Special issue of the International Conference on Computational and Experimental Science and Engineering (ICCESEN 2014)

Sequential Pattern Discovery Algorithm for Malaysia Rainfall Prediction

A.M. Ahmed[a], A.A. Bakar[a,*], A.R. Hamdan[a], S.M. Syed Abdullah[b] and O. Jaafar[b]

[a]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology

[b]Institute of Climate Change, University Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia

This study proposes a sequential pattern mining algorithm to discover sequential patterns of Malaysia rainfall data for prediction. The apriori based algorithm is employed to find the sequential patterns from the time series data. The frequent episodes of rainfall sequences are discovered and classified by the expert into four main events namely, No rain, Light, Moderate and heavy. The sequential rules of ten rainfall stations from the duration of 33 years are analysed. The proposed algorithm is able to generate higher confidence and support of frequent and sequential patterns. Generally, the proposed study has shown its potential in producing methods that manage to preserve important knowledge and thus reduce information loss in weather prediction problem.

DOI: 10.12693/APhysPolA.128.B-324

PACS: 92.40.Zg, 92.60.Wc

1. Introduction

The dynamic nature of the weather has led to many researches relating to weather predictions such as weather forecasting warning system [1], rainfall and flood forecasting [2–3]. The condition of the weather gives impact to many sectors, for example in the power usage of industry, residential facilities, agriculture [4–5]. The continuous change in climate highlights the need of weather prediction systems that is up to par with the current technology existed [6], including the Malaysian weather prediction system. Rainfall prediction is one of the areas being actively researched all over the world [7–9]. The rainfall data in Malaysia have been used previously in researches such as gauging the size of the rainfall cells [10–11]. Sequential pattern mining is one of data mining techniques that potentially give new insights to the weather prediction problems. Previously it has been used in other application such as alarm log analysis, financial events, and stock trend relationship analysis [11]. Several works by Katoh et al. [12–13] propose algorithms that find frequent episodes from the input sequence. In this paper we employ a sequential pattern discovery algorithm for prediction of Malaysia rainfall dataset. This study focuses on rainfall data collected from Institute of Climate Change University Kebangsaan Malaysia (UKM).

2. Material and methods

The rainfall data sequences are classified by experts into four main events, namely no rain (N), light (L), moderate (M) and heavy (H). Let $C = [e_1, \ldots e_m]$ $(m \geq 2)$ be a finite alphabet with total order that denotes the rainfall events over N. Each element $e \in C$ is called an event. $S = s_1, \ldots s_w$, $(w \geq 2)$ is a serial of events where each s is numbers of events (e), for example (L $\rightarrow$ M $\rightarrow$ H) where $e_1 = L$, $e_2 = M$, $e_3 = H$ and the width of this episode is 3. Table I shows an example of serial episodes in the rainfall data and the number of observations.

TABLE I

Example of episodes and number of observations.

Episodes/Events	Events	#Observations
LLM	LLM	3
MMH	MMH	2
LLM	NNL	1
NNL	NML	1
LLM	MHH	2
MHH		
NML		
MHH		

The frequent pattern algorithm is employed to find the most repeated episodes among those that were generated in the Allen operation process [13–14].

Algorithm 1:
Episodes (frequent sequence) generation

Input: rainfall data sequences Output: frequent episodes
Step 1: read the rainfall data sequences
Step 2: Generate episodes for every 3 events (use Allen interval concept)
Step 3: read each episodes and count the frequency
Step 4: calculate the confidence of the episodes (conf_e)
Step 5: compare with the min_conf = 0.1
Step 5: If the conf_e $\geq$ min_conf then input to frequent episode list
Step 6: generate frequent episodes

*corresponding author; e-mail: azuraliza@ukm.edu.my

The output of the algorithm is the most frequent patterns, i.e., those that satisfy the minimum support predefined by the user. The frequent algorithm is shown as algorithm 1. The frequent episode is counted based on number of occurrence and it is evaluated based on its confidence level (minimum confidence is set to 0.1). The numbers of frequent episodes are produced and adapted as new patterns for rainfall data sets.

3. Results

This section presents and discusses the results of the application of the frequent patterns algorithm. The rainfall produced *23228* events from different 10 rainfall stations. Table II shows an example of frequent serial episodes in the month of January for the 10 rainfall stations. High confidence was found in extracted episodes 'LML' which occurs 101 times in January, which means that when L comes before M and L comes after L with confidence 0.65 this indicates new rules for that specific month (January) and the given data sets. Another frequent set of episodes that can be seen is 'LLH', with confidence 0.64.

Table III presents the normal and non-normal patterns of rainfall episodes. It is important to remark that the patterns are frequent episodes where each episode denotes three events, and each event is a period of rainfall points. The patterns are divided into types of patterns: normal patterns (more frequents occur) which denotes the rainfall in class L and M, non-normal (unusual patterns which rarely occur and are very interesting to the experts) that denote the patterns with more N and H classes. Normal patterns can be seen in 9 months of the year. The non-normal patterns can be seen in eight different months of the year. It indicates N to H rain condition in certain duration. The expert verifies that pattern of April are interesting and very similar to the current time.

