Middle-East Journal of Scientific Research 24 (7): 2375-2384, 2016
ISSN 1990-9233
© IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.07.23799

Bag of Patterns Representation Technique of Constructed Detection Temporal Patterns for Particular Climatic Time Series

[1]Yahyia BenYahmed, [1]Azuraliza Abu Bakar,
[1]Abdul RazakHamdan and [2]Sharifah Mastura Syed Abdullah

[1]Center for Artificial Intelligence, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600 Bangi, Selangor DarulEhsan, Malaysia
[2]Institute of Climate Change, Universiti Kebangsaan Malaysia,
43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract: Pattern detection is one of the critical problems in time series analysis which try to detection subsequences of time series for the analysis patterns and trends of time series. In this paper, a proposed method to detect benefit patterns to identify the unusual or suddenly events from large historical time series data. The proposed insight by analysing univariate time series based on a histogram-based representation for weather data using Bag of Pattern representation (BoP) that is commonly applied by the information retrieval communities and text mining. Our approach first segments a time series into subsequences using a sliding windows approach and then converts the local segments into symbolic representation considered as patterns by using Symbolic Aggregate Approximation (SAX). Finally, the obtained patterns are represented by BoP representation which may be provided to group in terms of domain experts or statistical features to identify each pattern. The presented experiments on 35 years of Malaysia weather data. In this context, we will select these data to show the benefit and correlation between the variables, which consider for rainfall and river flow data for trying to make prediction of future river flow values. The datasets are also in order to evaluate the effectiveness of the proposed method. The BoP is proved to be more efficient in reserving the trends of time series which identify suddenly changes, repeated or unrepeated patterns on weather time series.

Key words: Time Series · Pattern Detection · Sliding Windows · Symbolic Approximation · Bag of Pattern

INTRODUCTION

Temporal pattern mining is useful as tool to extract nontrivial, implicit, previously unknown and potentially useful information or pattern from large databases. It is the process of applying the mining tasks like clustering, classification, prediction, novelty and motif detection to data with the intention of uncovering hidden patterns [1]. In recent years, a growing amount of time series data stored with the high dimensions have encouraged the researchers to make a great interest in the patterns detection or creating models of enormous data sets, also referred to as knowledge discovery and data mining. Furthermore, many of the scientific applications that primarily serve the temporal sequences to predict future events effectively [2]. Many algorithms have been presented for patterns detection [3-5]. Extracting this kind of patterns is that each analysis is related with a time. The idea of detecting repeated patterns or suddenly changes that occur over a specified interval that is less than the collected entire database based on the first and last events of each pattern.

Pattern detection problem, purposed to detect a various type of significant patterns from a large time series data has been defined in [6]. The goal is to provide the frequent patterns between the variables of large time series data, which give a real correlation and meaningful information about the area. Using these patterns may be for weather prediction to identify new and unexpected events or trends and hidden relations in time series data

Corresponding Author: Yahyia BenYahmed, Center for Artificial Intelligence, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor DarulEhsan, Malaysia.

and use them to find out more information about natural of the weather. In weather analysis, the extraction of frequent patterns leads to extract sets of normal of abnormal patterns that predict future events for human life. If these patterns are identified correctly from the domain experts or statistical features then this patternsets are considered as perfect information about the future [7].

In this work, investigating the issue of mining multidimensional of time series for detecting and learning new patterns, which group in terms of domain experts or statistical features to identify meaningful pattern, using Bag of Pattern representation (BoP) that is commonly applied by the information retrieval communities and text mining [8, 9]. Our approach first segments a time series into subsequences which considered as local segments, the time series segmentation is established in a Sliding Windows approach (SW) [10] and then converts the local segments into symbolic representation considered as patterns by using Symbolic Aggregate approXimation (SAX) [11]. Finally, the obtained patterns are represented by BoP representation which may be provided to group in terms of domain experts or statistical features to identify each pattern, for instance, the patterns of rainfall variable might define as no rain, light, moderate, or heavy. This is essentially useful dealing with the detection of unusual trends or events time series. The goal is to provide information about patterns trend of time series in the suddenly changes or unusual behaviour of the weather in spite of the observed difference of the natural phenomena. In this context, to detect normal or danger patterns in weather status, we aim to learn the human to make weather prediction in future, which is complex to any different domain and then to discover any suddenly or unusual events in comparison with this prediction.

Related Work: The pattern detection is the process of finding temporal symmetries within the time series and the aim of analysing a time series is to detect benefits frequent patterns is repeated within time intervals to identify the unusual or suddenly events from a large amount of data sets in historical time series [1]. Find the similarity is a useful tool to explore the time series databases to find a specific pattern in large time series. The Efficiency and accuracy of similarity searching is an important problem for time series representation. Many of the Dimensional reduction techniques [11-15] have been proposed for effective representation of time-series data. Symbolic Aggregate Approximation (SAX) was proposed for symbolic time series representation [11]. SAX allows the discretization of original time sequences into symbolic strings and distance measures to be defined on the symbolic approach. However, SAX is based on the Piecewise Aggregate Approximation (PAA) representation that minimizes dimensionality by the mean values of equal sized frames[14].

