Analysis of Iterative Dichotomiser 3 Algorithm Uses Fuzzy Curves Shoulder as a Determinant of Grade Value

Arina Prima Silalahi[1], Zakarias Situmorang[2], Syahril Efendi[3], and Eva Darnila[4]
{primaarinasilalahi@gmail.com}

[1]Magister of Information Technology, Universitas Sumatera Utara, Medan, Indonesia
[2]Department of Computer Science, University of Katolik Santo Thomas, Medan, Indonesia
[3]Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia
[4]Department of Informatics, Universitas Malikussaleh, Aceh Utara, Indonesia

Abstract. Data mining is a process that combines statistics, artificial intelligence, mathematics and machine learning to extract data on a large scale in the database. Data mining is always able to analyze the data so as to find the relevance of data that has a meaning and have a tendency to check large-scale data stored in the database to find a meaningful pattern or rules. The increasing availability of data is often not utilized to provide new knowledge so that large data accumulate is meaningless. The purpose of this research is to extract the information so as to produce knowledge through the decision tree and show the accuracy or influence of Iterative Algorithm Dichotomiser 3 which is used to predict a situation. The classes or attributes in the Iterative Algorithm Dichotomiser are continuously broken into relative categories. Fuzzy Curve Shoulder will be used as a function to form the categories of each attribute value. Using a fuzzy shoulder curve, the dataset is processed using a decision tree that is useful for extracting large amounts of data and searching for hidden links between multiple potential input variables with a target variable. The results of this study are decision trees that will provide predictive data with Iterative Dichotomizer (ID) Algorithm 3.

Keywords: Data mining, Fuzzy Curve Shoulder, Iterative Dichotomizer Algorithm.

1 Introduction

Knowledge Discovery in Database (KDD) is a method to gain knowledge from existing databases. In the database there are tables-tables related to each other. The knowledge obtained in the process can be used as a knowledge base for decision making purposes.

Data mining concerns theories, methodologies, and in particular, computer systems for knowledge extraction or mining from large amounts of data. Data mining is a method to extract the knowledge and information from a large number data such as incomplete, noisy and random [1]. Data mining explorers data from within the database to find hidden patterns, searching for information to predict data. Data Mining techniques are used to examine large databases as a way to discover new and useful patterns.

The decision tree is considered one of the most popular approaches. In the decision tree classification consists of a node that forms the root [2]. Decision tree can describe the

WMA-1 2018, January 20-21, Indonesia

DOI 10.4108/eai.20-1-2018.2281862

relationship between the factors that affect each other a particular problem and find the best solution by taking into account the factor. The existence of a decision tree that is able to analyze the value of risk and value of an information can produce an alternative solution to a problem. Decision trees are also useful for exploring data, finding the hidden relationship between a number of potential input variables with a target variable. Because decision trees combine data exploration and modeling, decision trees are excellent as a first step in the modeling process even when used as the final model of some other technique [3].

The root node in the decision tree is the top node in a decision tree. The decision node is a node to test the attributes of each branch representing the result of the test process on the decision node whereas the leaf node is the last node labeled classification class.

The ID3 algorithm is a recursive procedure, where in at each step there is an evaluation of a subset and creation of decision node, based on a metric called Information Gain, until the subset in evaluation is specified by the same combination of attributes and its values. ID3 algorithm creates a tree by using top-down approach by using the given set of values by checking each attribute at every node. Information gain is used as metric to generate tree to select the best attribute at each step [4], [5].

The fuzzy set theory was introduced by Lofti A.Zadeh, as a mechanism for representing the vagueness and imprecision of the concepts used in natural language[6]. In the analysis of condition these vague concepts equipment type is alse presented. Our reasoning is based on information, it must consider the linguistic form as a variable, whose values can be expressed in terms of natural language. Fuzzy sets were defined as an extension of the classic sets that allows modeling the imprecision of the concepts that manages specialist inductive equiptment, the fundamental change proposed bt Zadeh is to introduce a membership degree(compliance), it is expressing the comformity of an element to a set as a real number in the interval 0 and 1 [7], [8]. In this case, the fuzzy value grouping uses the membership function of the shoulder curve. each class value will be grouped into three values that is good, enough or bad. As time goes by, the availability of data in an agency or system gets bigger. Data that has been processed, often not utilized to provide a new science so that large data is piled up and has no more meaning. From the concept of data mining techniques can be shown that data mining analysis runs on data that tends to continue to enlarge and the best techniques used then oriented to the data is very large to get the conclusion of even the most feasible forecasting.