TABLE II

Examples of frequent serial episodes in January.

No.	episodes	freq	conf.	No.	episodes	freq	conf.
1	LLM	91	0.59	9	HMH	31	0.15
2	LMM	47	0.22	10	MHL	41	0.20
3	MMM	31	0.15	11	HLH	51	0.24
4	MML	50	0.24	12	LHL	91	0.59
5	MLM	39	0.19	13	NLL	36	0.17
6	MLL	93	0.60	14	LLN	37	0.18
7	LLH	100	0.64	15	LNL	39	0.19
8	LHM	51	0.24	16	LML	101	0.65

TABLE III

Examples of normal and non-normal patterns from symbolic rainfall data sets.

normal patterns			non-normal patterns		
January: light	April: moderate	August: moderate	January: heavy	May: heavy	September: heavy
L M→ L (0.65)	L L → M (0.46)	M L→L (0.56)	L L→ H (0.64)	H L→ L(0.50)	L H→ L (0.61)
M L→ L (0.60)		L L→ M(0.56)		L L→ H(0.49)	H H→ L (0.55)
February: light	June: light	October: light	February: heavy	June: no rain	October: heavy
M L→ L (0.51)	M L→ L (0.61)	& moderate	L L→ H(0.69)	L L→ N(0.53)	H L→ L (0.55)
	L M→ L (0.50)	L M→ L(0.52)	H L→ L(0.65)	L N→ L(0.50)	L L→ H((0.51)
March: light	July: moderate	December: moderate	April: heavy	July: heavy	November: heavy
M L → L (0.57)	M L→ L (0.59)	M L→ L (0.56)	H L→ L(0.53)	L L→ H (0.55)	H L→ L (0.60)
L L → M(0.56)		L M→ L (0.52)	H L→ H(0.51)	L H→ L (0.56)	L L→ H (0.61)
L M → L (0.61)			H H→ L (0.50)		L H→ L (0.53)
				August: heavy	
				L L→ H0.43)	
				H L→ L(0.44)	

4. Conclusion

In this paper, interesting and new patterns of Malaysia rainfall are discovered through sequential pattern mining. Series of episodes are detected that explain the overall rainfall pattern over the 12 months of the year. It shows that rainfall is light in four months, moderate in four months, and heavy in six months of the year.

Some patterns can be used as rules for prediction. In conclusion, the mining of rainfall in both phases for frequent and sequential patterns results in very useful patterns. Those patterns can be used to help experts to build prediction models for all rainfall stations.

References

[1] I.S. Isa, S. Omar, Z. Saad, N.M. Noor, M.K. Osman, in: *2nd International Conference on Computational Intelligence, Communication Systems and Networks*, 2010, p. 96.

[2] I.I.A. Rahman, N.M.A. Alias, in: *IEEE High Capacity Optical Networks and Enabling Technologies*, 2011, p. 323.

[3] R. Lee, J. Liu, *IEEE T. Syst. Man. Cy. C* **34**, 369 (2004).

[4] V.M. Zavala, E.M. Constantinescu, M. Anitescu, in: *IEEE PES*, 2010, p. 1.

[5] V.G.N. Nguyen, H.X. Huynh, T.T. Vo, A. Drogoul, in: *MEDES'11*, 2011, p. 150.

[6] H. Saima, J. Jaafar, S. Belhaouari, T.A. Jillani, in: *National Postgraduate Conference (NPC 2011)*, 2011, p. 1.

[7] A.K. Tripathy, S. Mohapatra, S. Beura, G. Pradhan, *Int. J. Scientific Eng. Res.* **2**, 1 (2011).

[8] J. Gill, B. Singh, S. Singh, in: *8th International Symposium on Intelligent Systems and Informatics*, 2010, p. 465.

[9] C. Li, Y. Wang, X. Liu, in: *4th International Congress on Image and Signal Processing*, 2011, p. 1775.

[10] H. Mannila, H. Toivonen, A.I. Verkamo, *Data Min. Knowl. Discov.* **1**, 259 (1997).

[11] T. Katoh, K. Hirata, M. Harao, *Mining Frequent Diamond Episodes from Event Sequences*, MDAI, 2007, p. 477.

[12] T. Katoh, H. Arimura, K. Hirata, *Mining Frequent Bipartite Episode from Event Sequences*, Eds. J. Gama, V.S. Costa, A.M. Jorge, P.B. Brazdil, *Discovery Science*, Springer, Berlin 2009, p. 136.

[13] F. Höppner, in: *Proceedings of the ECAI'02 Workshop on Knowledge Discovery*, 2002, p. 25.

[14] J.F. Allen, *Commun. ACM* **26**, 832 (1983).