Patterns detection part should be carried out before time series mining. The key concept of pattern detection is to find the number of change points first and identify the class of those points as one window (pattern). Detection has also been studied for a long time in statistics literature where signal with a series of best fitting lines and return the end points of the segments as change points or sequence of time points known as a window (pattern) [16, 17]. Pattern detection of time series was a major problem in any application domain where the gather data is the temporal aspect. Traditional methods [18, 19] detect each data instance (a univariate record) independently and ignoring the sequence aspect of the data. Often, in time series, patterns can be detected only by analysing data instances together as a sequence and hence cannot be detected by the traditional detection techniques. Recently, [20] proposed a clustering approach to clustered similar trends and to compare such clusters based on Self Organizing Maps (SOMs) algorithm. Furthermore, during the detection phase, time series need to be passed through processes segmentation and classification.

In [21] have consider subsequences of fixed and random lengths The random length subsequences can potentially detect patterns that appear with different lengths and be split across the time points. As an alternative, [22] subsequences are generated in a fixed, uniform manner. Moreover, time series segmentation approach [23] is presented to make smaller segments represent as subsequences, the sliding windows algorithm used to capture meaningful patterns along the time series. The approach slides the time series in term of specific stopping criteria is met to extract subsequences from time series which consider as local segments [10]. The approach of sequence analysis [6] have presented Apriori algorithm to define a frequent pattern in terms of the changing frequency of frequent patterns with time, to detect the relationship between the non-trivial patterns or events in the temporal database, which identify the change point or trends in the data [16]. BoP representation have applied to extract local temporal or frequency information to characterize time series. BoP is suitable performance for pattern detection, mostly in pattern recognition [24-26], computer vision as image processing [27, 28] and natural images classification [29].

Our work is based on the BoP approach in which complex sequence are characterized by pattern vectors of subsequences. A BoP is referred to as a Bag of Instances in multiple instance learning [30, 31] and a Bag of Frames in speech and audio recognition [32, 33] and a bag-of-words (BoW) [22] that is inspired by text documents analysis [34, 35] to represent time series on sequence level. Local segments extracted from time series are encoded as words using the symbols produced by SAX algorithm.

MATERIAL AND METHODS

Temporal pattern mining is indicated to data, which includes a series of events that is difficult to detect from large amounts of temporal data. Pattern detection is one of the essential problems in time series analysis which try to find subsequences of time series for the analysis patterns and trends of time series.

In this paper, a proposed method to detect benefits frequent patterns to identify the unusual or suddenly events from large historical time series data. The proposed insight by analysing time series based on a histogram-based representation for weather data using Bag of Pattern representation (BoP) that is commonly applied by the information retrieval communities, text mining and computer vision approach [22, 36, 37].

Bag of Patterns Representation (BoP): Bag of Patterns (BoP) representation is text mining and computer vision approach developed to extract high structural information from text document. The BoP representation is efficient to detect the patterns similarity in time series [22, 36, 37]. The BoP used in time series for classification, clustering and anomaly detection to disregard the repeated information of patterns and count the occurrence of patterns in time series to produce a pattern frequencies. The BoP represents as a dictionary that have a set of codebook and shows a time series as a histogram of codebook occurrence which indicate the number of a codebook occurred in the time series [34, 35].

Using BoP representation to detect structural information in weather time series. The data is converted into a frequency patterns. The patterns are created by sliding the time series to detect all subsequences in a window size *w* using sliding windows approach [10], the subsequences considered as local segments and then converting the local segments into a SAX representation [38] that uses a window size w to extract patterns in local subsequences and count the occurred patterns in time series to create a frequency information.

Time Series Segmentation: Time series segmentation is commonly used routines of time series data aim to make an accurate approximation of time series, by reducing its dimensions with preserving the important features [11, 39]. Segmentation used to determine suitable periods of time and making a good approximation of the time series. According to [23], There exist a heuristic approach to extract corresponding segments that can be classified into three groups; first sliding window algorithm [10], a segment is merged until specific error criterion is happened, second top-down algorithm [40], a time series is recursively separated until some error measure is happened and bottom-up algorithm [41], starting with small segments that are grown until some error criterion is happened. Least Squares Error (LSE) is often used to measure the error criterion. Sliding window algorithms that are given very fast time series analysis and are appropriate for many variant applications.

The segmentation problem can be framed in several ways. For instance, time series t produces s segments based on the maximum error for any segment. Some segment does not exceed some user specified threshold. Whereas the combined error of all given time series produces segments is less than some user specified threshold [15]. A number of segmentation algorithms not only defines segment boundaries also mentioned to as segmentation points but also a Piecewise Linear Approximation representation (PLA) of the time series within a segment, where linear interpolation or linear regression is statistical methods use to represent the time series as segments.

SAX Representation: The local segments which obtained from Sliding Windows algorithm convert to symbolic representation using SAX representation. The process primarily of SAX algorithm the data is normalized using Z-Score normalization that the average is 0 and the standard deviation is 1. After that, converts the data into the PAA representation and later signifies the PAA representation into a significant pattern [15].

