PATTERN DISCOVERY ALGORITHMS FOR WEATHER PREDICTION PROBLEM

YAHYIA MOHAMMED M. ALI BENYAHMED

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

ALGORITMA PENEMUAN CORAK UNTUK CUACA RAMALAN MASALAH

YAHYIA MOHAMMED M. ALI BENYAHMED

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this dissertation is my own except for quotations and summaries which have been duly acknowledged.

17 May 2018

YAHYIA M. M. ALI BENYAHMED
P67404

ACKNOWLEDGMENT

First and foremost, all praises and thanks to the mighty Allah for giving me enough strength and blessing to complete this thesis.

Indeed, there are many wonderful people who have contributed significantly to the completion of this thesis. I owe a great deal to them.

I would like to express my genuine gratitude, appreciation, and sincere acknowledgment to my exceptional supervisor *Professor Dr. Azuraliza Abu Bakar* for her valuable guidance, generosity and freedom throughout the completion of this research. It was a great honour and privilege to work with her during my PhD program. She has always been very supportive and I am so grateful for that. Without her help and supervision, I could not cope with the difficulties and pressures that I faced during my study. This thesis in fact is the result of her extraordinary help, which directed my efforts toward achieving the objectives of this research.

Also, I would like to express my high appreciation to my co-supervisor *Professor Dr. Abdul Razak Hamdan*, for continuous aid in every aspect, academic, helpful suggestions and the good advice.

Also, I gratefully acknowledge support provided by *Professor Dato' Dr. Sharifah Mastura Syed Abdullah,* at climate change institute (IPI), UKM University.

I would also like to express my appreciation to my family for their unlimited love and kindness. I would like to give special thanks to my parents for their spiritual support, and I would like to thank them again for bringing me up to who I am today. My success symbolizes and reflects the support and love from both of them. My deepest appreciation goes to my beloved wife for being so kind hearted patient and understanding throughout PhD journey. Her tolerance of my occasional moods is a testament in itself of devotion and love. She has been my inspiration and motivation for continuing to improve my knowledge and move my career forward. I also thank my wonderful children: *Rawan, AbdAlkafi, Albassel and Ahmed* for always making me smile. Their love inspired me and empowered my spirit from a long distance. I hope that one day they can read this thesis and understand why we spent so much time out of the country away from sibling and kindred. I love you all.

Last but not least, I am very grateful to all friends who have assisted me whenever I needed them, to UKM and all the staff and members of the School of Computer Science for help. It is my pleasure to have a special appreciation to all of the DMO research group members for their friendship

ABSTRACT

Pattern discovery in weather mining are performed to predict future events based on meaningful patterns extraction from previously observed events. The pattern is describing the behavior present at a specific time that required being useful and understandable for superior performance of prediction. Various algorithms are performed to process operational prediction based on pattern discovery. The tasks are complex and difficult since the weather data is collected events over time for unusual and surprised phenomena. The weather data problem is the high-dimensional data, which are dealing in its raw format is memory consuming, noisy, complexity searching and huge events. It is needed to develop representation techniques that can reduce the dimensionality of data without substantial loss of information. The aim of study presents an effective weather pattern discovery algorithm which includes the processes of relevant and interesting pattern, to improve the effectiveness of weather prediction, with three main objectives: i) time series representation; ii) prediction; iii) pattern discovery. The first is to enhance Symbolic Aggregate approXimation (SAX) algorithm for weather representation called ESAX^{+} that can reduce the dimensionality with less loss information. The second is to propose predictor model based on Naive Bayesian algorithm (NB) for prediction pattern that integrated with ESAX^{+} representation and sliding windows approach to extract the most suitable patterns to find subsequence patterns of weather data. The third is to enhance dynamic pattern detection approach using a sliding window algorithm for weather data segmentation called ESW^{+}, which is using change-point detection proposed to extract the meaningful patterns from weather data that influence to weather changes and pattern discovery algorithm applied for the frequent pattern and sequential patterns of weather data. The performance of the proposed algorithm was evaluated using UCR time series data, and by the Malaysian weather data for rainfall and river flow applications, which were collected from 13 stations for 35 years period. The proposed solutions achieve encouraging performance of pattern prediction and pattern discovery that were supported and trusted from the experts. Experimental results show that the proposed symbolic representation has superior performance in several existing algorithms. The ESAX^{+} Algorithm shows better results as average as 0.426 in terms of error rate based on optimal word and alphabet size of time series data, the algorithm improved the SAX algorithm. In pattern prediction, NB approach was able to generate significant patterns and rules for pattern prediction with superior prediction accuracy up to 79% and was supported by the experts. In pattern discovery, the proposed combined mining algorithms were able to generate higher confidence as high as 70% and, minimum support 0.01 for frequent and sequential patterns. The proposed study has shown its potential in generating algorithms that have the ability to maintain vital knowledge and reduce information loss. Therefore, the weather prediction task has exposed more essential information that can support the decision-making process.

