International Journal of Engineering & Technology, 7 (4.34) (2018) 394-397

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analyzing Climate Variability in Malaysia Using Association Rule Mining

Rabiatul A. A. Rashid[1], Puteri N. E. Nohuddin[1]*, Zuraini Zainol[2]

[1]*Institute of Visual Informatics, National University of Malaysia, 43600 Bangi, Selangor, Malaysia*
[2]*Department of Computer Science, Faculty of Science and Defence Technology, Universiti Pertahanan Nasional Malaysia, Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia*
**Corresponding author E-mail: puteri.ivi@ukm.edu.my*

Abstract

Previous surveys proved that data mining is one of the methods that can be utilized for climate prediction, predominantly clustering and classification are the most applied methods in data mining to build a model to predict changes in the climate. Unlike the climate change, climate variability is a phenomenon where the occurrence of climate uncertainty is according to the changes year to year basis. This study is focusing to look at the effectiveness of the Association Rule Mining (ARM) techniques in predicting climate variability events in Malaysia. In this report, it explained how the patterns that exist within climate data is discovered using ARM and how the extracted pattern is used to predict climate variability. In this report also, a framework is developed to explain how ARM can generate rules and extract patterns from the data and how the extracted rules and patterns is used to develop a model for predicting climate variability event.

Keywords: *Association rule mining; Climate prediction; Climate variability.*

1. Introduction

Knowledge Discovery in Database consists of a few process and one of the main process is Data Mining [1] where it is a process aim to discover unknown information from the massive amount of data. Data mining techniques are different from standard statistical method, where data mining can be programmed to find hidden knowledge without relying on the previous information of the data, however, the unknown knowledge exists in the data were obtained using data mining task [2].

Data mining consists a few techniques that can be applied individually or combined for performing sophisticated processes. Classification is the systematic approach of the classification model based on a set of input data [3]. This technique is used in classifying the data from the database to some classes based on certain criteria [4]. The algorithms that are commonly used in this method are neural networks and decision tree [5]. Clusterization is a technique in which the data are divided into groups of similar objects. Each cluster constituted group data based on certain features [6]. The algorithms in clusterization techniques are K-means and hierarchical clustering [7]. ARM is often used to find patterns that exist in the itemset. It aims to extract the interesting relationship, frequent patterns and coloration existing in the set of items in the data repository [8]. ARM has been applied in many real world problems such as finding patterns in documents [9-10], predicting floods [11], trend analysis of social networks [12], monitoring elderly people [13], etc.

Various research has been done to see the suitability of data mining in predicting the climate. The most commonly used techniques are ARM, classification and also clustering. Therefore, these methods are commonly used in research in getting the most appropriate and accurate techniques for modeling climate forecasting. This is because climate forecasting is known as a complex analysis due to various elements in weather and climate such as rainfall, wind speed, temperature, humidity and etc. [14].

ARM is capable to identify the relations exist within the datasets. ARM has several algorithms and the popular algorithms which is often used are Apriori and FP-Growth [15-16]. Many researchers have used ARM in developing a model for climate prediction. Using previous data, model for climate prediction were developed using ARM [17-18]. In classification and clusterization, the data will be divided into specific groups. However, it is different with ARM, where patterns and rules are obtained from relationship discovered within the database [19]. This method is considered as useful for climate prediction due to the fact that climate data consist of various elements and factors.

2. Methodology

2.1. Association Rule Mining

In ARM, it is known that the rules are measured by the value of support and confidence. In the datasets of A →B, where A and B are item sets, the value of support is measured as:

Support, Supp (A) = A / T

Supp (A) shows how many times that the item occurs in total of the transactions (T).

The confidence is value of percentage in T that contains A that also contains B. The conficence value can be calculated as:

Confidence (A →B) = Supp (A ∪ B) / Supp (A)

Therefore, ARM will extract rules-based by:

- Support ≥ minsupp threshold
- Confidence ≥ minconf threshold

where minsup and minconf are the corresponding support and confidence thresholds.

The lift value of the rules is also analyzed in this study and it is one of the parameters to in ARM that can be used in the analysis. The value of the lift is calculated as below.

Lift = Confidence/Expected Confidence

Therefore, ARM will provide the information about the probability of generated rules based by the lift value.

2.2. Climate Variability Prediction Framework (CVPF)

The CVPF is developed to extract rules from climate data variables and the rules is clusters according to their characteristic. The framework (see Figure 1), is consists of Stage (i): Data Processing, Stage (ii): Rule Analysis and Stage (iii): Prediction Model. The process of data cleaning and data normalization will be done in Stage (i). The inaccurate data will be identified in the data cleaning process and the data will be either replaced, or modified, or deleted from the dataset. Then, in normalization the set will be organized based by the columns and tables to reduce the data redundancy and therefore will improve the integrity of the data. The results from stage (i) will be applied in the normalization process.