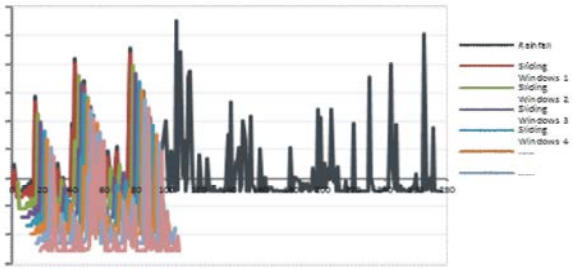

Fig. 1: Sliding windows algorithm using windows size is 90 for rainfall data

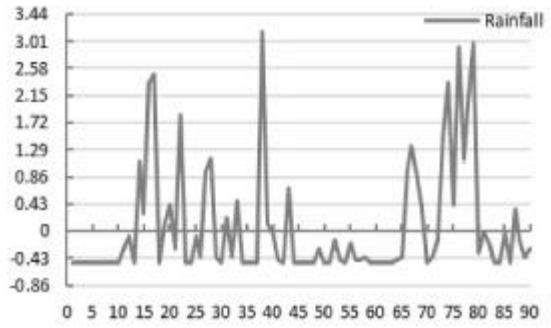

Fig. 2: Time series converted by SAX algorithm, adopted from [9].

After applying PAA, the values are later converted into specific values by means of a probability distribution function. The SAX algorithm, to convert a real valued time series into a SAX string representation has three parameters, time series length (n), word size (w), alphabet size (a). An efficient number of this parameters are inherited from [9]. Then, SAX was run with the same parameters a=3 and w=7. Fig. 2 illustrates SAX representation processing, a time series is converted to SAX symbols by getting PAA estimation and mapping the PAA coefficients to SAX codes. An example, with n = 90, w = 7, a = 3 and the separations occur at -0.43 and 0.43 in a normalized Gaussian distribution, the time series is associated with the word 'acbbacb. Note that in this example, the three characters a, b and c is about equally likely as desired. In effect, this is a concatenation of symbols representing the sequence of word to generate sufficient reduction of time series data. The output of the SAX representation is a new value of the time series represented as a time series data reduction. The new time series length (N) for each year generates 276 values approximately N=276.

BoP Representation: The representation takes the symbolic univatiate time series representation to generate the local patterns and use a histogram representation of pattern as the new patterns representation, which can be equated to the patterns in time series and counts the occurrence of these patterns to create a vector of patterns frequencies [22, 42]. After the local pattern detection, a new dataset is created where the patterns from each subsequence provide an instance and each time series forms the bag. The local patterns are represented by BoP representation which may be provided to group in terms of domain experts or statistical features to identify the class label for each instance is the class of the homologous time series.

In BoP representation, the frequency patterns are stored in the bag of patterns vector and extract the codebook as the summary of the weather time series. A BoP representation allows integrating local information from a vector of patterns frequencies of the time series in an efficient way. Table 1 illustrates the frequency vectors of rainfall time series for 35 years. The similarity measure between the vectors can be calculated using Euclidean distance.

The useful of a pattern can be measured with a maximum confidence threshold. Counting frequencies of patterns is comparatively straightforward, patterns are sequences of events that occurrence orderly and the confidence is counting the number of patterns in which they occur. BoP representation is applied to extract the most frequency patterns that are measured by support and confidence. Assuming every c_i in time series C appears at least once. Then, we define equation to represent confidence $conf(c_i)$ and support $sup(c_i)$ to count the maximum occurrence of each C. Eq. (1) show the confidence.

$$conf(c_i) = \frac{sup(c_i)}{sup(C)} \quad (1)$$

where $sup(c_i) = \{ci \in C\}$.

In pattern detection, BOP representation is applied to detect the patterns with fixed width. Then the frequency patterns are designed to detect patterns with the most confidence. This relation compares number of events containing patterns to the number of all events in time series.

Experimental Design: Temporal frequent pattern detection is one of the important problems in time series analysis which try to find structural information for the analysis patterns and trends of time series. Experiments were conducted in which the SAX representation was integrated into the BoP representation to generate sufficient patterns of weather time series. The useful of a pattern can be measured with a maximum confidence threshold. The performance of BoP was showed different time series representation and proved to be more efficient in reserving the trends of time series which identify suddenly changes, repeated or unrepeated patterns on weather time series.

The performance of the BoP representation was applied on 2 time series datasets for 35 years of weather data, the data consisting of daily rainfall data refers to name station is 8R-Empangan Sg.Semenyih_Sel, 16R-SK Sg.Lui_Sel and 53R-Ladang Dominion. While river flow

Table 1: An example of BoP representations of rainfall time series for 35 years [9].

		Time Series data																
		1	2	3	4	5	6	7	.	.	.	.	31	32	33	34	35	Total
Codebook	accbacc	2	.	.	3	.	.	.	6	.	.	9	.	5	.	.	7	32
(SAX representation)	aacbbcb	.	.	.	.	.	.	.	.	.	.	2	.	.	5	.	.	25
	aababcc	1	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	18
	abababc	.	1	.	.	.	1	.	.	5	.	.	.	3	.	.	.	20
	ababbab	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
	ababbbb	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	2
	ababbbc	.	.		.	3	.	.	.	.	1	.	.	3	.	.	.	15
	babcbbb	2	.	.	.	.	.	9	.	.	8	6	.	.	4	.	.	40
	babcbbc	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1
	babccbb	.	9.	1	.	5	.	5	.	.	1	9.	6	4	3	7	.	50
	bacaabb	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12
	bbaabab	.	.	.	.	.	2	.	.	.	3	.	.	.	.	1	.	8
	cccbbbb	.	.	.	.	.	.	.	1	1	.	.	2	.	.	.	.	4
	cccbbcb	.	.	.	.	.	.	7	.	7	.	.	1	.	.	.	.	21
	cccbcbb	.	.	.	.	.	1	.	.	.	2	.	.	.	.	.	.	3
	ccccbbb	.	.	.	.	1	.	.	.	.	.	.	9	.	.	7	.	24
	cccbcaa	.		.	.	.	.	.	.	.	7	.	.	1	.	.	.	10

data refers to name station is OrgQ-SgSemenyih-KgRinching in Malaysia from 1975 to 2009 that collected from Institute of Climate Change, UKM.

