Journal of Engineering and Applied Sciences 13 (15): 6157-6161, 2018
ISSN: 1816-949X

A Review of Several Algorithms for Data Mining

Yasella Dina Aprilia, Roswan Latuconsina and Tito Waluyo Purboyo
Department of Computer Engineering, Faculty of Electrical Engineering,
Telkom University, Bandung, Indonesia

Abstract: Data mining is means of processing explanation from the database to find a pattern from big data. Data mining is needed in the world of business and non-business. There are two techniques of data mining that is classification and clustering. Any data to be extracted using data mining would require a suitable method, therefore this study explain a review is needed to determine which method is suitable.

Key words: Classification, clustering, suitable, processing, explanation, non-business

INTRODUCTION

The process for processing information from a database to find patterns of big data is called data mining. Data mining is used for various purposes for example for banking, insurance. Data mining is aimed at reducing costs, accelerating computing time and etc.

According to MIT Technlogy Review data mining is needed to change the world. Larose (2005) In fact there are many methods not only classification and clustering but in this study will be limited ie only classification an klustering because two methods are often used.

The classification method aims to identify classes that have objects of some of the same descriptive features. They find usefulness in various human activities and especially in automated decision making (Hssina *et al*., 2014). Different from classification, clustering techniques to automatically group data without needing to give class labels.

Each method must have the advantages and disadvantages. From the review will be known what the difference is from all the methods.

MATERIALS AND METHODS

In this study, review will discuss about data mining algorithms that exist for classification and clustering. Classification is studying a set of data, so that, rules are generated that can recognize new data that have never been found. The classification model is built on the knowledge of an expert. The algorithm belonging to the deepest of this classification group are KNN, MKNN, ID3, C4.5, CART, Naive Bayes, SVM.

KNN: KNN is a simple algorithm. This algorithm is just looking for the closest K object training data to a new data object.

This Euclidean distance formula below is used to calculate the distance of two training and test objects (Singh and Patel, 2014):

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

Where:

$d(x, y)$ = Euclidean distance
x_i = Data 1
y_i = Data 2
i = i-data
n = Amount of data

KNN is very easy to implement just by setting one parameter k and working locally but the weakness is very sensitive to noise and data.

MKNN: Of the lack of KNN is very sensitive to sound and data, so, made a new algorithm is Modified KNN algorithm (MKNN). The MKNN phase is as follows:

Do the distance calculation: To do the distance calculation using euclidean formula, so that, between two points in training data and re data can be calculated (Eq. 1) (Singh and Patel, 2014).

Validity of training data: To calculate the amount of all data points of training that is done validity. The affect in this validity is the nearest neighbor of any data (Parvin *et al*., 2008):

Corresponding Author: Yasella Dina Aprilia, Department of Computer Engineering, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia

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

To calculate the sameness between points x and number of data to i using the function S (Singh and Patel, 2014; Thakur and Sahayam, 2013):

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

Weight voting: The validity of each training data multiplied with the weight of Euclidean distance will be determine weight voting. In the MKNN method, the counting of the sound weights of each of the neighbors is in Eq. 4 (Mutrofin *et al*., 2015):

$$W(x) = \text{Validalitas}\ (x) \times \frac{1}{d_{e+0.5}} \quad (4)$$

The comparison between KNN and MKNN: The difference between KNN and MKNN is in the ratio of accuracy. The need for KNN and MKNN analysis is to know the optimal data pattern. From various experiments that have been done, MKNN has a higher accuracy of KNN.

ID3 algorithm: ID3 is a decision tree learning that searches gluttonously, so, it is not necessarily optimal. The ID3 algorithm uses the recursive function. To classify upcoming samples using trees. Stage to create a tree using the acquisition of information from the data training and then do classification test data. The attribute used by ID3 is the nominal attribute to classify without losing the value (Hardikar *et al*., 2012).

The power of ID3 is to have a clear concept and easy to understand and easy to implement by using recursive algorithm. The pseudo code in ID3 is simple (Hssina *et al*., 2014). The weakness of ID3 is difficult to implement with a very big data set and simple to overfit.