When handling with time series data, the values for the attribute will be divided into a number of sub-ranges in the normalization stage. Usually a low value of range will be used as it will have the effect of producing fewer columns in the output data which in turn provide computation efficiency benefits. The larger number of sub-range value, the more output attributes will be generated during the normalization process and this will affect the efficiency of the ARM algorithm used in the analysis stage.

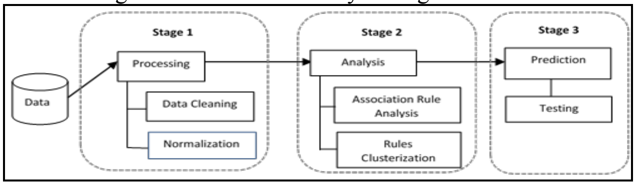

Fig. 1: Climate Variability Prediction Framework (CVPF)

In stage (ii), the data will be analyzed using ARM and meaningful rules from the data sets will be extracted by using the FP-growth algorithm. This algorithm is known to be one of the popular and fastest association rule algorithms because compact data structure is used in the algorithm and eliminate the need of repeated database scan [20]. Due to that, the information within the data set produced by FP-Growth is greatly compressed and FP-Growth algorithm is used in this study to extract rules and patterns for the prediction model.

The relevant rules and patterns will be identified, and then the selected rules will be grouped into different clusters in the rules-based clustering process. All significant rules are clustered according to the rules features and similarity. In the last stage, prediction model based by the results from stage (ii) will be built. Then, the model will be tested using the previous climate data and the results from the model will be evaluated to measure the accuracy of the developed model.

3. Results and Discussion

Previous weather data of Petaling Jaya, Selangor is used in this study and the data were collected from the Institute of Climate Change, The National University of Malaysia. The monthly data sets for year 2014 and 2015 consist of humidity, temperature, wind speed, rainfall and number of rain days.

During the experiment, the generated rules were observed. Several support and confidence value is used in the experiment and the support value of 15% and confidence value of 70% show more significant and meaningful patterns. The details of the ARM experiment based on the number of lift and confidence rules produced are shown in Table 1.

Table 1: Results from ARM Experiments

Year	Support Value	Confidence Value	Lift Rules Generated	Confidence Rules Generated
2013	15	70	188	94
2014	15	70	288	160
2015	15	70	88	22

Table 1 shows that ARM extracted more rules from the data in 2013 and 2014. In both years, the rules extracted shows more information on the relationship between variables and produced more significant patterns to be used in the prediction model. Lift value shows the strength of the rules and indicates that lift with higher value is more reliable for prediction.

The results show rules and patterns extracted from the dataset and the significant rules show detailed information that is related to climate that can be used to predict season changes in Malaysia. In Table 2, the rules generated shows features in the itemset to predict rainfalls and rainy days.

Table 2: Extracted Rules for Rainning Season

Year	Generated Rules
2013	{rainday>=21.6} -> {rainfall<527.80, humidity<82.3, temperature<27.62 windspeed<1.06} 100.0
	{rainday>=21.6} -> {humidity<82.3, temperature<27.62} 100.0
	{rainday>=21.6} -> {humidity<82.3, windspeed<1.06} 100.0
2014	{humidity<82.5 temperature<27.46, windspeed<0.9400000000000001 rainfall<624.0} -> {rainday>=21.4} 100.0
	{humidity<82.5 temperature<27.46 windspeed<0.9400000000000001} -> {rainday>=21.4} 100.0
	humidity<82.5, windspeed<0.9400000000000001, rainfall<624.0} -> {rainday>=21.4} 100.0
2015	{humidity<74.61999999999999, temperature<28.64} -> {17.0<=rainday<20.0} 100.0
	{humidity<74.61999999999999, windspeed<1.2 rainfall<437.96} -> {17.0<=rainday<20.0} 100.0
	{humidity<74.61999999999999} -> {17.0<=rainday<20.0} 75.0

In Table 4, significant rules that show details features of rainy days were chosen and from the rules it shows that:

- In 2013, rainy days are >=21.6 days, and the related features during the rainy days are when the temperature is <27.62℃, the wind speed measure is <1.06m/s, the humitdity is <82.3% and the total amout of rainfall is < 527.80mm.
- In 2014, the total of rainy days based by the rules generated >=21.4 days. During the raining season the related variables recorded are temperature<27.46℃, the wind is <0.94m/s, the humidity is <82.5% and the amount of rainfall is <624.00mm.
- In 2015, the rules show that rain days happen between 17 to 20 days, and the related variables during the raining days are temperature<28.64℃, the percentage of humidity is less than 74.619%, the wind speed is less than 1.2m/s and amount rainfall is less than 437.96mm".

