Data Mining utilization in Determining Strategic Business Area Restaurants by Using C4.5 Algorithm

Tryadi Christianto[1*], Deden Abdul Wahab[2]
{ tryadichristianto@mahasiswa.unikom.ac.id[1;] , dedenwahab@unikom.ac.id }

[1]Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Indonesia
[2]Departemen Magister of Management, Universitas Komputer Indonesia, Indonesia

Abstract. The purpose of this study is to help in analyzing the feasibility of business locations quickly. The research method used is qualitative research to describe the decision tree with C4.5 algorithm to easily solve the problem by analyzing the data mining usage. The results of this study are in the form of knowledge as a basis for decision making. In this research, data mining will be carried out so it can produce decisions in determining strategic business locations. It can be concluded that the use of Data Mining in Determining the Strategic Restaurant Business Area by Using C4.5 Algorithm can help in analyzing the feasibility of business locations quickly.

Keywords: Data mining, C.45 Algorithm, Business.

1. Introduction

Currently, the number of few and large businesses that have sprung up have caused competition between each other. One of the things that need to be considered in running a business is seeing opportunities in determining a good business area. This causes many businesses to not running well or suffer losses, due to the lack of knowledge and experience in running a business.

Business plan is a document that is written and contains the business objectives of the sales plan and financial plan [1]. So far, there are still many entrepreneurs that still determine their feasibility of a strategic business location manually which takes more time, especially for a few businesses.

Therefore, data mining methods which is very useful in processing large data will be used to solve existing problems. Data mining can also be said as a class of analytical techniques that go beyond the statistics and aims to conclude data patterns or models [2].

Data mining contains the process of finding trends or patterns in large databases for future decision making [3]. Data mining is used to convert large amounts of data into patterns or rules that has a valuable value which can be useful as a basis for decision making [4].

Data mining has various techniques such as estimation, classification, association, and grouping. Among various types algorithm, classification algorithm plays an important role in predictive analysis. The classification algorithm aims to share objects that only assigned to one of the categories called classes [5].

Classification is defined as a work or training to make a model that can be used to predict unknown data [6]. Classification process of textual documents with sentiment analysis is done

INCITEST 2019, July 18, Bandung, Indonesia

DOI 10.4108/eai.18-7-2019.2287540

by dividing the document types into three categories, namely positive, neutral, and negative sentiment [7].

In classification, there are several methods, namely decision trees, artificial neural networks, rough set theory, fuzzy theory and logic that have their respective functions in the algorithm [8]. Decision tree is a process of classifying unknown data by performing a top-down search strategy for the solution [9].

In the decision tree method, there is a C4.5 algorithm that helps in making decision trees by selecting attributes as root which creates a branch for each value, the process will be repeated for each branch until the case in the branch has the same class [10].

Many studies have been reported by Rosenfeld [11], Qian Y et al. [12], Brook et al [13], and Lou [14]. Although their models have been referenced by many reports, the method they used still has limitations, especially in Utilizing Data Mining in Determining Strategic Business Area Restaurants by Using the C4.5 Algorithm.

Entrepreneurship is a creative effort built on innovation to produce something new which creates work and results, adding value, and provides benefits for others. The purpose of this study is to help in analyzing the feasibility of business area quickly in the form of knowledge as a basis for entrepreneurs in decision making.

2. Method

The research method used in this paper is qualitative research which describes the decision tree with C4.5 algorithm in data mining that can easily solve problems by analyzing the data usage. Decision tree is a prediction model that uses tree structures or hierarchical structures and it is one of many popular classification methods that are easier to be interpreted by people. (Figure 1)

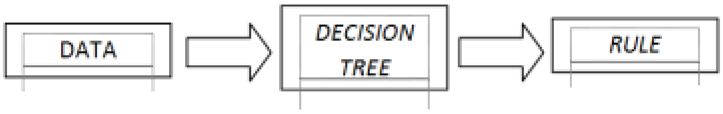

Fig. 1. Decision Tree Concept

Decision tree is also an excellent means of classifying and predicting the outcome by modeling the problems from the processes and provide the solutions.