The proposed insight by analysing univariate time series based on a histogram-based representation for weather data. The time series segmentation approach and symbolic representation were used in the experiment to obtain meaningful patterns using BoP representation to detect benefit patterns and identify the unusual or suddenly events from large historical time series data. The BoP representation may be provided to group in terms of domain experts or statistical features to identify each pattern. In this context, we select these data to show the benefit and correlation between the variables, which consider for rainfall and river flow data for trying to make mining task as classification, prediction or anomaly detection. The datasets are also in order to evaluate the effectiveness of BoP representation. The representation is proved to be more efficient in reserving the trends of time series which identify suddenly changes, repeated or unrepeated patterns on weather time series.

RESULTS AND DISCUSSION

The experimental results showed that the BoP representation provide useful information with different time series representation in terms of identifying suddenly changes, repeated and unrepeated patterns on weather time series. BoP representation yielded extremely meaningful patterns in most datasets that have small alphabet (a) and word sizes (w). The obtained patterns are represented by BoP representation which may be classified to group in terms of domain experts or statistical features to identify each pattern, which may help the experts to make future prediction. Based on sliding windows algorithm for univariant rainfall and river flow data for 35 years with sliding windows $s = 90$, we get 9669 subsequences which consider as local segments, this subsequences may contain rich local information about the rainfall or river flow data, we use $s = 90$ as a season period per year to identify new and unexpected events or trends and hidden relations in time series data and use them to find out more information about natural of the weather. Fig. 1 illustrates sliding windows algorithm. After that, Then, SAX representation was run with the same parameters the alphabet size $a = 3{:}4$ and word size $w = 6{:}8$ to represent the local segments as SAX symbols that are considered as patterns. We consider $a = 3$ and $w = 7$ which given the length of pattern is 7 with 3 symbols is a, b and c. The efficient number of intervals for a and w of SAX parameters inherited from [9]. Furthermore, we extract whole patterns in term not be repeated in 9669 patterns, which generates a dictionary of as the summary information of the rainfall or river flow data that consider as codebook. The result of size codebook in rainfall data is 400 patterns and river flow is 520 patterns. Finally, The BoP representation applied to represent as a dictionary that has a set of codebook and shows a time series as a histogram of codebook occurrence which indicate the number of a codebook occurred in the time series. The resulting codebook of descriptors is quantized through the codebook as the summary of the rainfall or

Table 2: An example of BoP of rainfall data and number of observations

No	Time Codebook	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total
1	*abbabab*	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	7
2	*abbabba*	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
3	*abbabbb*	0	0	0	0	2	3	1	0	0	0	4	7	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0	21
4	*abbabbc*	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
5	*abbabcb*	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
6	*abbacbb*	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	*abbbabb*	0	0	0	3	8	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	5	1	0	1	0	0	0	0	0	0	0	0	21
8	*abbbabc*	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	5
9	*abbbacb*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
...	*abbbbab*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6
...	*abbbbbb*	3	19	3	3	3	7	0	0	10	0	13	12	6	0	6	5	6	8	6	16	3	4	4	15	1	4	8	1	6	3	5	0	2	0	1	183
...	*abbbbbc*	0	0	0	3	9	0	0	2	0	0	0	2	2	0	0	0	1	3	0	1	0	2	9	0	1	9	1	0	0	1	0	0	0	0	0	46
....	*abbbbca*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
....	*abbbbcb*	0	0	0	4	0	0	2	0	0	0	0	0	0	0	0	8	2	4	0	0	0	0	0	0	1	3	0	0	0	6	1	0	0	0	0	31
....	*abbbbcc*	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	8
396	*abbbcba*	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
397	*abbbcbb*	0	0	0	0	0	4	0	0	0	0	4	0	0	0	0	5	4	0	0	6	1	0	0	0	0	0	0	4	0	1	0	0	4	1	0	34
398	*abbbcbc*	0	0	0	1	0	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	9
399	*abbbcca*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
400	*abbbccb*	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	7

