# Implementation Association Rules with Apriori Algorithm of Student Attendance Record in Islamic University of Indonesia

Rachmad Febrian[1)], Ayundyah Kesumawati[2)]

*Statistics Department, Faculty of Mathematics and Natural Sciences,*
*Islamic University of Indonesia (UII), 55584 Sleman, Yogyakarta, Indonesia*)
*[1]14611246@students.uii.ac.id, [2]ayundyah.k@uii.ac.id*

ARTICLE INFO

*Article history:*
Received March 10, 2017
Revised May 15, 2017
Accepted June 02, 2017

*Keywords:*
Data Mining
Association Rules
Student Attendance Record

ABSTRACT

Data has become an indispendable part of every subject in the world such as in economy, industry, business function, individual and also organization in education such as in university. Data mining software is one of a number of analytical tools for analyzing data. One of techniques in data mining is Association Rules. This paper aim is to known the association between student attendance record and 3 variable which are course, time, and credit semester. Student attendance record need to be analyzed due to see the effectiveness of student attendance record. The result of this paper, association rules is the best method for this case. Courses is not significant for association rules because didn't present in output. There are 9 rules that has lift ration beyond 1.0000000.

## I. Introduction

Data has become an indispendable part of every subject in the world such as in economy, industry, business function, individual and also organization in education such as in university. Recently data mining became one of the solution to extraction of hidden predictive information from a large database. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resource.

In this paper, discussed how to determine the hidden predictive from student attendance report by 3 variable which are course, time and credits in semester. One of techniques in data mining is Association Rules. One of the main techniques in data mining that used to know the relationship between patterns from a dataset is Association Rules [1].

Several organizations have collected massive amounts of such data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information [2]

Discipline is one of factors that important for study in college. Student that underachieving not only ability factor but can also not discipline. Discipline is an attitude and behaviour in obeys all rule in gets behaviour. If concerned by studying therefore studying discipline is an attitude or someone behaviour in obey norm and manner in learned.

Student are required to have high discipline attitude especially discipline in classroom. The discipline learn important applied at lecturing by student. Discipline will add student softskill who will be easily facing the workforce after graduating.

Absence is a system of activities to prove the present or not. Absence is associated with the discipline that determined by the university. The use of the technology in university always use long system (manual). One of implementation of the technology to reach the increases college absence effectiveness by use absence machine fingerprint (fingerprint). Fingerprint absence machine (fingerprint) are absence machine that use someone fingerprint where fingerprint any one not same.

Mechanically fingerprint automatic can not be manipulated and data automatic into the system. In this paper, used the record of student attendance class.

## II. Methodology

### *2.1. Data Mining*

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information – information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is process of finding correlations or pattern among dozens of fields in large relational databases. [3]

Data mining as known by knowledge discovery from data (KDD) while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process are [2] : Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

### *2.2. Association Rules*

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to indentify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

An association rule is an implication expression of the form $X \rightarrow Y$, where $X$ and $Y$ are disjoint itemsets, i.e., $X \cap Y = \emptyset$. The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in $Y$ appear in transactions that contain $X$. The formal definitions of these metrics are [4]

$$\text{Support}, s(X \rightarrow Y) = \frac{\sigma(X \cap Y)}{N} \quad (1)$$

$$Confidence, c(X \rightarrow Y) = \frac{\sigma(X \cap Y)}{\sigma(X)}. \quad (2)$$

Association rule generation is usually split up into two separate steps: First, to find all frequently itemset in a database, minimum support is applied. Second, the rules that used were from these frequent itemsets and the minimum confidence constraint.

While the second step is straight ahead, the first step need more attention. Finds all frequent itemset in a database is difficult because of involving the search for all possible itemset (combination items). The set from itemsset what could be set over power in I and has size *2n - 1* (excluding the empty set is not valid itemset).Although size powerset growing exponentially in the number of items *n* on the *I*, search is made possible by efficient use of property support go down (also called anti monotonisitas) ensuring to frequent itemset, all set pieces also often and thus to itemset rarely, all supersets must also rare.Using the property, algorithm efficient (e.g. Apriori, and Eclat) can find all itemset often [6]