The decision tree composition consists of several parts called nodes. (Figure 2)

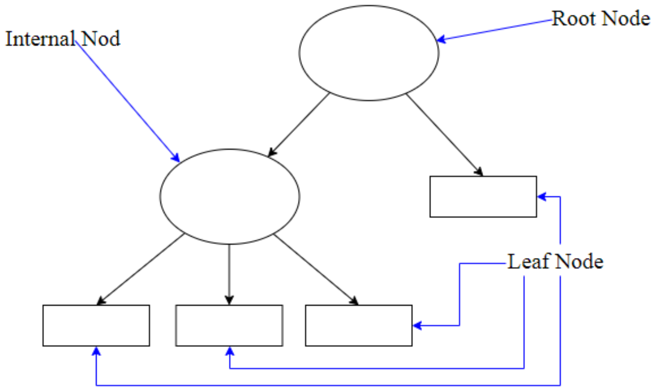

Fig. 2. Decision Tree Composition

Starting from the root node to the leaf node where each branch states the conditions that must be met and at the end of the tree states the data class. The decision tree consists of three parts: root node, internal node, and leaf node. C4.5 algorithm is the right algorithm in viewing strategic business locations. (Figure 3)

$$\boxed{\text{Entropy (S)} = \sum_{j=1}^{k} - Pj\ Log2\ Pj} \tag{1}$$

$$\boxed{\text{Gain (A)} = \text{Entropi (S)} - \sum_{j=1}^{k} \frac{|Si|}{|S|}\ \text{X Entropy(Si)}} \tag{2}$$

Fig. 3. Entropy and Gain formula

The things that need to be known are the profit value and the attributes that become the root node using the Entropy and Gain formula.

3. Results and Discussion

3.1. Algorithm Analysis

The C4.5 is an algorithm for making decision trees by calculating the value of profits [15]. The value of benefits can be obtained by determining the entropy value, whereas the attributes for the root nodes were taken from the highest gain value.

3.2. Criteria for determining strategic business location

The criterias used in determining strategic restaurant locations are:

1. The residents that live in this location
   a. Many
   b. Enough
   c. Few
2. Costs of building or leasing business premises
   a. Expensive
   b. Medium
   c. Cheap
3. Number of competitors
   a. Many
   b. Few
4. Access to the location
   a. Difficult
   b. Easy
5. Target consumers
   a. Many
   b. Few
6. Completeness of business
   a. Complete
   b. Enough
   c. Less

3.3. Algorithm Implementation

We can see through a strategic business location feasibility table to get values that will become a reference in the decision tree. The following are some of the samples obtained in literature: (Tabel 1)

Tabel 1. Strategic Business Location

No	Population	Cost	Competitor	Access to Location	Customers	Completeness of business	Advisability
1	Many	Medium	Few	Easy	Many	Enough	True
2	Enough	Expensive	Many	Difficult	Few	Enough	False
3	Many	Expensive	Many	Easy	Few	Complete	False
4	Few	Medium	Few	Easy	Many	Less	True
5	Enough	Expensive	Many	Difficult	Few	Less	False
6	Many	Cheap	Few	Easy	Many	Enough	True
7	Few	Expensive	Few	Difficult	Few	Less	False
8	Few	Medium	Many	Difficult	Few	Enough	False
9	Enough	Expensive	Few	Easy	Many	Complete	True
10	Few	Expensive	Many	Easy	Few	Enough	False
11	Enough	Expensive	Many	Easy	Few	Enough	False
12	Many	Cheap	Few	Easy	Many	Complete	True
13	Few	Medium	Many	Difficult	Few	Less	False
14	Few	Medium	Few	Easy	Many	Enough	True
15	Enough	Medium	Few	Difficult	Many	Enough	False

Based on table 1, the next step is to calculate node 1 or the root node. The root node calculation aims to determine the highest gain value which become the starting point in making a decision tree. From the results in table 2. It can be seen that the attribute with the highest gain is Customer for 0.7342. thus, Customer is included to the root node. A few attribute values have been classified into one case and the decision is "False" which do not need to do further calculations. (Table 2)

