The Comparison of CBA Algorithm and CBS Algorithm for Meteorological Data Classification

Mohammad Iqbal*, Imam Mukhlash*, Hanim Maria Astuti**
* Department of Mathematics, Faculty of Mathematics and Science, Institut Teknologi Sepuluh Nopember
** Department of Information Systems, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember

Keywords:

Meteorological Data
Classification
Association
Sequential pattern
Data mining

ABSTRACT

An increase in the growth of data as a result of the widely use of applications as well as information systems has made data mining an important task in knowledge discovery field of research. Several methods in data mining such as classification has been proposed based on renowned learning methods such as decision tree or neural network. Few studies explored the topic of classification combined with another task in data mining such as association and sequential pattern. Several algorithms that combine classification with other data mining tasks are Classification Based on Associations (CBA) that combines classification with association rules and Classify by Sequence (CBS) algorithm which combines classification with sequential patterns. However, none of studies analyzes the comparison between these two algorithms. In this paper, we explore and compare the performance of CBA algorithm and CBS algorithm in term of accuracy and running time. To do so, we use meteorological data for rain or dry season classification with average temperature, wind speed, relative humidity and air pressure set as our parameters. Based on our experimental evaluation, the CBS algorithm results in high accuracy than of the CBA algorithm. In term of runtime, the CBS algorithm is more efficient than the CBA algorithm.

Corresponding Author:

Iqbal, Mohammad
Departement of Mathematics, Faculty of Mathematics and Science,
Institut Teknologi Sepuluh Nopember,
Jalan Raya Kampus ITS, Gedung U, Sukolilo, Surabaya, Indonesia.
Email: iqbalmohammad.math@gmail.com

1. INTRODUCTION

Data mining is a method to find useful information from large databases. In recent years, various data mining techniques were developed, such as association rules, clustering, classification, sequential pattern, and others. From many data mining problems, classification and prediction are considered as important tasks due to its large applications. Classification is a data mining task intended to learn function or model that maps an item of data into a class from known classes [4]. It is then used to predict a new item data. Several common techniques for classification are decision tree, neural network and others. Other data mining techniques are association rule and sequence pattern. Association rule is a technique used to find intra transactional patterns in databases only for an event while sequence pattern is a data mining technique intended to find series pattern of event at time.

All this time, the development of a single data mining technique has increased due to the widely used of data mining techniques for business, research and other purposes. However, along with the development of this single technique, the development of multiple techniques that integrates several data mining techniques is also rapidly increasing. Some of algorithms based on multiple techniques are classification based association rule (CBA) [1] and featured based sequence classifier [3]. Furthermore, there is also Classify by Sequence (CBS) algorithm that integrates classification and sequence pattern for temporal data [6]. These combination methods can effectively combine the advantages of single mining methods to improve the performance in complex data mining [6].

A study about data mining for weather prediction has been conducted using a single technique such as classification or association rule [1]. However, none of studies explores weather prediction using multiple

data mining techniques. CBA and CBS are included as algorithms based on multiple techniques. These algorithms have similarities, those are: 1) both of them is constructed based on apriori algorithm, and 2) both of the algorithms can process temporal data [6][7][8]. Based on the two aforementioned reasons, a mining process uses either CBA or CBS can be possibly conducted. In addition to that, comparing both algorithms can also be done. In this paper, we use meteorological data to build a classifier using association and sequence pattern for predicting weather. Among two types of meteorological data—a station and various stations—we use station meteorological data. Based on this meteorological data, in this paper, we analyze the performance of CBA algorithm in compared to CBS algorithm in term of accuracy and running time consumption. For this classification, we use two classes on meteorological data: rain season class and dry season class.