The apriori achieves good performance by reducing the size of candidate sets. However, in situations with very many frequent itemsets, large itemsets, or very low

minimum support, it still suffers from the cost of generating a huge number of candidate sets [5]

```
Procedure Apriori (T, minSupport) { //T is the database and minSupport is the minimum
support
        L1= {frequent items};
        for (k= 2; Lk-1 !=∅; k++) {
                Ck= candidates generated from Lk-1
                //that is cartesian product Lk-1 x Lk-1 and
                //eliminating any k-1 size itemset that is not
                //frequent
                for each transaction t in database do {
                        #increment the count of all candidates in Ck that are contained in t
                        Lk = candidates in Ck with minSupport
                }//end for each
        }//end for
        return ;
```

Fig. 1. Frequent itemset generation of the Apriori Algorithm

And scanning the database repeatedly to check a large set of candidate itemsets. In fact, it is necessary to generate 2100 candidate itemsets to obtain frequent itemsets of size 100.

## III. Numerical Result

This section discusses the result include descriptive statistics, and association rules. The result show as follow:

### *3.1 Descriptive Statistics*

Fig. 2 shows the bar diagram of the Student attendance record in Statistics Islamic University of Indonesia. It be can be seen that the highest student late for join class is in the class that start at 07.00 (first class).

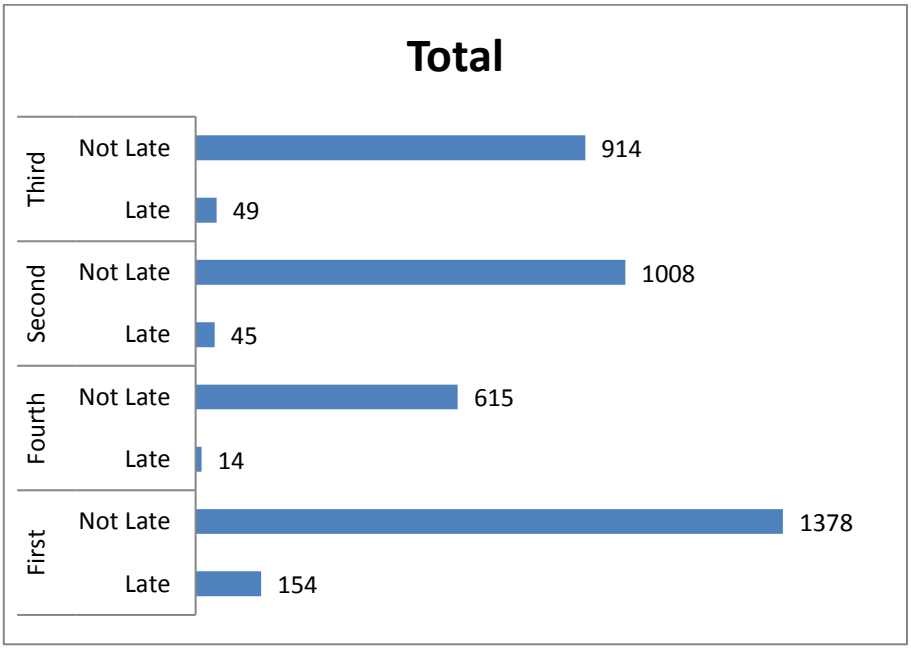


Fig. 2. Total of student attendance record

### *3.2 Association Rules*

In this paper discussed about association rules between four variables which are courses, credits, times of class, and decision (student attendance in courses). The association rules method in this paper using R 3.3.1 to solve this problem and the output R program as follow

Table 1. Result of association rules with apriori algorithm

| lhs | rhs | support | confidence | lift |
|---|---|---|---|---|
| {} | {Decision=Not Late} | 0.9372756 | 0.9372756 | 1.000000 |
| {Time=Four} | {Credits=Two} | 0.1505865 | 1.0000000 | 2.124618 |
| {Time=Four} | {Decision=Not Late} | 0.1472349 | 0.9777424 | 1.043175 |
| {Time=Third} | {Decision=Not Late} | 0.2188173 | 0.9491173 | 1.012634 |
| {Time=Second} | {Decision=Not Late} | 0.2413215 | 0.9572650 | 1.021327 |
| {Time=First} | {Decision=Not Late} | 0.3299018 | 0.8994778 | 0.959673 |
| {Credits=Two} | {Decision=Not Late} | 0.4563084 | 0.9694812 | 1.034361 |
| {Credits=Three} | {Decision=Not Late} | 0.4809672 | 0.9086386 | 0.969447 |
| {Credits=Two,Time=Fourth} | {Decision=Not Late} | 0.1472349 | 0.9777424 | 1.043175 |
| {Time=Fourth,Decision=Not Late} | {Credits=Two} | 0.1472349 | 1.0000000 | 2.124618 |
| {Credits=Three,Time=Third} | {Decision=Not Late} | 0.1378980 | 0.9305331 | 0.992806 |
| {Credits=Three,Time=Two} | {Decision=Not Late} | 0.1728513 | 0.9512516 | 1.014911 |
| {Credits=Two,Time=First} | {Decision=Not Late} | 0.1596840 | 0.9542203 | 1.018079 |
| {Credits=Three,Time=First} | {Decision=Not Late} | 0.1702179 | 0.8535414 | 0.910662 |