C4.5 algorithm: Created to overcome the disadvantages of the ID3 algorithm. The ID3 degradation is meant here is the ID3 algorithm is too sensitive to the features of a large amount of data. To conduct an internet search, this ID3 weakness must be overcome (Hssina *et al*., 2014). The formula to find gain ratio:

$$\text{Gain}(S, A) = \text{Entropy}(s) - \sum_{i=1}^{n} \frac{|s_i|}{|s|} + \text{Entropy}(S_i) \quad (5)$$

Where:

S = Space (data) sample used for training
A = Attribute
$|S_i|$ = Number of samples for value V
$|S|$ = Total number of sample data
Entropy (S_i) = Entropy for samples having value i
n = The number of S partitions i starting from = 1, 2, ... n

where, Entropy is:

$$\text{Entropy}\ (S) = \sum_{i=1}^{n} -P_i.\log 2p_i \quad (6)$$

Where:

S = The set (dataset) of the case
n = The number of S partitions
p_i = The probability of getting from sum (Yes) divided by total case i starting from = 1, 2, ..., n

The formula to find splitInfo:

$$\text{SplitInfo}\left(S, A\right) = -\int_{i=1}^{n}\left(\frac{S_i}{S}\right)\log_2\left(\frac{S_i}{S}\right) \quad (7)$$

Where:

S = Space (data) sample used for training
A = Attribute
Si = Number of samples for attribute i starting from = 1, 2, ..., n
n = The number of S partitions

C4.5 uses a set of data training as in ID3 to create a decision tree. The stage of making a decision tree is choosing an attribute as a root, creating a subsidiary in every case, dividing the matter in a subsidiary, repeating the means of every subsidiary, so that, it has the same class. Calculates total entropy, then calculates the gain. The highest gain becomes the root attribute (Hssina *et al*., 2014).

CART: CART is a supple process to illustrate how variable Y dispart vector X. The CART Model uses a binary tree to splite the approximate into a particular set where the Y distribution is continuous. The leaf node of the tree corresponds to a different division region determined by the separation rule corresponding to each internal node. The Y distribution at the node is determined by moving the root of the tree to the leaf node. To specify the CART branch attribute using Gini Index. Select an attribute whose Gini index is minimum after separated (Hssina *et al*., 2014):

$$\text{Gini}(S) = \sum_{i=1}^{K} \frac{|S_i|}{|S|}\left(1-\frac{|S_i|}{|S|}\right) = \int_{i\neq j}^{\frac{|S_i|x|S_j|}{S^2}} \quad (8)$$

Naive Bayes classifier Bayesian: Classifier that is classified in statistics in probability. Bayesian classifications show high accuracy probability equation:

$$\frac{P(H|X) = (P(X|H)\,P(H))}{(P(X))} \quad (9)$$

X is a data tuple, in terms of "proof". H is some hypothesis when it wants to decide P(H|X), the probability that hypothesis H continues is specified "proof" or observed tuple data. X. Looks for the possibility that X is a class X tuple, we must know the illustration attribute X (Fatahillah *et al*., 2017).

Support vector machine: SVM is maximizing margin maximizing minimum distance to nearest instance. SVM algorithm uses training data to produce functions used in new cases.

SVM only supports binary classifications but is used to solve multi-class classification problems. Binary SVM K*(K-1)/2 uses one-to-one encoding (Bishop, 2006; Rehana, 2017).

The SVM power of having a high data generalization capability, capable of producing a good classification model and relatively easy to implement.

The svm weakness is difficult to implement with a large number of samples and generally hanay formulated to solve the problem of classification of two classes.