From the results, rules and pattern with similar features are identifed and used to predict an event that caused by climate variability. In Figure 2, the rules generated shows similarity in the variable's value that can be used to predict raining season. For all three years, the patterns show that period of rainy days are between 19-21 days, and in Figure 2 it shows the similarity of the variables. From the pattern, we can conclude that Petaling Jaya is expected to receive a huge amount of rainfall measured from 440mm to

620mm during raining season. The high amount of rainfall during this period is could bring risk to the public especially the of flash floods. Therefore, prediction of the long rainy period can be used by the authorities to predict events such as flash flood and landslides. They can take precautionary actions such as ensuring the river water level is in a safe state and to maintain a good drainage system in Petaling Jaya.

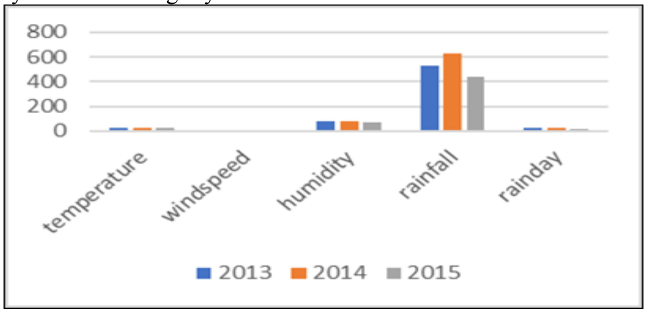

Fig. 2: Pattern of Raining Season in Petaling Jaya

Table 3: Extracted Rules for Dry Season

Year	Generated Rules
2013	{humidity<73.47999999999999} -> {temperature<28.66 rainfall<222.64} 100.0
	{humidity<73.47999999999999 temperature<28.66} -> {rainday<11.4} 100.0
	{windspeed<1.3} -> {rainday<11.4} 100.0
2014	{rainday<7.6} -> {humidity<68.34 rainfall<158.72} 100.0
	{temperature<30.1} -> {humidity<68.34 rainfall<158.72} 100.0
	{temperature<30.1} -> {humidity<68.34} 100.0
2015	{temperature<29.02 windspeed<1.2} -> {17.0<=rainday<20.0} 100.0
	{humidity<71.78} -> {temperature<29.02} 100.0
	{17.0<=rainday<20.0} -> {rainfall<356.91999999999996} 100.0

Table 3 shows the generated rules that stated the features that indicates the dry season. The selected rules show that:

- In 2014, when the period for rainy days <11.4 days, it is considered as dry season and during this period the temperature is <28.66℃, the windspeed measured <1.3m/s, humidity is less <73.5% and rainfall received is <222.64mm.
- For 2014, the rules show that when rainy days happen >=7.6days, the related variables during this period is the temperature is <30.1℃, the humidity is <68.34% and the amount of rainfall is <158.72mm.
- Meanwhile, for the year 2015, the rules show the range of rainy days between 17 to 20 days, and the related variable value are temperature is <29.02℃, the humidity is <71.78%, the windspeed is <1.2m/s and the amount of rainfall is <356.91mm.

Meanwhile, in Figure 3, the resulting rules show the decreasing amount of rainfall received during the period from 350mm to 160mm of rain only. This pattern indicates that the shorter rainy day period can reduce the amount of raindall in the area. During this period, the shortage of rainfall will affect the water supply to the public and therefor, this information can be used to predict the dry season.

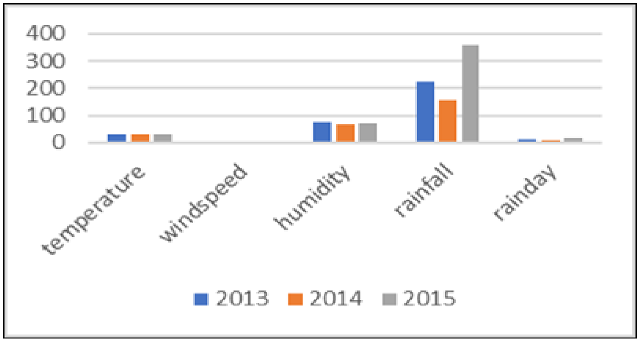

Fig. 3: Pattern of Dry Season in Petaling Jaya

From the analysis conducted, huge number of rules and patterns were extracted from the dataset using ARM. In this experiment, rules and patterns were generated using FP-Growth, and the results were analyzed to identify significant patterns and rules. Based on the analysis, the prediction of the seasons is based on the climate features indicated by the rules. The rules show significant features that can be used to identify types of climate and in Table 4 is the summary of the climates features that has been identified in this study.