Table 3: Frequent patterns in different time periods

No	Sequence	Frequent	Conf	No	Sequence	Frequent	Conf	No	Sequence	Frequent	Conf
1	accbacc	32	0.003	16	bccbbbb	59	0.006	31	bbcbbbb	349	0.036
2	aacbbcb	25	0.003	17	bccbbbc	19	0.002	32	bcbbaca	1	0.000
3	aababcc	18	0.002	18	bccbbcb	9	0.001	33	cbbbabb	38	0.004
4	abababc	20	0.002	19	bcbcbbb	27	0.003	34	cbbbabc	4	0.000
5	ababbab	1	0.000	20	bcbcbbc	4	0.000	35	cbbbbab	11	0.001
6	abababb	2	0.000	21	bcbcbcb	10	0.001	36	cbbbbba	14	0.001
7	ababbbc	15	0.002	22	babcbbb	40	0.004	37	cbbbbbb	245	0.025
8	abbbbba	6	0.001	23	babcbbc	1	0.000	38	cbbbbbc	42	0.004
9	abbbbbb	183	0.019	24	babccbb	50	0.005	39	cbbbbca	2	0.000
10	abbbbbc	46	0.005	25	bacaabb	12	0.001	40	cbbbbcb	66	0.007
11	bbbbbab	156	0.016	26	bbaabab	8	0.001	41	cbbbbcc	21	0.002
12	bbbbbac	12	0.001	27	bcbbbbb	274	0.028	42	cbbbcaa	6	0.001
13	bbbbbba	205	0.021	28	bbccbbb	88	0.009	43	cccbbbb	4	0.000
14	bbbbbbb	2787	0.288	29	bcbbcbb	37	0.004	44	cccbbcb	21	0.002
15	bbbbbbc	337	0.035	30	bcbbbab	40	0.006	45	cccbcbb	3	0.000

river flow data. A BoP representation allows integrating local information from subsequences of the time series data in an efficient method. Table 2 shows the sample of result for BoP representation is applied on rainfall data for 35 years.

As examples, first pattern is *abbabab* has repeated 7 times during 35 year, in 1985 year is 4 times, in 1991 year is 1 time and in 2002 year is 2 time. In other pattern is *abbbbbb* is repeated 183 times, which may be recurring patterns, or periodical patterns. At variance, *abbbacb*, *abbbacb* and *abbbcba* are repetitive 1 time, which drives to these patterns are surprising patterns, anomaly patterns or suddenly change patterns. The useful pattern can be measured with a extreme confidence threshold. Counting frequency patterns is relatively clear, the patterns are sequences of events that occurrence orderly and the confidence measure is counting the number of patterns in which they occur.

Table 3 shows an example of the frequency patterns with different years in 8R-Empangan Sg.Semenyih_Sel. BOP representation generate 9669 patterns for rainfall data set in 35 years. In the table extracted pattern *ababbab* occurred one times, which means the ababbab is occurrence one time in 35 years with 0.000 confidence that would indicate new rules for specific period and given data sets. Another frequent pattern can be seen in Table 3 is *bbbbbbb*, that indicates 2787 times occurred with 0.028 confidence, we can conclude the patterns which have 0.000 confidence indicate to surprise patterns or sudden changes events and more than 0.50 confidence which denote to normal or natural events during 35 years.

Form the domain experts define the pattern that have repeats extremely may be regular or normal patterns and the patterns that is repeated slightly may be surprising or abnormal patterns. Fig. 3 explains some affecting patterns, which may give more information about