Clustering is grouping data automatically without being notified of class labels but if the class label is known it can be used. The algorithm belonging to the deepest of this clustering group are k-means, k-modes, k-medoids.

k-means: k-means to minimizes the amount of squared distance between the data and the cluster center to do the grouping. The k-means algorithm works in four steps:

- The set of data to be clustered, selected randomly selected units as initial centroids
- Any object that is not a centroid, inserted into the nearest cluster based on the specific size of a given
- Each centroid is updated based on the average of the objects present in each cluster
- The second and third steps are repeated until all centroids are stable

The disadvantage of using k-means is that with the initial centroid generated randomly then repaired based on the mean distance, k-means not able to produce good centroid (Manivannan and Pevi, 2017).

k-modes: k-modes is k-means an algorithm that uses the average size that can only be applied to the set of objects with attributes converted to numeric from which k-modes algorithm is formed. Denoting X = $\{X_1, X_2, ..., X_n\}$ are clusters (sets), Q is the center for X:

$$D(X, Q) = \sum_{i=0}^{n} d1\,(X_i, Q) \quad (10)$$

The D(X, Q):

$$fr(Aj = q|X) \geq fr(Aj = ck, j|X)\ \forall q = ck, j\ \forall j = 1, 2, ..., K \quad (11)$$

Toan and Yasushi, (2016) k-medoids is computationally more difficult than k-means as it calculates k-medoids using frequency occurrences (Park and Jun, 2009). k-medoids have potentially important characteristics that the center lies between the data points themselves. The k-medoids algorithm is the grouping algorithm associated with k-means algorithm and medoids algorithm. K-means and k-medoids algorithms are clustering.

But there are differences between k-means anda k-medoids, k-means to minimize the total square root and k-medoids to minimize the number of point differences in clusters and points at the center.

k-medoids selects datapoint as the center (Batra, 2011; Gazalba and Reza, 2017). The consequence of k-medoids is that the computational complexity becomes higher.

RESULTS AND DISCUSSION

In Table 1 discuss the algorithms used in data mining and include the researcher of the reference. There are 10 methods listed in Table 1 in accordance with the above explanation. Table 1 is useful to facilitate the reader in understanding this study.

Table 1: Resume algorithm for data mining

Methods	Researcher name	Discussion
KNN	Okfalisa, Mustakim, Ikbal Gazalba, Nurul Gayatri Indah Reza	KNN is very easy to implement just by setting one parameter k and working locally but the weakness is very sensitive to noise and data
MKNN	Okfalisa, Mustakim, Ikbal Gazalba, Nurul Gayatri Indah Reza	MKNN algorithm is better than KNN algorithm in the ratio of accuracy
ID3	Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali	ID3 is a decision tree learning that searches gluttonously, so it is not necessarily optimal. The ID3 algorithm uses the recursive function. To classify upcoming samples using trees
C4.5	Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali	The C4.5 algorithm is created to resolve the disadvantages of the ID3 algorithm. The ID3 degradation is meant here is the ID3 algorithm is too sensitive to the features of a large amount of data. To conduct an internet search, this ID3 weakness must be overcome
CART	Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali	In CART, the Y distribution at the node is determined by moving the root of the tree to the leaf node. CART branch attribute using Gini Index
Naive Bayes	Naufal Riza Fatahillah, Pulut Suryati,Cosmas Haryawan	Classifier that is classified in statistics in probability. Bayesian classifications show high accuracy
SVM	Hasin Rehana	SVM is maximizing margin maximizing minimum distance to nearest instance. The svm power of having a high data generalization capability, capable of producing a good classification model and relatively easy to implement. The SVM weakness is difficult to implement with a large number of samples and generally hanay formulated to solve the problem of classification of two classes
k-means	P.Manivannan, Dr. P. Isakki @ Devi	k-means with the initial centroid generated randomly then repaired based on the mean distance, k-means not able to produce good centroid
k-modes	Nguyen Mau Toan, Inoguchi Yasushi	K-modes is k-means an algorithm that uses the average size that can only be applied to the set of objects with attributes converted to numeric, from which k-modes algorithm is formed
k-medoids	Prihandoko Bertalya, Muhammad Iqbal Ramadhan	Denoting $X = \{X_1, X_2, ..., X_n\}$ are clusters (sets), Q is the expected center for X K-medoids have potentially important characteristics that the center lies between the data points themselves. The consequence of k-medoids is that the computational complexity Iqbal becomes higher

CONCLUSION

In this study, we have review about algorithm for data mining. Algorithm for data mining has two method, there are classification and clustering. There are differences in each algorithm. Each algorithm has advantages and disadvantages. From the level of accuracy algorithm in classification there are KNN, MKNN, ID3, C4.5, Cart, Naive Bayes, SVM. From the level of accuracy in classification, C4.5 is better.