ABSTRAK

Corak penemuan dalam perlombongan cuaca dilaku untuk meramal peristiwa masa depan berdasar pengekstrakan pola yang bermakna dari peristiwa yang diperhatikan sebelum ini. Corak menerangkan tingkah laku kini pada masa tertentu yang diperlukan supaya berguna dan mudah difahami untuk prestasi yang tinggi daripada ramalan. Pelbagai kaedah dilaku untuk memproses ramalan operasi berdasar penemuan corak. Tugasan adalah rumit dan sukar kerana data cuaca adalah peristiwa sepanjang masa yang dikumpul untuk fenomena luar biasa dan mengejut. Masalah data cuaca ialah data dimensi tinggi yang berurusan dalam format mentah iaitu penggunaan memori, bising, kerumitan mencari dan acara besar. Ia diperlukan untuk membangunkan teknik perwakilan yang dapat mengurangkan dimensi data tanpa kehilangan maklumat yang besar. Tujuan kajian ini membentangkan algoritma penemuan corak cuaca berkesan yang merangkumi proses corak relevan dan menarik bagi meningkat keberkesanan ramalan cuaca dengan tiga objektif utama: i) perwakilan siri masa; ii) penemuan corak; iii) ramalan. Pertama adalah meningkatkan algoritma simbolik Agregat Penghampiran (SAX) bagi perwakilan cuaca dipanggil $ESAX^{+}$ yang boleh mengurangkan kematraan dengan maklumat kerugian kurang. Kedua adalah mencadang model peramal berdasarkan algoritma Bayesian Naif (NB) untuk corak ramalan yang disepadu dengan $ESAX^{+}$ perwakilan dan pendekatan tingkap gelongsor bagi mendapatkan corak yang paling sesuai untuk mencari corak data cuaca berikutnya. Ketiga bertujuan meningkat pendekatan pengesanan corak dinamik menggunakan algoritma tetingkap gelongsor untuk pemecahan data cuaca yang dipanggil ESW^{+} yang menggunakan pengesanan titik perubahan yang dicadangkan untuk mengekstrak pola yang bermakna dari data cuaca yang mempengaruhi perubahan iklim dan algoritma penemuan corak yang digunakan untuk corak yang kerap dan corak data cuaca yang berurutan. Prestasi algoritma yang dicadangkan telah dinilai menggunakan UCR siri masa data dan data cuaca di Malaysia untuk adalah hujan dan aliran sungai domain yang dikutip dari stesen yang 13 selama tempoh 35 tahun. Penyelesaian yang dicadang mencapai prestasi yang menggalakkan bagi ramalan corak dan penemuan corak yang disokong dan dipercayai dari pakar. Hasil eksperimen menunjuk perwakilan simbolik yang dicadang mempunyai prestasi lebih baik dalam beberapa algoritma sedia ada. Algoritma $ESAX^{+}$ menunjukkan hasil yang lebih baik sebagai purata nilai 0.426 dari segi kadar ralat berdasarkan perkataan optimum dan abjad yang bersaiz data siri masa seterusnya meningkatkan algoritma SAX. Dalam ramalan corak, pendekatan NB dapat menghasilkan corak dan peraturan penting untuk ramalan corak dengan ketepatan ramalan yang cekap sehingga 79% dan disokong oleh pakar. Dalam penemuan corak, gabungan algoritma perlombongan yang dicadangkan dapat menjana keyakinan yang lebih tinggi sehingga 70% dan sokongan minima sehingga 0.01 untuk corak yang kerap dan berurutan. Kajian yang dicadang telah menunjukkan potensi dalam menghasilkan algoritma yang mempunyai keupayaan untuk mengekalkan pengetahuan penting dan mengurang kehilangan maklumat. Justeru, tugas ramalan cuaca telah mendedahkan lebih banyak maklumat penting yang boleh menyokong proses membuat keputusan.