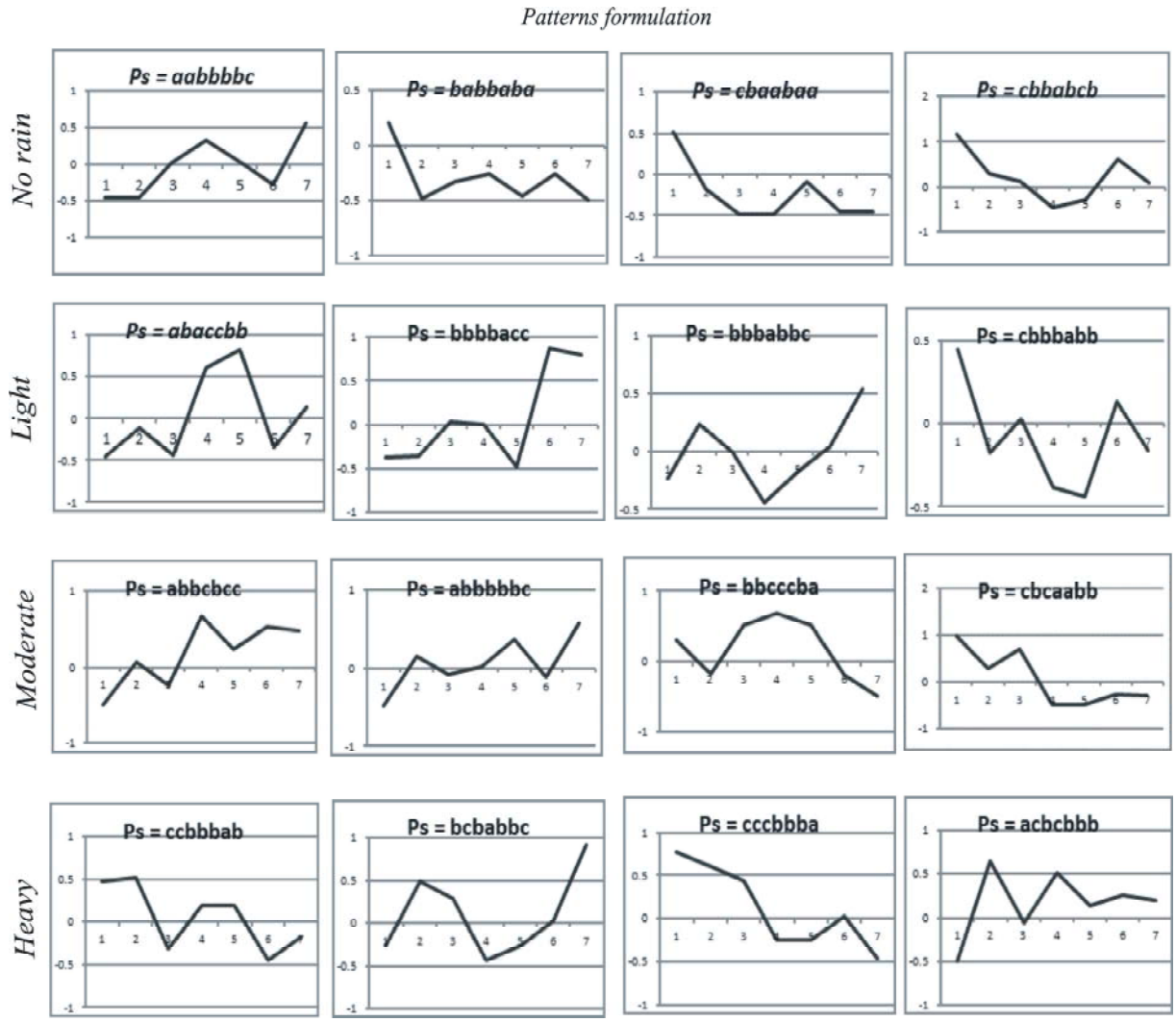

Fig. 3: Patterns formulation for rainfall data

the patterns. The obtained patterns are may be classified to cluster in terms of domain experts or using classifier approach to identify each pattern, which may help the experts to make future prediction. However, in this work founded on domain experts to identify the rainfall patterns in 4 group are no rain, light, moderate and heavy. Fig. 3 displays the groups of rainfall impact into different patterns formulation.

An example, we see the motion of different rainfall patterns are *abbcbcc*, *abbbbbc*, *bbcccba* and *cbcaabb* when the rainfall is moderate, the patterns are *ccbbbab*, *bababbc*, *cccbbba* and *acbcbbb* define as heavy rainfall and so that. With that same approach, we have applied BoP representation into river flow data to demonstrate the different patterns, which identify as normal, alert, warning and danger.

Based on the patterns results of rainfall data indicated in Fig. 3, the similarity analysis considered in this work. Fig. 4.a shows frequencies patterns on 35 years of rainfall data, the number of pattern is 400 patterns. From the histogram, we see which patterns are repeated considers as regular patterns and unrepeated pattern considers as surprising patterns. For instance, on time 190 the pattern defines as *bbbbbbb* and the rainfall status is Light which repeated 2787 time during 35 years. The patterns *bbbbcbb* is Light, *bbcbbbb* is Moderate and *bbbcbbb* is Moderate which repeated 355, 349 and 348 time on times 195, 225 and 204 respectively during 35 years. Whole these patterns may define as regular patterns. At variance, the patterns *cbbcbcc* is Light on time 354, *cbbccab* is Moderate on time 355, *cbbcccb* is Moderate on time 358 and *cccbbba* is Heavy on time 396

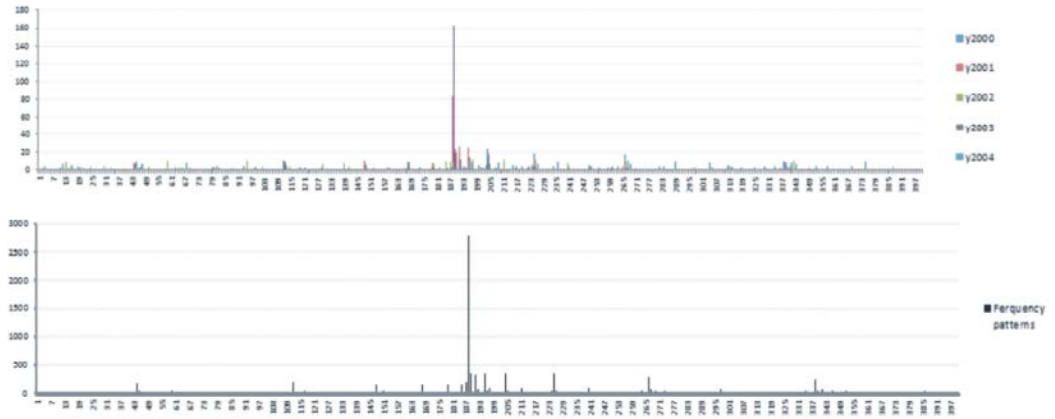

Fig. 4: Frequencies Patterns of rainfall data for 35 years.