2. LITERATURE REVIEW

a. Classification Based on Association (CBA) Algorithm

The basic idea of Classification Based on Association Rule (CBA) technique is generated from association rule that integrated with classification rule to produce subset from effective rules [2]. In recent years, many researchers show that CBA has a high potential to build classification system with more predictive and accurate result than of traditional methods such as decision tree [1]. The concept of CBA algorithm consists of two phases [1].

i. CBA-RG (CBA-Rule Generator)

CBA-RG is called a rule generator that built based on Apriori algorithm to find association rule. CBA-RG is used to find all *ruleitems* that at least similar to the *minimum support* or *minsup*. $\langle condset, y \rangle$ is a *ruleitem* where *condset* is a set of items and $y \in Y$ is a class label. *condsupCount* is number of cases in *D* dataset that consisted of *condset*. *rulesupCount* is a numerous case in *D* that consisted of *condset* and label of classes *y*. Each *ruleitems* represents a rule $condset \rightarrow y$.

$$support = \frac{rule\_sup\_count}{|D|} x100\% \quad (1)$$

and

$$confidence = \frac{rule\_sup\_count}{cond\_sup\_count} x100\% \quad (2)$$

which |*D*| is the dataset size. *Ruleitem* that has support values more than *min_sup* is called *frequent ruleitem* whereas *frequent ruleitems* used to generate *set possibility frequent rule items* is called *candidate ruleitems*. The last step in this procedure is to find the *candidate ruleitem frequent* which has *support* more than *min_sup*.

ii. CBA-CB (CBA-Classifier Builder)

CBA-RB is a classifier builder procedure. To build classifier, set rules evaluate all possible subset training data and choose the correct subset rule sequence with the smallest error. Given two rules r_i and r_j where $r_i \succ r_j$, if *confidence* $r_i \geq$ *confidence* r_j but *support* $r_i \geq$ *support* r_j so that r_i built first than r_j. And *classifier* : $\langle r_1, r_2, ..., r_n, default\ class \rangle$, $r_i \in R$, *R* is set of generated rules $r_a \succ r_b$ if $b \succ a$. If there is no ruleitems in class, it will generate the default class.

b. Classify by Sequence (CBS)

Classify by sequence is integrating sequential pattern mining based on Apriori algorithm and probabilistic induction for classification on temporal data. The advantages are simplicity implementation and the availability of result classification due to the pattern based architecture of CBS. Considering a large database *D* that stores numerous instances, where each instance is comprised of a sequence of temporal data, the following rules are set. Let D_i represents all temporal data instances belonging to class *i*, $D = \{D_1, D_2, ..., D_N\}$. For each class dataset, temporal data instances are represented by $I_j = \langle a_{j1}, a_{j2}, ..., a_{jk} \rangle$, where a_{jk} represents the value of the *k*th time point of instance I_j. The CBS algorithm consists two phases [6]:

a. CSP Miner Algorithm

In this phase, the algorithm discovers all *classifiable sequential pattern* (*CSP*) for temporal classification. To do so, first, we discover all frequent items of the whole dataset as length 1-frequent sequences. We build a root containing these items as leaves. Using this architecture, *CBS* feature mining can discover all possible sequences. As the Apriori-like algorithm, generate all candidate sequences by tree extension. It adds all possible elements to the leaves in order to compose the tree by all frequent sequences of length *k*, and subsequently followed by other candidate sequences of length *k+1*. After candidate tree is built, count the tree by the whole dataset for pruning the non-frequent parts by tree tracing. After counting, remove all sub trees that do not contain any leaves with value more than *min_sup*.

b. Classifier Builder Algorithm

In this phase, build a classifier by removing *CSP*. Before building a classifier, remove the *CSP* that has no classification information. After pruning, the remaining *CSPs* are the features of the class they belong to. Then, the *CSP* can be classified. Because each *CSP* has its own class, it shows that the instances which contain this *CSP* have some possibilities of being classified as a class of *CSP*. Most of the *CSP* contains with one temporal instance belonged to one class, so that this instance should be classified as the class.

3. RESEARCH METHOD

In this paper, we compare the performance between CBA Algorithm and CBS Algorithm in Meteorological data for weather classification. The procedures our research are shown in Fig 1.