Table 4: Summary of Climate Features

Raining Season	Temperature <28℃ + Raindays > 20 days + Rainfall >400mm / month
Dry Season	Temperature > 29.00℃ + Raindays < 20 days + Rainfall <200mm / month

4. Conclusion

The aim of this study is to prove that ARM can extract significant rules and patterns within the climate data. This study also shows that the patterns and rules produced using ARM can be applied to construct a prediction model. The significant patterns and the rules of the association is measured by the high confidence and lift value. From the analysis, it shows that the rain and dry season can be determined based on the rules produced by ARM. However, there is a limit during this study because the data period is based on monthly data. More detailed analysis results can be obtained if more detail climate data such as daily data is used in the future work.

Going onward, this written report will concentrate on how to apply the clustering method based on the rules and pattern produced by ARM. In the clustering method, each cluster will be based on the rules' characteristic and behavior. The prediction model will be built based on the outcome of the association rule-based clustering method. The prediction model will be tested with previous data to measure the ability and accuracy of the model in predicting climate.

Acknowledgement

Authors would like to thank Institute of Climate Change, The National University of Malaysia for providing the climate data to be used in this study.

References

[1] Ramamohan Y, Vasantharao K, Chakravarti CK, Ratnam ASK. A study of data mining tools in knowledge discovery process. International Journal of Soft Computing and Engineering, 2012, 2(3): 191-194.

[2] Olaiya F, Adeyemo AB. Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 2012, 4(1): 51-59.

[3] Tarmizi ND, Jamaluddin F, Bakar AA, Othman ZA, Hamdan AR. Classification of dengue outbreak using data mining models. Research Notes in Information Science, 2013, 12: 71-75

[4] Ngai EWT, Xiu L, Chau DCK. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 2009, 36(2): 2592-2602.

[5] Ozer P. Data mining algorithms for classification. Bachelor thesis, Radboud University Nijmegen, 2008.

[6] Berkhin P. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), Grouping Multidimensional Data. Berlin: Springer, 2006, pp. 25-71.

[7] Amin MNM, Nohuddin PNE, Zainol Z. Trend cluster analysis using self organizing maps. Proceedings of the 4th World Congress on Information and Communication Technologies, 2014, pp. 80–84.

[8] Karthikeyan NRT, Ravikumar N. A survey on association rule mining. International Journal of Advanced Research in Computer and Communication Engineering, 2013, 3(4): 318-321.

[9] Zainol Z, Nohuddin PNE, Jaymes MTH, Marzukhi S. Discovering "interesting" keyword patterns in hadith chapter documents. Proceedings of the Interntional Conference on Information and Communication Technology, 2016, pp. 104–108.

[10] Chua S, Nohuddin PN. Frequent pattern extraction in the Tafseer of Al-Quran. Proceedings of the IEEE 5th International Conference on Information and Communication Technology for the Muslim World, 2014, pp. 1-5.

[11] Harun NA, Makhtar M, Aziz AA, Zakaria ZA, Abdullah FS, Jusoh JA. The application of Apriori algorithm in predicting flood areas. International Journal on Advanced Science, Engineering and Information Technology, 2017, 7(3): 763-769.

[12] Nohuddin PNE, Coenen F, Christley R, Setzkorn C, Patel Y, Williams S. Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps. Knowledge-Based Systems, 2012, 29: 104-113.

[13] Azam MA, Loo J, Naeem U, Khan SKA, Lasebae A, Gemikonakli O. A framework to recognise daily life activities with wireless proximity and object usage data. Proceedings of the 23rd IEEE International Symposium on Personal, Indoor and Mobile Radio Communication, 2012, pp. 2553–2558.

[14] Joshi A, Kamble B, Joshi V, Kajale K, Dhange N. Weather forecasting and climate changing using data mining application. International Journal of Advanced Research in Computer and Communication Engineering, 2015, 4(3): 19-21

[15] Liu X, Zhai K, Pedrycz W. An improved association rules mining method. Expert Systems with Applications, 2012, 39(1): 1362-1374.

[16] Kumbhare TA, Chobe S V. An overview of association rule mining algorithms. International Journal of Computer Science and Information Technologies, 2014, 5(1): 927–930.

[17] Gouda KC, Chandrika M. Data mining for weather and climate studies. International Journal of Engineering Trends and Technology, 2016, 32(1): 29-32.

[18] Rashid RAA, Nohuddin PNE, Zainol Z. Association rule mining using time series data for Malaysia climate variability prediction. Proceedings of the 5th International Visual Informatics Conference, 2017, pp. 120-130.

[19] Rana DP, Mistry NJ, Raghuwanshi MM. Novel usage of gujarati calendar in temporal association rule mining for temperature analysis of Surat, India. Proceedings of the International Conference on Soft Computing and Machine Intelligence, 2014, pp. 38–41.

[20] Nohuddin PNE, Zainol Z, Lee ASH, Nordin AI, Yusoff Z. A case study in knowledge acquisition for logistic cargo distribution data mining framework. International Journal Advanced Applied Sciences, 2018, 5(1): 8-14.