Tabel 2. Node 1 Calculation

Node			**Total Case**	**False (S1)**	**True (S2)**	**Entropy**	**Gain**
1	Total		15	9	6	0.9709	
	Population						0.1460
		Many	4	1	3	0.8112	
		Enough	5	4	1	0.7219	
		Few	6	4	2	0.9182	
	Cost						0.2948
		Expensive	7	1	6	0.5916	
		Medium	6	3	3	1.0000	
		Cheap	2	0	2	0.0000	
	Competitor						0.5382
		Many	7	7	0	0.0000	
		Few	8	2	6	0.8112	
	Access to Location						0.4811
		Difficult	6	6	0	0.0000	
		Easy	9	3	6	0.9182	
	Customers						0.7342
		Many	7	1	6	0.5916	
		Few	8	8	0	0.0000	
	Completeness of Business						0.0619
		Complete	3	1	2	0.9182	
		Enough	8	5	3	0.9544	
		Less	4	3	1	0.8112	

Most of the attribute value still needs to be calculated again because there are still contain "True" and "False". (Figure 3)

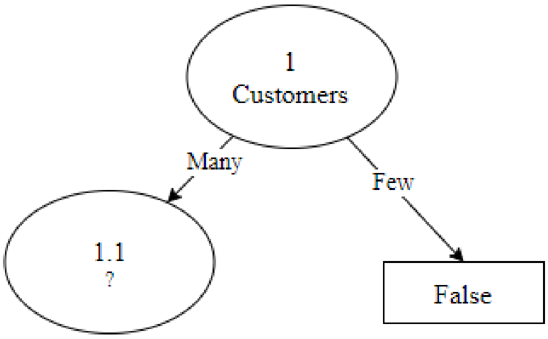

Fig. 3. Temporary Decision Tree

The next step is to calculate node 1.1 or the Internal Node. From the results in table 3. It can be seen that the attribute with the highest gain is Access to location is 0.5916. Thus, customer can be categorized as the internal node 1.1. Difficult attribute value such as few or expensive has been classified into “false” decision and Easy attribute such as many or cheap is classified into “true” decision. So, there is no need to do further calculations. (Table 3)

Tabel 3. Node 1.1 Calculation

Node			**Total Case**	**False (S1)**	**True (S2)**	**Entropy**	**Gain**
1	Costumers	Many	7	1	6	0.5916	
	Population						0.3058
		Many	3	0	3	0.0000	
		Enough	2	1	1	1.0000	
		Few	2	0	2	0.0000	
	Cost						0.1280
		Expensive	1	1	0	0.0000	
		Medium	4	1	3	0.8112	
		Cheap	2	0	2	0.0000	
	Competitor						0.000
		Many	0	0	0	0.0000	
		Few	7	1	6	0.5916	
	Access to Location						0.5916
		Difficult	1	1	0	0.0000	
		Easy	5	0	5	0.0000	
	Completeness of Business						0.1280
		Complete	2	0	2	0.0000	
		Enough	4	1	3	0.8112	
		Less	1	0	1	0.0000	

The next step of making decision tree is to see the second highest gain which is Population.(Figure 4)