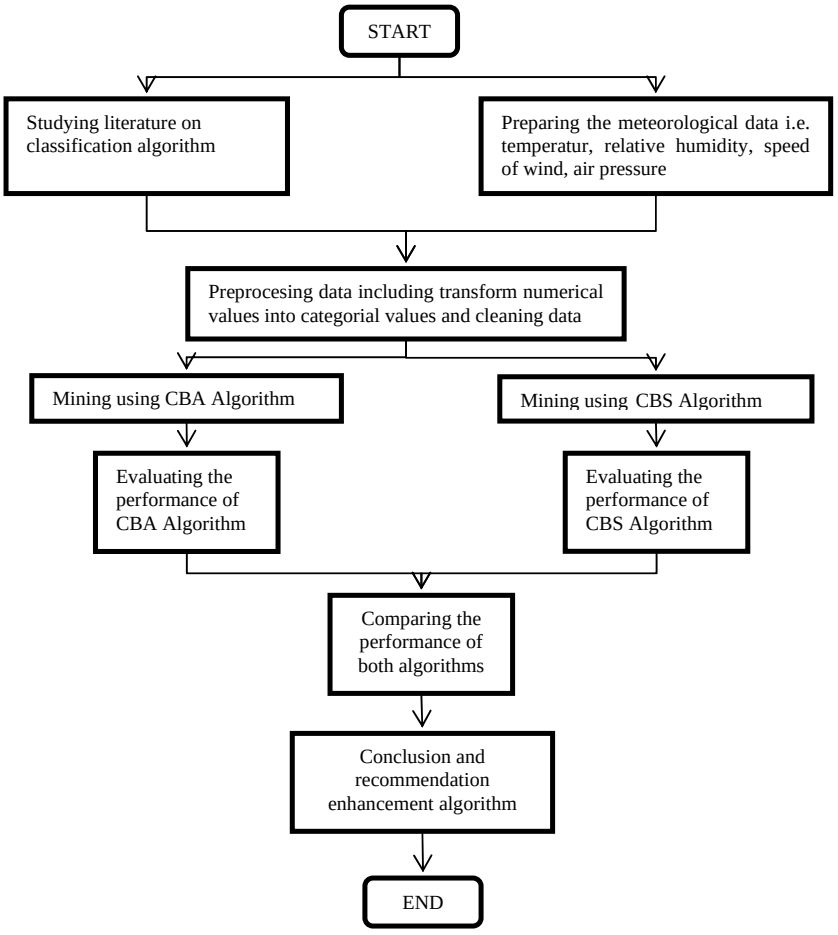

Figure 1. Flow chart of research method

4. RESULTS AND ANALYSIS

In experimental, we use meteorological data with a single station. It consists of 13.515 datasets of climate data at Perak Surabaya, East Java-Indonesia started from 1973 until 2009, obtained from the Bureau

of Meteorology, Climatology and Geophysics of Indonesia (BMKG). All datasets are in numerical values so we need to transform them into categorical values. Since there were some missing values in the attributes, we deleted records containing the missing values and filled with average values taken from other attributes. After this preprocessing stage, the number of datasets decreases into 9.885 datasets.

The attributes of the datasets consist of temperature (average, maximum, and minimum), relative humidity, air pressure and speed of wind (average, maximum, and minimum), dew temperature, rainfall, etc. In this paper, we used 4 attributes as class attributes, i.e. average temperature, relative humidity, average speed of wind and air pressure, and rainfall. We built classes into two classes, "rain season" and "dry season". Like with other general data mining experimentations, we divided datasets into a training part and a testing part. The training datasets are used to build the classifier and the testing data that is intended for accuracy evaluation. To do so, we randomly selected 80% instances of datasets for training data, and the remaining 20% for testing data.

The next stage, the meteorological data is processed using CBA and CBS algorithm. The results obtained from both algorithms are then compared. Both of the algorithms have similarities, as both is constructed based on apriori algorithm. However, the meteorological data used in CBA algorithm is perceived as a sequence to obtain classifier by paying attention to each class label [5]. As for CBS algorithm, daily data obtained from meteorological data is perceived as a sequence or event where rainy or sunny is set as the class label.

