Modification of A priori Algorithm focused on confidence value to association rules

IOP Conf. Series: Materials Science and Engineering **420** (2018) 012125 doi:10.1088/1757-899X/420/1/012125

Modification of A priori Algorithm focused on confidence value to association rules

Dewi Sartika Ginting[1], Herman Mawengkang[2], Syahril Efendi[3]

[1]Student in Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara, Medan, Indonesia
[3]Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

E-mail: [1]dewigintingdg90@gmail.com, [2]hmawengkang@yahoo.com, [3]syahril1@usu.ac.id

Abstract. A priori algorithm is one of the data mining algorithm in formation of rule mining association. A priori algorithm is the process of extraction of information from a database, followed by frequent item / itemset and candidate generation in formation of association rule mining in order to obtain minimum value of support and minimum confidence value. The value of confidence has a big effect on rule the resulting, where the rule generated by the k-item sets pattern needs to be calculated on the level of confidence or certainty of the k-item sets pattern that has complied with the rules. Therefore, this research discusses about a priori algorithm modification which focuses on giving confidence value for each rule generated. Modifications are made by substituting the Bayesian method on a standard A priori confidence formula. And for the next process there is difference of confidence value between a priori standard and modification, where the value of confidence generated a priori modification is bigger, and after calculated for some rules taken according to minimum support condition then there is average difference of confidence value equal to 10,50%.

1. Introduction

Currently, the concept of data mining is increasingly recognized as a tool important in information management because of the increasing amount of information. Data mining itself is often referred to as knowledge discovery in database (KDD) is an activity that includes the collection, use of historical data to find regularities, relationship patterns in large data sets. The output of this data mining can be used for future decision making (Alfina et al, 2012). Data mining is the process of finding patterns or interesting information in selected data by using a particular technique or method. One technique in data mining is market basket analysis that produces asociation rule.

The A priori algorithm was developed in 1994 by Agrawal and Srikan. A priori is an algorithm in performing search frequent itemset by using knowledge of frequent itemset known to process further information. Usually the candidates that may appear are determined first with respect to the minimum support.

Many previous studies on modification of a priori algorithm modify the determination section frequent item set or the determination of a good combination of item sets. And for this occasion the author tries to look on the side of the value of confidence, because the value of

confidence also plays an important role in the a priori algorithm, where the value of confidence will find associative rules that have a greater level of trust than the minimum requirement that has been determined. Based on the above, the writer tries to modify the a priori algorithm from the determination of the value of confidence by substituting a method to get the value of confidence against the rules generated by a priori.

2. Basis Theory

Han and Kamber (2006) briefly data mining can be interpreted as extracting or digging knowledge of large amount of data. Understanding data mining is exploring and analyzing large amounts of data to find patterns and rules that mean Berry and Linoff (2004). Thus, with the growing need for information, more and more areas apply the concept of data mining.

Classification aims to classify data items into one of several standard classes. For example, an email program can classify legitimate emails with spam emails. Some classification algorithms include decision trees, nearest neighbor, naïve bayes, neural networks and support vector machines.

2.1. A priori performance

a. Determine minimum support,
b. Iteration 1: count items from support (transactions that contain all items) by scanning database for 1-itemset, after 1-itemset is obtained, from 1-itemset is above minimum support, if it meets the minimum support, 1- the itemset will become pattern frequent high,
c. Iteration 2: to get 2-itemset, k-itemset before, then scan the database again to count items that load support. itemset meets the minimum support will be chosen as pattern frequent high of candidates,
d. Set the value of k-itemset of support that have met the minimum support of k-itemset,
e. Make the process for the next iteration until no k-itemset meet the minimum support.

2.2. Analysis of high frequency patterns with A priori algorithms

Look for combinations of items that meet requirements minimum of values support in the database. The value of support is items obtained using the following formula:

$$\text{Support (A)} = \frac{\text{Transaksi Mengandung A}}{\text{Total Transaksi}} \quad (1)$$

The value support of 2 items is obtained using the formula:

$$\text{Support (A,B)} = \text{P (A}\cap\text{B)}$$

$$\text{Support (A,B)} = \sum \frac{\text{Transaksi Mengandung A dan B}}{\text{Total Transaksi}} \quad (2)$$

The value confidence of the AUB rule is obtained by the following formula:

$$\text{Confidence} = \text{(B|A)} = \sum \frac{\text{Transaksi Mengandung A dan B}}{\text{Transaksi Mengandung A}} \quad (3)$$

To determine the association rules to be selected must be sorted by Support × Confidence. Rules are drawn as many as n rules that have the greatest results.