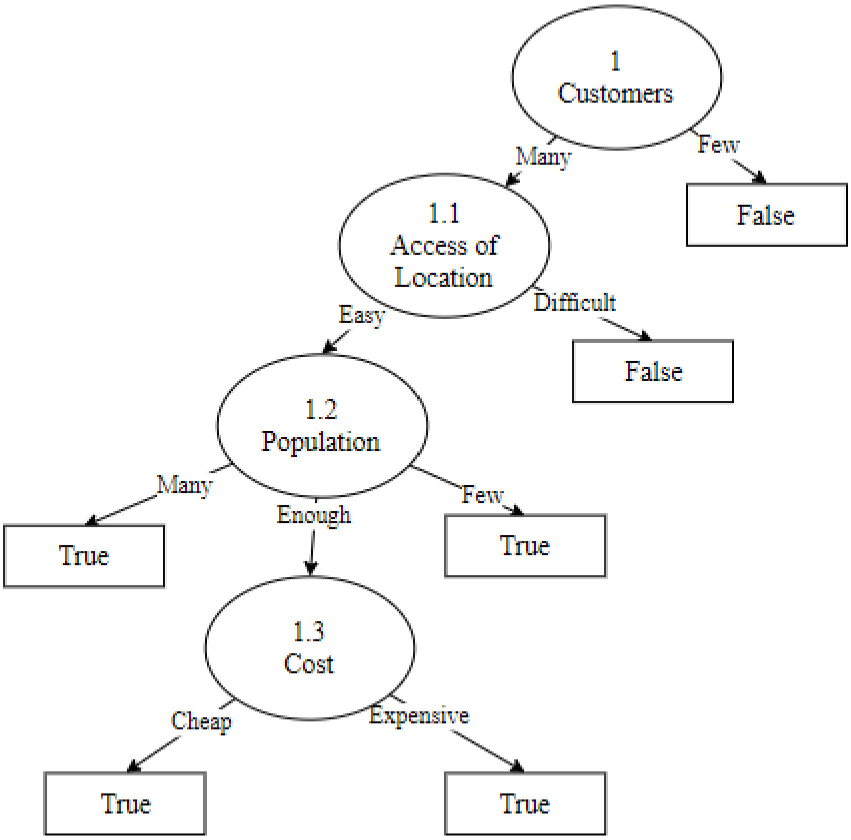

Fig. 4. Decision Tree

Decision models that have been successfully made can be transformed into rules that can be used as a basis for decision making.

4. Conclusion

By utilizing data mining and classification techniques with C4.5 algorithm can help analyze the feasibility of business area quickly, which will produce a model or basic knowledge in decision making that will be useful in determining strategic or less strategic business locations.

References

[1] Soegoto, E. S : Entrepreneurship menjadi pebisnis ulung edisi revisi. Elex Media Komputindo. (2014)

[2] Chouat, O., & Irawan, A. H : Implementation of Data Mining on Online Shop in Indonesia. In IOP Conference Series: Materials Science and Engineering (Vol. 407, No. 1, p. 012013). IOP Publishing. (2018)

[3] Wahyuni, S : Implementation of Data Mining to Analyze Drug Cases Using C4. 5 Decision Tree. In Journal of Physics: Conference Series (Vol. 970, No. 1, p. 012030 IOP Publishing (2018).

[4] Sudrajat, R., Irianingsih, I., & Krisnawan, D : Analysis of data mining classification by comparison of C4. 5 and ID algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 166, No. 1, p. 012031). IOP Publishing (2017).

[5] Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetiyo, B., & Alimah, S : Optimization of C4. 5 algorithm-based particle swarm optimization for breast cancer diagnosis. In Journal of Physics: Conference Series (Vol. 983, No. 1, p. 012063). IOP Publishing. (2018).

[6] Wijaya, E : Implementation Analysis of GLCM and Naive Bayes Methods in Conducting Extractions on Dental Image. In IOP Conference Series: Materials Science and Engineering (Vol. 407, No. 1, p. 012146). IOP Publishing. (2018)

[7] Permana, F. C., Rosmansyah, Y., & Abdullah, A. S : Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media. In Journal of Physics: Conference Series (Vol. 893, No. 1, p. 012051). IOP Publishing. (2017).

[8] Maysanjaya, I. M. D., Pradnyana, I. M. A., & Putrama, I. M. (2018). Classification of breast cancer using Wrapper and Naïve Bayes algorithms. In Journal of Physics: Conference Series (Vol. 1040, No. 1, p. 012017). IOP Publishing. (2017).

[9] As' ad, B. : Prediksi Kehadiran Menggunakan Metode Klasifikasi Naïve Bayes, One-r, Decision Tree. Jurnal Penelitian Komunikasi dan Opini Publik, 20(1). (2016)

[10] Anwar, N., Pranolo, A., Kurnaiwan,R : Grouping the community health center patients based on the disease characteristics using C4.5 decision tree. In IOP Conference Series: Materials Science and Engineering Vol 403, No. 1, p. 012084 (2018).

[11] Rosenfeld, D : Suppression of rain and snow by urban and industrial air pollution. Science, 287(5459), 1793-1796. (2000).

[12] Qian, Y., Gong, D., Fan, J., Leung, L. R., Bennartz, R., Chen, D., & Wang, W : Heavy pollution suppresses light rain in China: Observations and modeling. Journal of Geophysical Research: Atmospheres, 114(D7). (2009)

[13] Brook, R. D., Rajagopalan, S., Pope III, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., ... & Peters, A : Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation, 121(21), 2331-2378. (2010)

[14] Luo K M : Calculation of kinetic parameters from DTA curves using the characteristic temperature Thermochimica acta 255 pp. 241-254 (1995)

[15] Ruggieri, S : Efficient C4. 5 [classification algorithm]. IEEE transactions on knowledge and data engineering, 14(2), 438-444. (2002).