The main metric in evaluating *CBA* and *CBS* is accuracy, of which is defined as [6]:

$$\text{accuracy} = \frac{\text{the number of testing instances which are classified correctly}}{\text{the number of testing instances}} \times 100\% \quad (3)$$

Table 1 shows the accuracy of both methods by varying minimum support. The accuracy results of *CBS* are stable when the minimum support is set from 0.1 to 0.4, and slightly decreasing when the minimum support is set from 0.5 to 0.7. As for CBA, its accuracy results increases with minimum support started from 0.3 to 0.7. However, if we compare the results of both algorithms, the accuracy of *CBS* is higher than *CBA* when minimum support is set from 0.1 to 0.6.

Table 1. Comparison accuracy vs minimum support between CBA algorithm and CBS algorithm

Minimum support	Accuracy (%)	
	CBA	*CBS*
0.1	36.97	76.7
0.2	48.36	76.67
0.3	35.6	76.67
0.4	44.26	76.67
0.5	66.67	76.6
0.6	66.67	76.6
0.7	79.5	76.5

The graphic visualization of the above table is presented in the Fig.2 below. The performance of CBA algorithm is slightly increasing while the performance of CBS algorithm is quite stable since the first time the minimum support is set. However, the accuracy of CBS algorithm is better than CBA algorithm Based on this result, therefore, for the case of classification for meteorological datasets, CBS algorithm is more accurate than *CBA* algorithm.

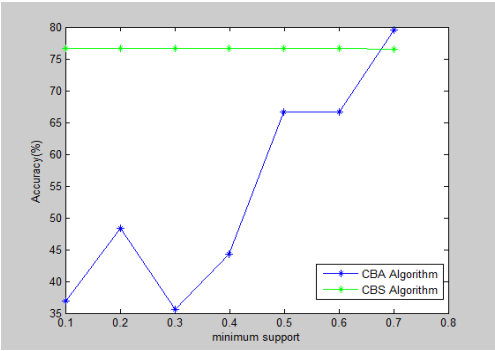

Figure 2. Comparison accuracy vs minimum support between CBA algorithm and CBS algorithm

The second comparison performance parameter is time consumption or runtime program between *CBA* and *CBS*. Table 2 below displays the run time results of both methods by varying minimum support settings. It can be seen that the runtime results of both methods decrease when the minimum support value is started from 0.2 to 0.5.

Table 2. Comparison runtime vs minimum support between CBA algorithm and CBS algorithm

Minimum Support	Runtime (in seconds)	
	CBA	*CBS*
0.2	1739	11.6
0.3	1620	6
0.4	553	4.2
0.5	175	3.1
0.6	372	2.3
0.7	233	1.4

Figure 3 is the graphical representation of Table 2. The figure shows that the runtime results of both methods decrease when the minimum support value is started from 0.2 to 0.5. The result tells us that in overall varying minimum support, the *CBS* runtime is more efficient than *CBA*. Thus, we conclude that the performance *CBS* in running time is better than of the *CBA*.

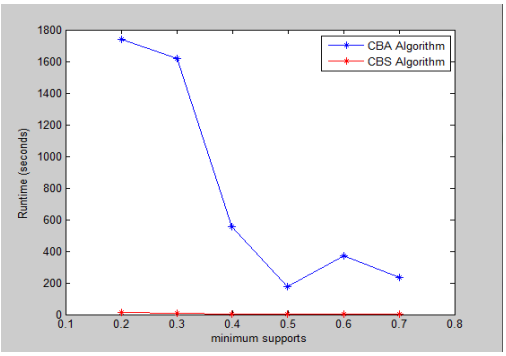

Figure 3. Comparison runtime vs minimum support between CBA algorithm and CBS algorithm

5. CONCLUSION

The research is aimed to compare the performance of multiple algorithms in classification technique, those are, CBA and CBS algorithms, in term of accuracy and running time. Through experimental evaluation, the results show that the running time of *CBS* is faster than of the *CBA*. This means that *CBS* algorithm is more efficient than the *CBA*. The results also show that *CBS* achieves higher and more stable classification results than *CBA* in term of accuracy. In addition to that, the empirical results show that for meteorological datasets, *CBS* algorithm can discover more classifiable features from training data than *CBA* algorithm. When minimum support is set to more than 0.7, no classifier is successfully built.