TABLE OF CONTECTS

Page

DECLARATION **iii**

ACKOWLEDGMENT Error! Bookmark not defined.**iv**

ABSTRACT **v**

ABSTRAK **vi**

TABLE OF CONTECTS **vii**

LIST OF TABLES **xi**

LIST OF FIQURES **xiii**

LIST OF ABBREVIATIONS **xv**

CHAPTER I INTRODUCTION

1.1 Background of The Study 1

1.2 Problem Statement 7

1.3 Research Question 11

1.4 Objective of The Study 12

1.5 Significance of Study and Contribution 12

1.6 Scope of Study 15

1.7 Research Design 16

1.8 Thesis Organization 19

CHAPTER II LITERATURE REVIEW

2.1 Introduction 21

2.2 Time Series Data 21

2.3 Time Series Representation Techniques 25

2.3.1 Piecewise Approximation 25

	2.3.2 Other Methods	28
2.4	Symbolic Representation Techniques	31
2.5	Time Series Prediction	37
	2.5.1 Naive Bayesian Approach	39
	2.5.2 Other Methods	46
2.6	Time Series Pattern Discovery	50
	2.6.1 Pattern Detection	51
	2.6.2 Frequent Patterns	55
	2.6.3 Sequential Patterns	59
2.7	Discussion	64
2.8	Summary	74
CHAPTER III	**RESEARCH METHODOLOGY**	
3.1	Introduction	76
3.2	Research Approach	76
	3.2.1 Phase 1: Identification of the Problem	78
	3.2.2 Phase 2: Data Collection and Pre-processing	79
	3.2.3 Phase 3: Data Representation	82
	3.2.4 Phase 4: Pattern Prediction	85
	3.2.5 Phase 5: Pattern Discovery	85
	3.2.6 Phase 6: Performance Evaluation	88
3.3	Summary	95
CHAPTER IV	**AN ENHANCED SAX PLUS ALGORITHM FOR WEATHER REPRESENTATION**	
4.1	Introduction	96
4.2	Proposed Algorithm	96
4.3	Experimental Design	108
4.4	Experimental Results and Analysis	110
4.5	Discussing the Results	118
4.6	Summary	120
CHAPTER V	**NAIVE BAYESIAN ALGORITHM FOR WEATHER PREDICTION MODEL**	
5.1	Introduction	122
5.2	Naive Bayesian predictor Algorithm	122
	5.2.1 Discrimination of Time Series	124
	5.2.2 Pattern Predictor	125

5.3	Experimental Results and Discussion		127
	5.3.1	Experimental Design	128
	5.3.2	Results Analysis and Discussion	130
	5.3.3	Experts Validation	140
5.4	Conclusion		142
CHAPTER VI	**AN ENHANCED SLIDING WINDOWS ALGORITHM FOR PATTERN DISCOVERY ALGORITHMS**		
6.1	Introduction		144
6.2	Proposed AlgorithmS		144
6.3	Pattern Detection		145
	6.3.1	Enhanced Sliding Window Plus Algorithm	146
	6.3.2	Change Detection	147
	6.3.3	Window Length	148
	6.3.4	Experiments Design	150
	6.3.5	Results and Discussion	151
6.4	Frequent Pattern Algorithm		153
	6.4.1	The Combination Algorithm	154
	6.4.2	Results Analysis and Discussion	157
6.5	Sequential Patterns Algorithm		167
	6.5.1	Results Analysis and Discussion	168
	6.5.2	Discussion	185
	6.5.3	Experts Validation	195
6.6	Summary		197
CHAPTER VII	**CONCLUSION AND FUTURE WORKS**		
7.1	Introduction		199
7.2	Research Conclusion		199
7.3	Contribution of Research		204
7.4	Future Work		207
REFERENCE			**209**
Appendix A	Data Representation for Benchmark Time Series datasets		240
Appendix B	Prediction Results		243
Appendix C	Pattern Discovery Results		244
Appendix D	Weather Data Representation of Pattern Prediction and		