Algorithm in clustering there are k-means, k-modes and k-medoids. From the level of accuracy in clustering, k-medoids is better.

REFERENCES

Batra, A., 2011. Analysis and Approach: K-means and K-medoids data mining algorithms. Proceedings of the 5th IEEE International Conference on Advanced Computing and Communications Technologies (ICACCT'11), November 5, 2011, IEEE, New Jersey, USA., pp: 274-279.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, Germany, ISBN:978-0-387-31073-2, Pages: 738.

Fatahillah, N.R., P. Suryati and C. Haryawan, 2017. Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. Proceedings of the 2017 International Conference on Sustainable Information Engineering and Technology (SIET), November 24-25, 2017, IEEE, Malang, Indonesia, ISBN: 978-1-5386-2183-7, pp: 128-131.

Gazalba, I. and N.G.I. Reza, 2017. Comparative analysis of k-nearest neighbor and modified K-nearest neighbor algorithm for data classification. Proceedings of the 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE'17), November 1-2, 2017, IEEE, Yogyakarta, Indonesia, ISBN:978-1-5386-0659-9, pp: 294-298.

Hardikar, S., A. Shrivastava and V. Choudhary, 2012. Comparison between ID3 and C4. 5 in contrast to IDS. VSRD. Intl. J. CS. IT., 2: 659-667.

Hssina, B., A. Merbouha, H. Ezzikouri and M. Erritali, 2014. A comparative study of decision tree ID3 and C4. 5. Intl. J. Adv. Comput. Sci. Appl., 4: 13-19.

Larose, D.T., 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, New Jersey, USA., Pages: 222.

Manivannan, P. and P.I. Devi, 2017. Dengue fever prediction using K-means clustering algorithm. Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS'17), March 23-25, 2017, IEEE, Srivilliputhur, India, ISBN:978-1-5090-4779-6, pp: 1-5.

Mutrofin, S., A. Izzah, A. Kurniawardhani and M. Masrur, 2015. [Optimization of the modified k nearest neighbor classification technique using genetic algorithms (In Indonesian)]. J. Gamma, 10: 1-5.

Park, H.S. and C.H. Jun, 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl., 36: 3336-3341.

Parvin, H., H. Alizadeh and B. Minaei-Bidgoli, 2008. MKNN: Modified k-nearest neighbor. Proceedings of the 2008 World Congress on Engineering and Computer Science (WCECS'08) Vol. 1, October 22-24, 2008, International Association of Engineers (IAENG), San Francisco, California, USA., ISBN: 978-988-98671-0-2, pp: 1-4.

Rehana, H., 2017. Bangla handwritten digit classification and recognition using SVM algorithm with HOG features. Proceedings of the 3rd International Conference on Electrical Information and Communication Technology (EICT'17), December 7-9, 2017, IEEE, Khulna, Bangladesh, ISBN: 978-1-5386-2308-4, pp: 1-5.

Singh, A. and S.K. Patel, 2014. Applying modified K-nearest neighbor to detect insider threat in collaborative information systems. Intl. J. Innovative Res. Sci. Eng. Technol., 3: 14146-14151.

Thakur, A.S. and N. Sahayam, 2013. Speech recognition using euclidean distance. Intl. J. Emerging Technol. Adv. Eng., 3: 587-590.

Toan, N.M. and I. Yasushi, 2016. Audio fingerprint hierarchy searching on massively parallel with multi-GPGPUs using K-modes and LSH. Proceedings of the 8th International Conference on Knowledge and Systems Engineering (KSE'16), October 6-8, 2016, IEEE, Hanoi, Vietnam, ISBN:978-1-4673-8930-3, pp: 49-54.