2 Method

One of the main advantages of decision trees is the ability to generate understandable knowledge structures, i.e., hierarchicaltrees or sets of rules, a low computational cost when the model is being applied to predict or classify new cases, the ability to handle symbolic and numeric input variables, provision of a clear indication of which attributes are most important for prediction or classification [9]. The adoption of the decision tree algorithm for a tree-based prediction model for RSWs with real manufacturing datasets collected from industry [10]. For this research, we use the Decision Tree algorithm to classify data and to extract the rules from the welding dataset. The decision tree resembles a tree structure. Tree is a hierarchical organization of collecting nodes and links, where each node, except the root node, has one incoming link. Each node is a predictive feature and the link represents the value of each conditional variable.Decision rules, decision trees and tests can be considered as a way of

knowledge representation, can be used for feature selection and for construction of classifiers. Based on decision trees and based on tests can construct decision rules [11], [12].

ID3 does not guarantee an optimal solution. It selects the best attribute from the given set. It then splits the dataset in each iteration. ID3 can overfit to the training data. As a solution to this, instead of larger trees smaller trees should be preferred. Though this algorithm specifies a solution, it does not always guarantee an optimal solution. A feature of this algorithm is that it is difficult to apply on continuous set of data. When the values of the attribute are continuous, it is harder to split the dataset into one specific point. Thus, searching for the best value to split becomes a time consuming job [13].

2.1 ID3 algorithm

Iterative Dichotomizes 3 (ID3) performs a thorough search (greedy) on all possible decision trees. The ID3 algorithm tries to build the decision tree top-down, starting with the question: "which attribute should first be checked and placed in root?" This question is answered by evaluating all the existing attributes by using a statistical measure (widely used is information gain) to measure the effectiveness of an attribute in classifying a data set of samples [5], [13].

ID3 method works by determining the weight value of each attribute, then proceed with the selection process of the best alternative from a number of alternatives, in this case the alternatives in question are the proposals that are entitled to follow up on the basis of the criteria specified. The process will continue to be used for the same process (recursive) and will later form a decision tree [9]. If an attribute has become a branch (node) then the attribute is not included in the calculation of the value of the information gain. This process will stop when all data from a branch has been included in the same class or if all attributes have been used but still remain in different classes.

2.2 Fuzzy

The fuzzy set is based on the idea of extending the range of characteristic functions such that the function will include real numbers at intervals [0,1]. The membership value indicates that an item is not only true or false. Value 0 indicates wrong, value 1 indicates true, and there are still values that lie between true and false. If x has a fuzzy membership value $\mu A [x] = 0$, then x is not a member of set A, also if x has a fuzzy membership value $\mu A [x] = 1$ means x becomes a full member in set A.

The term fuzzy logic has various meanings. One of the meanings of fuzzy logic is the expansion of crisp logic, so it can have a value between 0 to 1 [14].

Membership function is a curve that shows the input point into its membership value. To obtain a membership value can use the following function approach: Linear Representation, Representation of Triangle Curve, Representation of trapezoidal Curve, Shoulder Curve Representation, Representation of Bell Curve [15], [16].

2.3 Dataset

Dataset has 1724 records of data with 6 classes, namely Grade Point Semester I (IPSI), Grade Point Semester II (IPSII), Grade Point Semester III (IPSIII), Grade Point Semester IV

(IPSIV), Grade Point Semester V (IPSV) and Grade Point Semester VI (IPSVI). This data source will be used as the material of the analysis.

2.4 Tool

In this research, we apply ID3 algorithm in Rapid Miner open source application because it supports several data mining methods such as preprocessing data, clustering, regression classification, visualization and feature selection.

2.5 Research Framework

In this research data available will be noise data training. Before the dataset is processed with ID3 algorithm, the data is first in transformation. The data that has no information such as data that has no value, will be omitted. After the noising process is complete, the value of each class should be changed to a value dictated. In this case, the data must be categorized into good value, enough and less. To avoid random categorization, and expected to have a fairly accurate result, then categorization is done by processing with Fuzzy Shoulder Curve.Class values that have been grouped will be analyzed by ID3 algorithm to form a decision tree. To make predictions, the test data will be tested using the ID3 decision tree.