	Discovery algorithms	252
Appendix E	Questionnaire	257
Appendix F	List of Publications	271

LIST OF TABLES

Table No	Table
Table 2.1 Summarization of the notation used in SAX	33
Table 2.2 Summary of time series data representation technique	65
Table 2.3 Summary of symbolic representation techniques based on piecewise approximation	67
Table 2.4 Summary of pattern detection algorithms based on segmentation approaches	72
Table 3.1 UCR Time series data characteristics	79
Table 3.2 Rainfall and river flow time series data characteristics	80
Table 4.1 Segmentation bounds achieved from linear interpolation method	99
Table 4.2 Lookup table containing cut-off points, taken from	106
Table 4.3 Data description of Malaysian rainfall and river flow time series data characteristics	110
Table 4.4 Comparison of error rate obtained by SAX, ESAX and ESAX$^+$	111
Table 4.5 Comparison of ESAX$^+$, ESAX and SAX algorithms depending on error rate	113
Table 4.6 Comparison of SAX, ESAX and ESAX$^+$ algorithms depending on error rate using fitness function F(x)	114
Table 4.7 Alphabet and word size obtained by ESAX$^+$ algorithm	116
Table 4.8 Comparison of ESAX$^+$ with SAX and ESAX and based on p-Values	117
Table 4.9 Summary of representation algorithms on four criteria	119
Table 5.1 Precision evaluation of pattern prediction of daily rainfall time series of 16R – SK. Sg. Lui station with arbitrary size of prediction window and size of alphabet	130
Table 5.2 Precision evaluation of pattern prediction of daily river flow of Sg. Semenyih at Kg. Rinching station with arbitrary size of prediction window and alphabet size.	131
Table 5.3 Precision evaluation of pattern prediction of daily rainfall and river flow of Malaysian weather station with prediction window size 3 and alphabet size 3.	132

Table 5.4 Rule extraction of daily rainfall of Malaysian weather station with prediction window size 3 and alphabet size 3 135

Table 5.5 Rule extraction of river flow of Malaysian weather station with prediction window size 3 and alphabet size 3 138

Table 6.1 Sample river flow station 2 data for 50 days 149

Table 6.2 Number of segmented windows for SW and ESW^{+} 151

Table 6.3 An example rainfall with river flow station 3 patterns and number of observations in January 154

Table 6.4 Frequent patterns of rainfall and river flow in November 157

Table 6.5 Extraction regular patterns from rainfall stations with river flow station3 160

Table 6.6 Extraction irregular patterns from rainfall stations with river flow station3 164

Table 6.7 An example of an experiment with rules in January 170

Table 6.8 An example of the regular patterns of rainfall stations and river flow station 1 171

Table 6.9 An example of the regular patterns of rainfall stations and river flow station 2 174

Table 6.10 An example of the irregular patterns of rainfall stations and river flow station 1 179

Table 6.11 An example of the irregular patterns of rainfall stations and river flow station 2 182

Table 6.12 Average score evaluation of climate change experts 197

LIST OF FIGURES

Figure No	Figure
Figure 1.1 Summary of the research design of the study	17
Figure 2.1 The four classes comprising of the time series data in weather application.	22
Figure 2.2 Time series transformed using SAX symbols, as described in the SAX method	32
Figure 2.3 Some important points (shown in red) are missing	35
Figure 3.1 Research methodology	77
Figure 3.2 The neture of Malaysia data of for rainfall and river flow in fourth season 1978	82
Figure 3.3 Time series segmentation process based on the SW approach	86
Figure 4.1 Time segmentation on 10 points of time series data.	98
Figure 4.2 SAX represents the weather time series data. Some important points (shown in black point) are missing. (Rainfall data of 1 season) The SAX representation is acbbacb.	102
Figure 4.3 $ESAX^{+}$ represents the weather time series data. The $ESAX^{+}$ representation is aaaccacbbcbabaaccbabb.	103
Figure 4.4 Locating the positions of three important values (Top Mean, Mean, Bottom mean). pTm, pM, pBm are their respective positions in the segment.	103
Figure 4.5 Normal probability plot of the cumulative distribution of values from subsequences of length 90 from 10 different datasets. The high linear nature of the plot strongly suggests that the data came from a Gaussian distribution	105
Figure 4.6 The performance of $ESAX^{+}$, ESAX and SAX depending on error rate, (a) for rainfall (Si) and (b) for flow (Fi)	112
Figure 4.7 Evaluation of $ESAX^{+}$, ESAX and SAX depending on error rate, (a) for rainfall (Si) and (b) for flow (Fi) using Fitness function F(x)	114
Figure 5.1 A scheme of the proposed NBp algorithm for analysis of weather time series	124
Figure 5.2 Daily rainfall time series of weather from 1975 to 2009 (IPI,UKM).	128

