[2] Han, J., & Camber, M. 2006. Data Mining Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.

[3] Zimmerman, B. J. (2000). Attaining self-regulation: A social-cognitive perspective.In M Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 13–39).
[4] Amrane, M., Oukid, S.,Gagaoua, I. &Ensari, T. 2018. *Breast Cancer Classification Using Machine Learning*. IEEE 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 115-116
[5] Fayyad, U., G.P. Shapiro, and P. Smyth.1996. From Data Mining to Knowledge Discovery in Database. AI Magazine pp. 37-53
[6] Ihsan, A. &Maghari, A.Y. 2017. "Students Performance Prediction Using KNN and Naïve Bayesian" *2017 8 th International Conference on Information Technology (ICIT),* pp.13-19
[7] Jadhav, S.D., &Channe, H. P. 2016. "Comparative Study of K-NN , Naïve Bayes and Decision Tree Classification Techniques". *International Journal of Science and Research* vol. 5,no.1
[8] MacLennan, Z. Tang and B. Crivat. 2008."Data Mining with SQL Server 2008" Wiley Publishing, Indiana, 2009.
[9] Schunk, D, H. 2005. Self-regulated learning: the educational legacy of paulpintrich. Journal of Educational Psychology, 40(2), 85-94.
[10] Rani, Y., Manju.&Rohil.2013. Comparatie Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.8. The SIJ Transactions on Computer Science Engineering & its Applications, (CSEA), pp.1115-1122