2.3. Naive Bayes

Bayes method is a good method in learning machines based on data training, using conditional probabilities as the basis. (Sri Hartati and Sari Iswanti, 2008: 92). The Bayes method for evidence single and single hipotesis H is (Sutojo, et al, 2011: 189):

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \quad (4)$$

Where:

P (H | E) : The probability of hypothesis H occurs when evidence E occurs.
P (E | H) : The probability of the appearance of evidence E, if hypothesis occurs.
P (H) : The probability of hypothesis H regardless of any evidence.
P (E) : Probability evidence E regardless.

3. Research Method

Data used is the transaction data of February 2018 obtained from the super store. Initial data can be as many as 3000 records consisting of 800 transactions and 1397 types of items. The research design to see the percentage of confidence value generated in a priori non-modification and a priori modified by substitution of bayesian probability on calculating confidence value can be seen in figure 1.

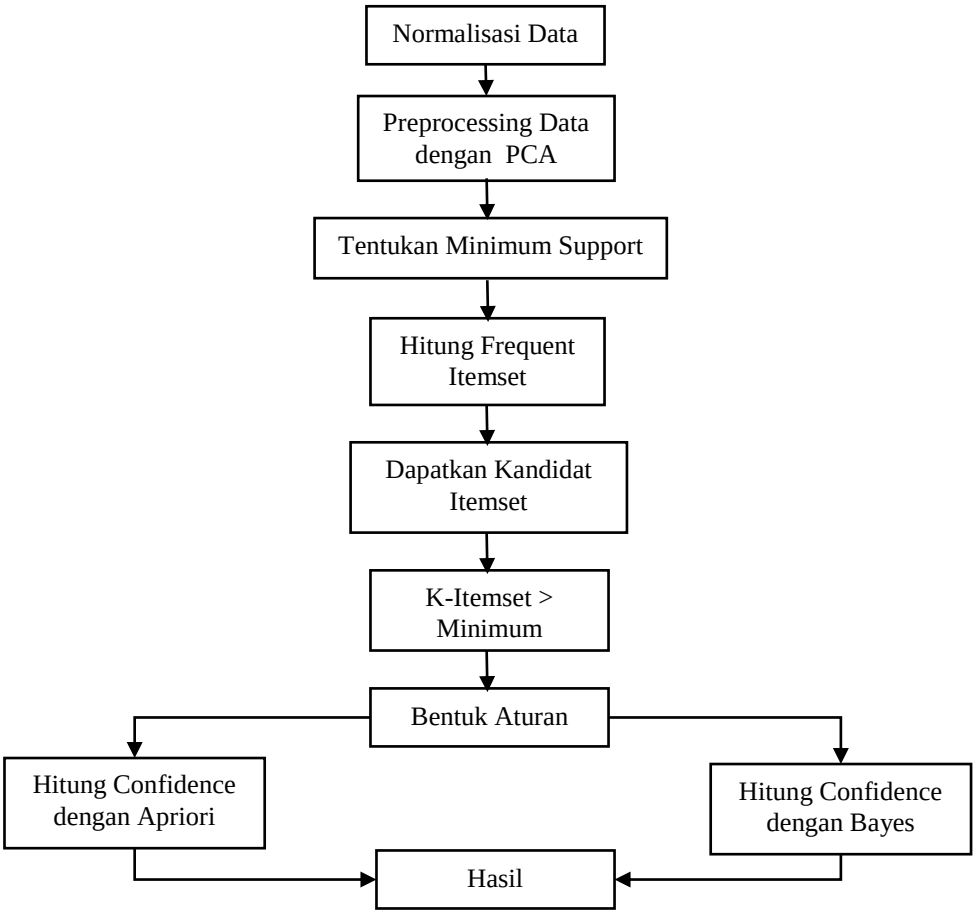

Figure 1. General research design

Figure 1 is an overview of the research design on non-modified a priori algorithm with a priori modification, and modification process using bayesian chance formula in determining its confidence value. The initial process is to collect data related to the data to be researched, then performed the normalization and reduction of items so as to produce data that is ready to use. This research will produce value confidence from non-modified a priori process and from modified a priori process. The final result of this research will be seen percentage level of each algorithm non-modification with which have been modified.