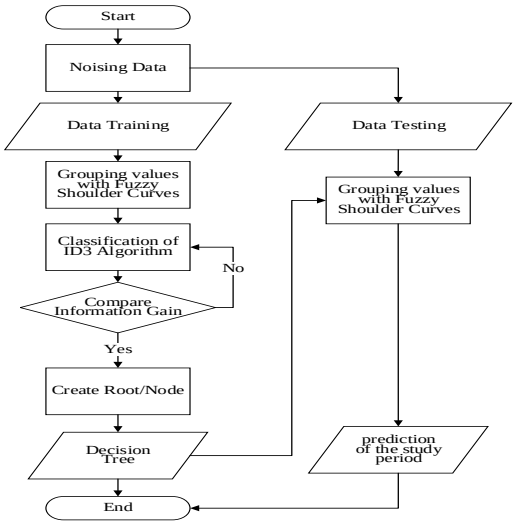

Figure 1. Research Framework

In general, the increasing availability of student data has an impact on data collection. Record student grades can be used as a reference to provide a new science that can be used to predict the accuracy of the study period. ID3 algorithm is machine learning method to classification. Our main goal in this study is to classify the class values in the dataset which is then analyzed with the ID3 Algorithm, get the accuracy of the prediction.

3 Results and Discussion

This research uses the data source of students as much as 1724 sample which will be divided into two parts, that is training data and test data. Student grade point data is used as a criterion by extracting the value of each semester based on the value of the group. The value of each semester class is grouped into good, sufficient and bad, using the fuzzy membership function of the shoulder curve.The area located in the middle of a variable that is represented in triangular form, on the right and left side will go up and down. But sometimes one side of the variable does not change. The fuzzy set of "shoulders", not triangles, is used to terminate a fuzzy region variable. The left shoulder moves from right to wrong, so the right shoulder moves from wrong to right. Representation of membership functions for shoulder curves is as follows:

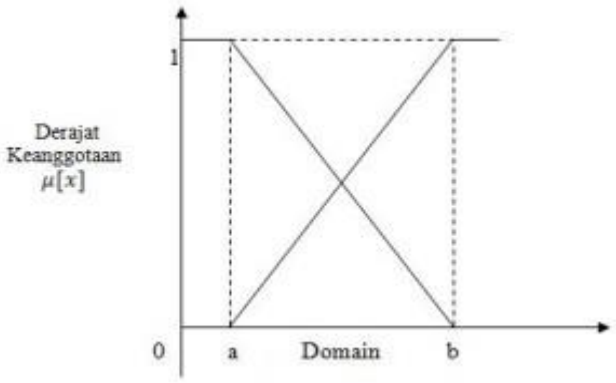

Figure 2. Representation of membership functions

The representation of the shoulder curve formula is:

$$[x,a,b] = \begin{cases} 0; & (x \leq b) \\ \dfrac{(b-x)}{(b-a)}; & a \leq x \leq b \\ 1; & x \geq a \\ 0; & x \leq a \\ \dfrac{(x-a)}{(b-a)}; & a \leq x \leq b \\ 1; & x \geq b \end{cases}$$

With the formula of the fuzzy function, all values of each class / attribute are categorized to be good, sufficient or bad. Training data is processed by using ID3 Algorithm to generate decision tree. Data training will be processed using the ID3 algorithm decision tree to derive predictions of student study classification and method accuracy.

Table 1. Class of Student Dataset

No	Name of Class	Code of Class
1	Grade Point of Semester I	IPSI
2	Grade Point of Semester II	IPSII
3	Grade Point of Semester III	IPSIII
4	Grade Point of Semester IV	IPSIV
5	Grade Point of Semester V	IPSV

6	Grade Point of Semester VI	IPSVI

Calculation Method of ID3 Algorithm can be done with steps as follows:

1. Calculating Overall Entropy Value

   Entropy is a measure of information theory that can know the characteristics of the impurity and homogeneity of the data set.

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

   Where: S: case set,
   n: number of partitions S,
   pi: the proportion of Si against S.