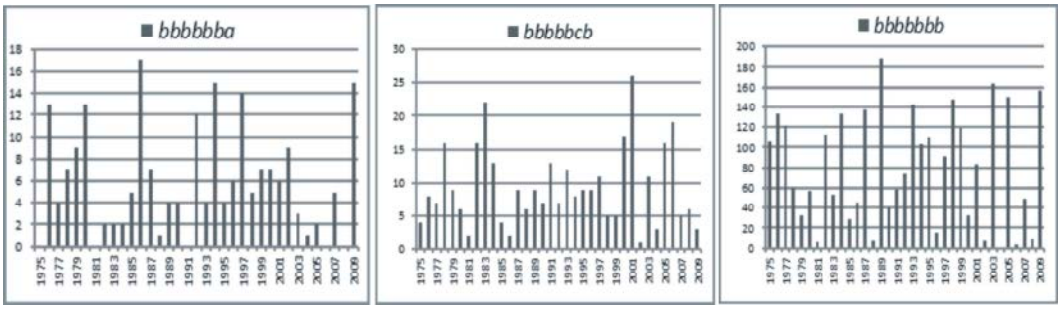

Fig. 5: Frequencies Patterns of rainfall data during 35 years.

have repeated one time during 35 years which meaning these patterns may define as surprising patterns. In Fig. 4.b illustrates the whole frequency patterns during 2000 year to 2009 year. In 190 time sows the *bbbbbbb* has repeated 119, 32, 83, 8 and 163 times respectively. Different pattern on time is *cbbbbbb*, which repeaters 16, 9, 9, 9 and 0 times respectively and so forth.

For more analysis, we can see each pattern how is effected during all 35 years. As the examples Fig. 5 show 3 patterns are *bbbbbba*, *bbbbbcb* and *bbbbbbb* how quantity of repeated time in each year. The *bbbbbba* pattern defines as Light rainfall and it is repeated 17 times in 1986 year, the 7 times in 2000 years and one time in 2004 and so forth. Different pattern is *bbbbbcb* consider also as Light rainfall is repeated 12 times in 1993, 26 times in 2001, 5 times in 2007 and one time in 2001 and so forth.

The representation of the patterns results by BoP representation may help the experts of Malaysian Climate Changes Institutes easy to understand and identify the diverse patterns to make future rainfall or river flow prediction. However, the performance of the BoP representation and fix segmentation of time series was allowed one to integrate benefit information from different patterns of the time series in an efficient way.

CONCLUSIONS

The task of representation in time series data mining is critical because the direct handling of continuous data with high dimensionality is extremely challenging to manage efficiently. Pattern detection is one of the essential problems in weather analysis which try to extract subsequences of time series to analysis the patterns and trends of time series, then applied Bag of Pattern representation (BoP) to detect benefit patterns to identify the unusual or suddenly events from large historical of weather data. In the first alternative, partitions a sequence into equal-length segments named intervals using sliding windows approach to extract subsequences from all locations with fix lengths of subsequences, which can be chosen based on certain domain knowledge. The partition into intervals allows detecting patterns represented by shorter time segments. Then, each subsequences represent by SAX algorithm which need to identify 2 parameters word size and alphabet size (w, a). The BoP generates as a dictionary represents as codebook for each univariate time series which shows as a histogram of codebook occurrence which indicate the number of a codebook occurred in the time series. Finally, the obtained

patterns are represented by BoP representation which may be provided to group in terms of domain experts to identify each pattern. The approach have applied in on 2 dataset for 35 years of Malaysian weather data, daily rainfall and river flow data that collected from Institute of Climate Change, UKM. The results of our experiments and analyses showed that the BoP representation can potentially prove to be more efficient in reserving the trends of time series which identify suddenly changes, repeated or unrepeated patterns on weather time series. This goal is very significant to achieve for instance when analysing weather time series data, in which patterns from each time series may contribute important information that would otherwise be lost due to the application of less effective methods. Although our focus in this study is on the similarity analysis of the weather time series, the BoP approach can be adjusted to mining tasks such as classification, clustering, anomaly detection and so forth.

ACKNOWLEDGMENT

This work is granted by project RGS/1/2012 /ST G07/ UKM/01/1, Ministry of Higher Education, Government of Malaysia.