Figure 5.3 Daily river flow time series of weather from 1975 to 2009 (IPI,UKM) 129

Figure 5.4 The performance of rainfall (Ri) and river flow (Fi) stations base on error rate measurement 133

Figure 5.5 Prediction results of daily rainfall stations with prediction window size 3 and alphabet size 3. 134

Figure 5.6 Prediction results of daily river flow for OrgQ-Sg. Semenyih Kg. Rinching station with prediction window size 3 and alphabet size 3. 138

Figure 6.1 A scheme of weather pattern discovery algorithm 145

Figure 6.2 Number of windows generated by SW and ESW^{+} for five stations 152

Figure 6.3 Approximation errors obtained by SW and ESW^{+}: vertical error (a), mean square error (b), and compression ratio (c). 153

Figure 6.4 Types of temporal relations between two events based on region connection calculus (RCC) 155

Figure 6.5 Map of the Selangor showing the rainfall and river flow stations 169

Figure 6.6 Frequent patterns for rainfall station Ri with river flow station F3 186

Figure 6.7 Frequent patterns for rainfall station 1 with river flow station F3 187

Figure 6.8 The signature patterns between rainfall stations and river flow station 3 189

Figure 6.9 Sequential patterns for rainfall station Ri with river flow station F2 191

Figure 6.10 Map of the four Selangor rainfall stations and river flow station 1 patterns in different time 192

Figure 6.11 Selangor Map of rainfall stations and river flow station 2 patterns in different time 194

Figure 6.12 Evaluation of climate changes experts 196

Figure 6.13 Score of evaluation of climate change experts 196

LIST OF ABBREVIATIONS

ABBREVIATION

C	Time Series Sequence
$\overline{C}$	Aggregated Time series n
$\hat{C}$	Symbolic sequence
n	Time Series Length
N	Reduced Time Series
t	Reduced Time Series
w	word size
a	Alphabet size
pM	Middle position based on ESAX^{+}
E_k	Eding position
pTm	Middle position of values over the average
$pmin$	minimum position
pBm	Middle position of values down the average
$pmax$	Maximum position
$pmid$	Middle position base on ESAX
B_k	Beginning position

CHAPTER I

INTRODUCTION

This chapter introduces in the area of identifying time series prediction and pattern discovery from climate change data. Recommending the pattern detection for weather through time series segmentation algorithm and conceptual symbolic representation is the key goal of this research. Presenting the executive summary of the research would be the main emphasis of this chapter. Moreover, the problem statement, background, objectives, research questions, significance, scope and research methodology of study would also be studied. The brief discussion of the study is found in the 'chapter organization' though which the readers would have the overview of the research study.

1.1 BACKGROUND OF THE STUDY

Discovery relevant and interesting information for historical knowledge is one of the essential tasks in weather prediction. The patterns from weather are describing the behaviour offer at a specific time that required being useful and understandable to get superior performance of prediction. Besides realizing the early warnings prediction of natural disaster occurrences (Li et al. 2011; Zschau & Küppers 2013), these natural trends result in a number of climate change researches, for instance, rainfall and flood prediction (Lee & Liu 2004; Alias 2011; Borga et al. 2014; Yucel et al. 2015; Chai et al. 2017), weather prediction warning system (Isa et al. 2010; Alfieri et al. 2013; de Groot et al. 2015; Breiholz et al. 2017) and agricultural area (De Silva et al. 2014; Keller & Arvidsson 2016). The events pertaining to severe decrease or increase in weather rainfall is expressed by weather prediction, by which the no rain, heavy or moderate rain is to be measured at certain area, or hot, cold or warm temperature is to be measured at specific location (Change 2015; McCrae 2016). According to the

Intergovernmental Panel on Climate Change (IPCC 2007), the conditions of the climate compositions, for example, the river flow, the rainfall amount, the average wind speed and the average temperature are needed by the meteorologists to forecast the weather (Pachauri & Reisinger 2007). Moreover, the Malaysian Meteorology Department (MMD 2012) is fully responsible to predict the weather. For flood warning, the concerned authorities didn't sufficiently calculate the percentage accuracy of the weather prediction.