   There are 1724 student data and it is known that the number of data on Graduate on Time (LTW) is 1323 people and the amount of data of Graduated Late (LT) is 401 people. Entropy for a collection of S data samples is

   Entropy (S) = - (1323/1724) log2 (1323/1724) – (401/1724) log2 (401/1724)
   = 0.782514

2. Calculate the entropy and information gain values of each attribute. Can be assumed attribute accuracy of study = "LTW" is a sample (+) which describes pass time, and attribute accuracy of study= "LT" is sample (-) which describes late pass. So the result of calculation node 1.1 as follows:

Table 2. The Result of Calculation Of Node 1.1

GPSI	LTW	LT	TOTAL	Inf.Gain
Good	896	278	921	0.098861
Sufficient	419	117	696	
Bad	8	6	107	
GPSII	LTW	LT	TOTAL	Inf.Gain
Good	769	219	988	0.000696
Sufficient	536	177	713	
Bad	18	5	23	
GPSIII	LTW	LT	TOTAL	Inf.Gain
Good	737	233	970	0.001439
Sufficient	568	158	726	
Bad	18	10	28	
GPSIV	LTW	LT	TOTAL	Inf.Gain
Good	817	255	1072	0.001465
Sufficient	500	146	646	
Bad	6	0	6	
GPSV	LTW	LT	TOTAL	Inf.Gain
Good	690	233	923	0.001868

Sufficient	620	164	784	
Bad	13	4	17	
GPSVI	LTW	LT	TOTAL	Inf.Gain
Good	958	308	1266	
Sufficient	350	89	439	0.001305
Bad	15	4	19	

From the results in Table 2 it can be seen that there is one attribute with the higest gain is GPSI. So the selected aattribute becomes the root node. The result of calculation with the Iterative Dichotomiser 3 algorithm can be indicated by the decision tree as follows :

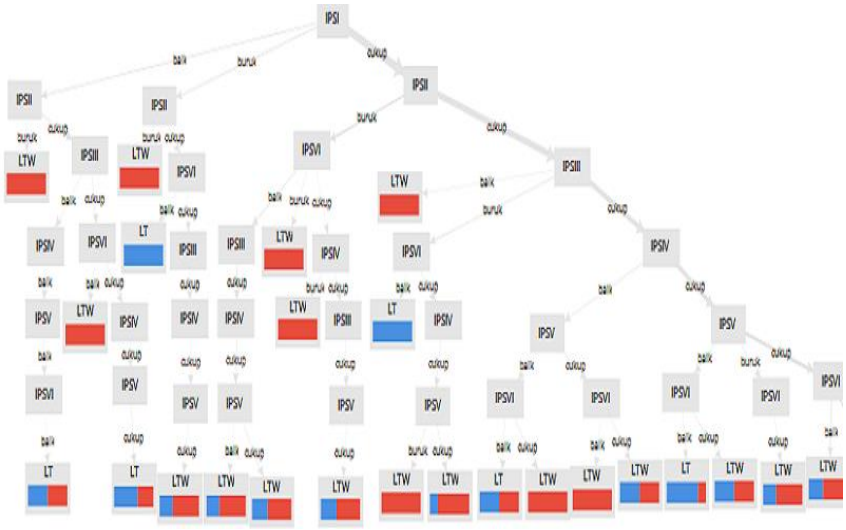

Figure 3. Decision Tree

Table 3. Table Matrix Configuration Testing Sample

	True LTW	True LT	Class Precision
Pred LTW	95	5	95%
Pred LT	0	0	0%
Class Recall	100%	0%	

Accuracy = $\frac{95+0}{100}$ x $100\% = 95\ \%$

The construction tree The ID3 algorithm is tested by inserting the testing data into the construction tree. Test sample size in this test is as much as 100 samples. The configuration matrix table in the test sample is as follows: Prediction results obtained using Iterative Algorithm Dichotomiser 3 (ID3) has an accuracy of 95%.

4 Conclusion

Accumulation of data can be processed into data learning in terms of forecasting. The value of each class on the student dataset can be categorized into either, either, sufficiently or poorly using the fuzzy membership function of the shoulder curve. Iterative Dichotomiser 3 algorithm has discrete attribute value, so used fuzzy curve of shoulder to help categorization not be done randomly. In the calculation using Iterative Dichotomiser 3 (ID3) produces the decision tree. Decision tree used for test data. With the result data and test data using. To forecast data/prediction.