4. Result and Discussion

4.1. Confidence transaction sales using A priori standart.

The process of determining the value of confidence on the rule generated in a priori standard is done by stages provide the minimum value of support that has been determined and form k-itemset pattern, then the rule generated will be calculated value of the level of confidence or certainty of the pattern formed by following the formula of a priori are described chapter 2.

4.2. Confidence transaction sales using modified A priori.

Furthermore, the process of determining the value of confidence on the rule generated in a priori modification, the initial stage is the same as the above process that is done with the stages give the minimum value of support that has been determined and form k-itemset pattern, to form rules appropriate. But the focus on the value of confidence the authors substitute bayesian techniques to obtain the value of confidence by applying a bayesian formula on the determination of confidence as described in the previous chapter.

4.3. Analysis of performance results

After running the association rule using a standard a priori and a modified a priori, there is a considerable difference in confidence value. But for the range of confidence changes from standard to modification is not significant which means that the jump is not fixed. But from the rule that we process the results there is always a difference in the value of certainty or confidence of a priori modified for each rule. The results data are shown in table 1.

Table 1. Results confidence A priori standart & modified

Rule	Standart	Modify	Value of Change
O, P → R	94%	100%	6%
J, BW → BO	94%	100%	6%
E, R, AA → AM	94%	100%	6%
E, P → Q	92%	98%	6%
I, AG → AM	92%	98%	6%
O, Q → R	92%	98%	6%
U, AG → AM	92%	98%	6%
AG, AK → AM	92%	98%	6%
E, P → Q	84%	91,2%	7,20%
I, AI → AM	84%	91,2%	7,20%
G, M → H	84%	91,2%	7,20%
M,O → H	78%	85%	7%
E, P, Q → R	78%	85%	7%
E, R, AM → Q	78%	85%	7%
AB, AJ, AN → AF	78%	85%	7%
BW → J	66%	80,4%	14,4%
D, P → Q	66%	80,4%	14,4%

P, AB → AM	66%	80,4%	14,4%
I, AF, AG → AM	66%	80,4%	14,4%
O, AA. AG → AM	66%	80,4%	14,4%
AB, AF, AK → AA	66%	80,4%	14,4%
AA, AI, AK → AM	66%	80,4%	14,4%
AK, AL, AM → AB	66%	80,4%	14,4%
J, E → BO	60%	73,7%	13,7%
I, AA → AM	60%	73,7%	13,7%
AA, AG → AM	60%	73,7%	13,7%
AG, AN → AF	60%	73,7%	13,7%
AI, AP → AF	60%	73,7%	13,7%
AJ, AR → AF	60%	73,7%	13,7%
E, AG → AM	60%	73,7%	13,7%
Z, AA → AF	60%	73,7%	13,7%
AI, AM, AN → AF	60%	73,7%	13,7%

Table 1 shows clearly the result of the value of confidence or certainty value for some rules taken in accordance with the initial provisions of the minimum support and confidence which is determined from the standard a priori with a priori modification. The table above also presents the range or leap value generated, where in the results above it appears that the confidence of a priori modifications larger than the standard a priori.