RERRENCES

1. AFan, W. and A. Bifet, 2013. Mining big data: current status and forecast to the future. ACM SIGKDD Explorations Newsletter, 14(2): 1-5.
2. AEsling, P. and C. Agon, 2012. Time-series data mining. ACM Computing Surveys (CSUR), 45(1): 12.
3. Bayardo Jr, R.J., 1998. Efficiently mining long patterns from databases. In: ACM SIGMOD Record. ACM: pp: 85-93.
4. Wang, J., J. Han and J. Pei, 2003. Closet+: Searching for the best strategies for mining frequent closed itemsets. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp: 236-245.
5. Han, J., J. Pei and Y. Yin, 2000. Mining frequent patterns without candidate generation. In: ACM SIGMOD Record. ACM: pp: 1-12.
6. Agrawal, R., T. Imieliński and A. Swami, 1993. Mining association rules between sets of items in large databases. In: ACM SIGMOD Record. ACM: pp: 207-216.
7. ASaleh, B. and F. Masseglia, 2011. Discovering frequent behaviors: time is an essential element of the context. Knowledge and information Systems, 28(2): 311-331.
8. Saeed, M. and R. Mark, 2006. A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association: pp: 679.
9. Lin, J. and Y. Li, 2009. Finding structural similarity in time series data using bag-of-patterns representation. In: Scientific and Statistical Database Management. Springer: pp: 461-477.
10. Palpanas, T., M. Vlachos, E. Keogh, D. Gunopulos and W. Truppel, 2004. Online amnesic approximation of streaming time series. In: Data Engineering, 2004. Proceedings. 20th International Conference on. IEEE: pp: 339-349.
11. Lin, J., E. Keogh, S. Lonardi and B. Chiu, 2003. A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM: pp: 2-11.
12. Agrawal, R., C. Faloutsos and A. Swami, 1993. Efficient similarity search in sequence databases: Springer.
13. Chan, K.P. and A.C. Fu, 1999. Efficient time series matching by wavelets. In: Data Engineering, 1999. Proceedings., 15th International Conference on. IEEE: pp: 126-133.
14. Akeogh, E., K. Chakrabarti, M. Pazzani and S. Mehrotra, 2001. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3): 263-286.
15. Akeogh, E., K. Chakrabarti, M. Pazzani and S. Mehrotra, 2001. Locally adaptive dimensionality reduction for indexing large time series databases. ACM SIGMOD Record, 30(2): 151-162.
16. ARoddick, J.F. and M. Spiliopoulou, 2002. A survey of temporal knowledge discovery paradigms and methods. Knowledge and Data Engineering, IEEE Transactions on, 14(4): 750-767.
17. Epifani, I., C. Ghezzi and G. Tamburrelli, 2010. Change-point detection for black-box services. In: Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering. ACM: pp: 227-236.
18. AHodge, V.J. and J. Austin, 2004. A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2): 85-126.
19. AChandola, V., A. Banerjee and V. Kumar, 2009. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3): 15.

20. Nohuddin, P.N., F. Coenen, R. Christley and C. Setzkorn, 2010. Detecting temporal pattern and cluster changes in social networks: A study focusing uk cattle movement database. In *Intelligent Information Processing V*: Springer, pp: 163-172.
21. AFu, T.C., 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1): 164-181.
22. AWang, J., P. Liu, M.F. She, S. Nahavandi and A. Kouzani, 2013. Bag-of-words representation for biomedical time series classification. Biomedical Signal Processing and Control, 8(6): 634-644.
23. Lin, J., E. Keogh, S. Lonardi, J.P. Lankford and D.M. Nystrom, 2004. Visually mining and monitoring massive time series. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: pp: 460-469.
24. ABabenko, B., M.H. Yang and S. Belongie, 2011. Robust object tracking with online multiple instance learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33(8): 1619-1632.
25. Dollár, P., B. Babenko, S. Belongie, P. Perona and Z. Tu, 2008. Multiple component learning for object detection. In Computer Vision–ECCV 2008: Springer. pp: 211-224.
26. Raykar, V.C., B. Krishnapuram, J. Bi, M. Dundar and R.B. Rao, 2008. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In: Proceedings of the 25th international conference on Machine learning. ACM: pp: 808-815.
27. Rahmani, R. and S.A. Goldman, 2006. MISSL: Multiple-instance semi-supervised learning. In: Proceedings of the 23rd international conference on Machine learning. ACM: pp: 705-712.
28. Zhang, C., X. Chen, M. Chen, S.C. Chen and M.L. Shyu, 2005. A multiple instance learning approach for content based image retrieval using one-class support vector machine. In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE: pp: 1142-1145.
29. Maron, O. and A.L. Ratan, 1998. Multiple-Instance Learning for Natural Scene Classification. In: ICML. Citeseer: pp: 341-349.
30. AChen, Y., J. Bi and J.Z. Wang, 2006. MILES: Multiple-instance learning via embedded instance selection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(12): 1931-1947.
31. ADietterich, T.G., R.H. Lathrop and T. Lozano-Pérez, 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence, 89(1): 31-71.
32. Briggs, F., R. Raich and X.Z. Fern, 2009. Audio classification of bird species: A statistical manifold approach. In: Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. IEEE: pp: 51-60.
33. AFu, Z., G. Lu, K.M. Ting and D. Zhang, 2011. Music classification via the bag-of-features approach. Pattern Recognition Letters, 32(14): 1768-1777.
34. AHofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1-2): 177-196.
35. ABlei, D.M., A.Y. Ng and M.I. Jordan, 2003. Latent dirichlet allocation. the Journal of machine Learning Research, 3: 993-1022.
36. Fei-Fei, L. and P. Perona, 2005. A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE: pp: 524-531.
37. Sivic, J., B.C. Russell, A.A. Efros, A. Zisserman and W.T. Freeman, 2005. Discovering objects and their location in images. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. IEEE: pp: 370-377.
38. ALin, J., E. Keogh, L. Wei and S. Lonardi, 2007. Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery, 15(2): 107-144.
39. Lonardi, J.L.E.K.S. and P. Patel, 2002. Finding motifs in time series. In: Proc. of the 2nd Workshop on Temporal Data Mining. pp: 53-68.
40. Lemire, D., 2007. A Better Alternative to Piecewise Linear Time Series Segmentation. In: SDM. SIAM: pp: 545-550.
41. Hunter, J. and N. McIntosh, 1999. Knowledge-based event detection in complex time series data. In Artificial Intelligence in Medicine: Springer. pp: 271-280.
42. Ordónez, P., T. Armstrong, T. Oates and J. Fackler, 2011. Classification of patients using novel multivariate time series representations of physiological data. In: Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on. IEEE: pp: 172-179.