References

[1] Azad, Mohammad. Zielosko, Beata. Moshkov, Mikhail, 2013. Decision rules, trees and tests for tables with many-valued decisions - Comparative study. Procedia Computer Science. **Vol.22**, 87-94

[2] Teli Shahrukh, Prashasti Kanikar, 2015. A Survey on Decision Tree Based Approaches in Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*. **Vol.5**, Issue 4, 613-617

[3] Thi, Thanh Dan, Bi Sihwi, Sari Widya. 2015. Implementasi Iterative Dichotomiser 3 Pada Data Kelulusan Mahasiswa S1 Di Universitas Sebelas Maret. Jurnal ITSMART, **Vol.4**, 84-91

[4] Hartanto David , Hansun Seng., 2014. Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. ULTIMATICS, **Vol. VI**, No. 1, 15-20

[5] Xian Liang, Fuheng Qu , Yong Yang and Hua Cai 2015 An Improved ID3 Decision Tree Algorithm Based on Attribute Weighted *International Conference on Civil, Materials and Environmental Sciences* 613-615

[6] Prof. Mr. A.M Bhadgale, Ms. Sharvari Natu, Ms. Sharvari G. Deshpande, Mr. Anirudha J. Nilegaonkar. 2017. Implementation of Improved ID3 Algorithm Based on Association Function. *International Journal of Pure and Applied Mathematics,* **Vol 114**,No. 10 2017, 1-9

[7] Velasquez Ricardo M Arias, Jennifer V.Mejia Lara. 2017. Principal Components Analysis and Adaptive Decision System Based on Fuzzy Logic for Power Transformer. *Fuzzy Information and Engineering*.**Vol.9**:493-514

[8] Kavakiotis, Ioannis Tsave, Olga Salifoglou, Athanasios Maglaveras, Nicos Vlahavas, Ioannis Chouvarda, Ioanna. 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. **Vol.15**:104-116

[9] Kusumadewi, Idham Guswaludin. 2005.Fuzzy Multi-Criteria Decision Making.*Media Informatika,* **Vol. 3** No. 1, Juni 2005, 25-38.

[10] Mohmad Badr Al Snousy, Hesham Mohamed El-Deeb, Khaled Badran, Ibrahim Ali Al Khlil., 2011. Suite of decision tree-based classification algorithms on cancer gene expression data**.** July 2011, Pages 73-82

[11] Sajida Perveena, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. SDMA Pages 115-121

[12] Fahim Ahmed and Kyoung-Yun Kim 2017Data-driven Weld Nugget Width Prediction with Decision Tree Algorithm *North American Manufacturing Research Conference 4*5. 1009-1019

[13] I. Chikalov, B. Zielosko, Decision rules for decision tables with many-valued decisions, in: J. Yao, S. Ramanna, G. Wang, Z. Suraj (Eds.), RSKT 2011, Vol. 6954 of LNCS, Springer, 2011, pp. 763–768.

[14] Rui-Min, Chai, and Wang Miao. "A more efficient classification scheme for ID3." In Computer Engineering and Technology (ICCET), 2010 2nd International Conference, **Vol. 1**, pp. V1-329. IEEE, 2010.

[15] Yuxun, Liu, and Xie Niuniu. "Improved ID3 algorithm." In Computer Science and Information Technology (ICCSIT), 2010 3rd *IEEE International Conference*, **Vol. 8**, pp. 465- 468.
[16] Putranto, Rizky. Wuryandari, Triastuti dan Sudarno. 2015. Perbandingan Analisis Klasifikasi Antara Decision Tree Dan Support Vector Machine Multiclass Untuk Penentuan Jurusan Pada Siswa Sma. *Jurnal Gaussian*. 4: 1007-1016
[17] Romansyah, F. Sitanggang, I. S. Nurdiati, S.2009. Fuzzy decision tree dengan algoritme ID3 pada data diabetes. Internetworking Indonesia Journal Vol.1:45-52
[18] IzakianH and AbrahamA 2011 Fuzzy C-Means and fuzzy swarm for fuzzy clustering problem Expert Systems With Applications, 38 (3), 1835-1838
[19 Hamasuna Y and Endo Y and Miyamoto S 2009 On tolerant fuzzy c-means clustering and tolerant possibilistic clustering. *Soft Computing*, 14(5), 487–49.