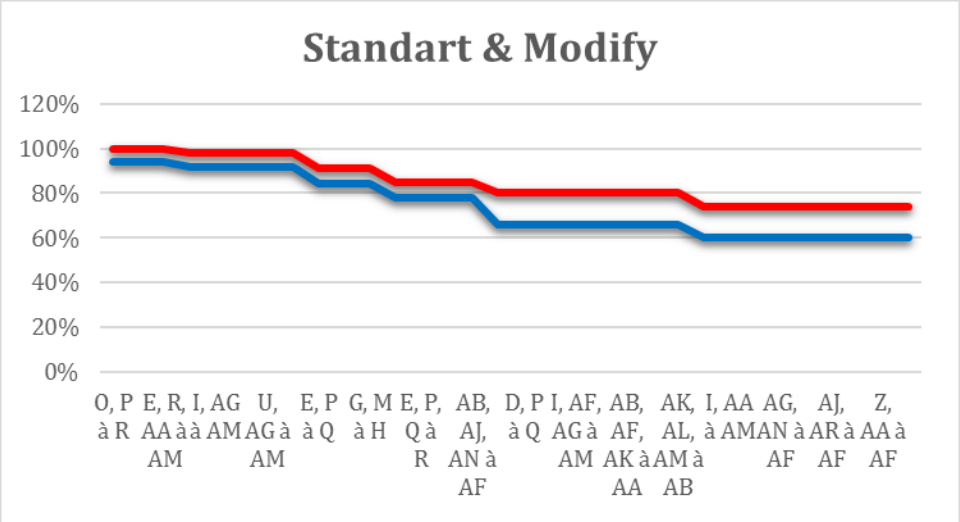

Figure 2. Graph of A priori confidence result standart & modification

From the comparison graph above confidence result, can be seen when rule O, P→ R in if with a priori standard then confidence generated 94%, and when rule the same if with a priori modification then the resulting confidence 100%. In rule this it looks a priori modification larger than 6% of the standard a priori. And overall rule, the value of certainty in the process with a priori standard and a priori modification has an average value of 10.50%. Confidence is increasing. For the range of different confidence values generated each rule of the standard and modification can be seen in figure 2.

5. Conclusion

1. In this study the authors modify from the value of confidence because the value of confidence is very important and a characteristic in a priori, where confidence tells the value of confidence or certainty of the k-itemset pattern that has been formed. Modifications are made by substituting bayesian opportunity work techniques on the formula of determining the value of confidence in the standard a priori.
2. From the data in though by the researcher then in can a priori algorithm standard and a priori modification has an average increase of 10.50. Where the confidence rule generated a priori modification 12.7% larger than the standard a priori.

6. References

[1] USA Erwin (Department of Informatics, Faculty of Computer Science, "Analysis Market Basket with Algorithms," J. Generic, vol 4, pp. 26-30, 2009.
[2] Wahyudi, Adi. "A priori Algorithm Modified with Technique Combination Reduction and Iteration Limitation on Shopping Cart. "Udayana University-2016.
[3] Jianwei Li, Ying Liu, Wei-Keng Liao, Alok Choudhary, Parallel Data Mining Algorithms for Association Rules and Clustering, 2001.
[4] Kusrini and Luthfi, ET, 2009. Algorithms Data Mining. Yogyakarta: Publisher Andi
[5] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New Algorithms for Fast Discovery of Association Rules. Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD'970, Newport Beach, CA), 283-296. AAAI Press, Menlo Park, CA, USA 1997.
[6] Jiawei Han, Michelin Kamber. 2006 "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, Champaign; CA9411.
[7] Wibisono, Y. 2006. Indonesia Language Classification Using the Naive Bayes Classifier. Bandung: FPMIPA Universitas Pendidikan Indonesia.
[8] Mahmudy, FY, Widodo, AW Automatic News Classification Using the Modified Naive Bayes Classifier Method. Tekno-Vol Journal: March 21, 2014.
[9] Kennedy, Hoga, Reza.Boby. Implementation of Data Mining Algorithm A priori On Health Supplies System Inventory. Journal of Information and Scientific Technology (INTI) - Volume: I, Number: 1, October 2013.
[10] Berry, Michael JA and Gordon S. Linoff. 2004. Data mining techniques for marketing, sales, customer relationship management. Second edition. Wiley publishing, Inc.
[11] Kurniasih. Friday, Kumala dewi Nia, Bakri. Analysis and Design of Data Mining With Method Market Basket Analysis For Analysis of Consumer Expenditure Patterns at Tendencies Store. Journal of Information Systems-2012.
[12] Prof. Paresh Tama, And Dr. Yogesh Ghodasara. "Foundation for Frequent Pattern Mining Algorithm", Implementation. International Journal of Computer Trends and Technology (IJCTI) -VOL.4 Issue 7 - July 2013.
[13] Abdi. H., and Williams, LJ, 2010, Principal Component Analysis, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 433-459.
[14] Jolliffe, IT, 2004, Principal Component Analysis second ed., Springer, New York