For future work we will further explore enhancement the algorithm to make pruning part more effectively for produced more *classifiable* rules, with aim to execution efficiency in the process. In addition to that, our research shows that the performance of CBS algorithm is better than of CBA. In our future research, we plan to improve the performance of CBS algorithm because this algorithm indeed has some weaknesses as stated in [6]. One of the rooms for improvement is in the more effective classifiable rules in the pruning process so that the execution time will be much more efficient.

ACKNOWLEDGEMENTS

This paper is part of Penelitian Pendukung Unggulan, a research grant funded by Institute of Research and Public Services, Institut Teknologi Sepuluh Nopember.

REFERENCES

[1] B. Liu, W. Hsu and Y. Ma. "Integrating Classification and Association Rule Mining," In *Proceeding of the 4th International conference on knowledge discovery and data mining* (KDD 1998), New York, USA, 1998., 1998.

[2] M. Nofal, S. Bani-Ahmad. "Classification Based On Association-Rule Mining Techniques: A General Survey and Empirical Comparative Evaluation". Department of Information Technology, Al-Balqa Applied University, Jordan.
[3] N. Lesh, M. J. Zaki, M. Ogiharra. "*Mining features for sequence classification*". In *Proceeding of the 5th ACM SIGKDD International conference on knowledge discovery and data miningSan Diego, California, USA*, 1999, pp. 31-36.
[4] N.P. Tan, M. Steinbach, V. Kumar. *Introduction to Data Mining*, Pearson Addison Weasly., New York , 2006.
[5] S.Nandagopal, S.Karthik, V.P.Arunachalam. "*Mining of Meteorological Data Using Modified Apriori Algorithm*". European Journal of Scientific Research., 2010.
[6] V.S. Tseng, C. H. Lee. "Effective *Temporal* Data Classification by Integrating Sequential Pattern Mining and Probabilistic Induction". *Journal of Expert System with Application*, 2009, pp 9524-9532.
[7] Hong Shu, Xinyan Zhu, Shangping Dai. "*Mining Association Rules in Geographical Spatio-Temporal Data*". International Society for Photogrammetry and Remote Sensing. 2012.
[8] Mennis, J. and Liu, J.W., "*Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change*". Transactions in GIS, 9(1): 5-17. 2005.

BIBLIOGRAPHY OF AUTHORS

Mohammad Iqbal earned B.Sc and a Master Degree in Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He is a lecturer in Mathematics Department, Institut Teknologi Sepuluh Nopember and a member of Computational Mathematics Laboratory in this department. His research interest includes data mining.

Imam Mukhlash is a Lecture in Mathematics Department, Institut Teknologi Sepuluh Nopember. In this department, he is an active member of Computational Mathematics Laboratory. He received his B.Sc in Mathematics from Institut Teknologi Sepuluh Nopember; a Master Degree and a Doctoral Degree both in Informatics and Electrical Engineering from Institut Teknologi Bandung, Indonesia. His research interests include data mining and intelligent systems. In addition to teaching, he has conducted some research projects funded by Institute of Research and Public Services, Institut Teknologi Sepuluh Nopember and National Research Program, the Ministry of Higher Education of Indonesia. He has also published papers both in international journal and international seminars.

Hanim Maria Astuti is currently an active member of the Planning and Development of Information Systems Laboratory in the Department of Information Systems, Institut Teknologi Sepuluh Nopember, Indonesia. She earned B.S. in Information Systems from Institut Teknologi Sepuluh Nopember, Indonesia; and M.Sc in Management of Technology from Delft University of Technology, the Netherlands. She has conducted some research projects granted by National Research Program, The Ministry of Higher Education, Indonesia, and has published conference papers. Her research interests include Mobile Service Innovation and Business Process Management.