Based on table 1., output from R program there are 14 rules, but the significant rules that used in this paper are rules that has a lift ratio beyond 1.00000. The rules for the lift ratio beyond 1 as follow:

1. {Time=Four}⟹{Credits=Two}

Rules with support value = 0.15 confidence = 1.000 and lift ratio = 2.1246. The meaning of support value 0.15 is 15% of the total data in fourth class with confidence 100% it happened in credit two.

2. {Time=Four}⟹{Decision=Not Late}

Rules with support value = 0.147 confidence = 0.977 and lift ratio = 1.043. The meaning of support value 0.147 is 14.7% of the total data in fourth class with confidence 97.7% it happened in not late

3. {Time=Third}⟹{Decision=Not Late}

Rules with support value = 0.218 confidence = 0.949 and lift ratio = 1.012. The meaning of support value 0.218 is 21.8% of the total data in three class with confidence 94.9% it happened in not late

4. {Time=Second}⟹{Decision=Not Late}

Rules with support value = 0.241 confidence = 0.957 and lift ratio = 1.021. The meaning of support value 0.241 is 24.1% of the total data in second class with confidence 95.7% it happened in not late

5. {Credits=Two}⟹{Decision=Not Late}

Rules with support value = 0.456 confidence = 0.969 and lift ratio = 1.034. The meaning of support value 0.456 is 45.6% of the total data in credits two with confidence 96.9% it happened in not late

6. {Credits=Two,Time=Fourth}⟹{Decision=Not Late}

Rules with support value = 0.147 confidence = 0.977 and lift ratio = 1.043. The meaning of support value 0.147 is 14.7% of the total data in credits two and fourth class with confidence 97.7% it happened in not late

7. {Time=Fourth,Decision=Not Late}⟹{Credits=Two}

Rules with support value = 0.147 confidence = 1.000 and lift ratio = 2.124 The meaning of support value 0.147 is 14.7% of the total data in fourth class and not late with confidence 100% it happened in credits two

8. {Credits=Three,Time=Two}⟹{Decision=Not Late}

Rules with support value = 0.172 confidence = 0.951 and lift ratio = 1.014 The meaning of support value 0.172 is 17.2% of the total data in credits two and two class with confidence 95.1% it happened in not late

9. {Credits=Two,Time=First}⟹{Decision=Not Late}

Rules with support value = 0.159 confidence = 0.954 and lift ratio = 1.018 The meaning of support value 0.159 is 15.9% of the total data in credits two and first class with confidence 95.4% it happened in not late.

## IV. Conclusion

In this paper, association rules is the best method for this case. Courses is not significant for association rules because it didn't show in output.

## References

[1] Kantardzid, M., Data Mining : Concept, Models, Method, and Algorithm, John Wiley & Sons, New Jersey, 2003.

[2] Agrawal, R., Imielinski, T., and Swami, A., Mining Association Rules Between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. New Yorka :ACM Press, 1993.

[3] Han, J., Kamber, M., Pei, J., Third Edition Data Mining Concept and Techniques, Elsevier, USA, 2011.

[4] Pang-Ning, T., Steinbach, M., Kumar, and Vipin. Association Analysis Basic Concepts and Algorithm : Introduction to data Mining. Addison-Wesley, 2005.

[5] Wu, X., et all. Top 10 Algorithms in data mining. Springer. London. Pp 13-14. 2007.

[6] Han, J. and Kamber, M. 2006. *"Data Mining Concepts and Techniques Second Edition"*. Morgan Kauffman: San Francisco.