In Malaysia, the continuous change in weather has also highlighted the need of meaningful patterns for good weather forecasting systems (Saima et al. 2011; Alshareef et al. 2016). The rainfall distribution patterns over the country are determined by the local topographic features together with the seasonal wind flow patterns. The heavy rain spells are experienced by the exposed areas like the Western Sarawak, east coast of Peninsular Malaysia and the northeast coast of Sabah during the northeast monsoon season. According to seasons, it is ideal to define and express the rainfall distribution of the country. The regions of Malaysia has a constant temperature, high humidity, a lot of rainfall in addition to winds, which are weak in the function of time series data with variable regions (Alam et al. 2012; Pachauri et al. 2014). Additionally, the problem in weather pattern discovery and prediction requires researchers to conduct experiments involving multivariate parameters such as temperature, humidity, sunshine, rainfall amount, river flow and water level (Isa et al. 2010; Saima et al. 2011; Dominick et al. 2012; Ghani et al. 2016; Khan et al. 2016; Yaseen et al. 2016). Besides the univariate findings such as, amount of rainfall (Alias 2011; Abdullah et al. 2016; Pineda & Willems 2016), velocity and wind speed (Vafaeipour et al. 2014; Santamaría-Bonfil et al. 2016), a role is played by these dynamic parameters in the uncertainties that prevail in predictions, due to which weather pattern discovery and prediction becomes a challenge (Gundawar et al. 2014). Earlier, a number of time-series statistics were analyzed to make weather predictions, such as, the amount of rainfall and the temperature, where the future weather patterns were interpreted and predicted by meteorologists with the help of their scientific understandings. Nonetheless, many intelligent techniques have been emerged, due to which, researchers are encouraged to be proactive for developing a more accurate weather prediction system. By emphasizing on rainfall and river flow prediction, the

researchers have thus been inspired to investigate the confidence and accuracy of the weather pattern discovery and prediction in Malaysia and this motivation is in reference to the intelligent techniques.

Data mining denotes to the nontrivial extraction of implicit information, was not previously known and could be useful data in databases (Fayyad et al. 1996; Han et al. 2011). With the huge amount of data stored in files, databases and other repositories, developing new advancements to analyze, understand and extract interesting knowledge that can help in decision-making is very important. To prepare time series data, researchers extract the information in such a way that it takes the temporal data. Time series is a sequence of data represents the recorded values of the phenomenon over time. And often include data recorded irregularly for the formation of values at regular intervals, as monthly, weekly, daily or hourly before it can be time-series analysis (Esling & Agon 2012). In addition, it also obtains a time series of recording observations of different types of phenomena, for instance, stock prices, temperature, household income, and heart rate of patient, the number of bits transferred and the product sales over a period of time. To extract meaningful data and its other attributes, time series data is featured in time series analysis (Mitsa 2010). 'Univariate' and 'Multivariate' analyses are the two types of time series analysis methods. In the univariate analysis, only one variable is taken into account for the observation and analysis purpose. While in the multivariate analysis, more than one variable is examined at a time.

There are several important challenges which have to be considered when dealing with time series analysis: the first pre-processing transformation, where time series contain some distortions that are misrepresentation. The task of the pre-processing transformations is to remove different kinds of distortions. Some of the most common pre-processing tasks are: offset translation, amplitude scaling, removing linear trend, removing noise etc. (Keogh & Pazzani 1999). Pre-processing transformations can greatly improve the performance of time-series applications by removing different kinds of distortions. The second time-series representation, which is generally high-dimensional data and a direct dealing with such data in its raw format is very time and memory consuming. Therefore, it is highly desirable to

develop representation techniques that can reduce the dimensionality of time series and third similarity or distance measure, which similarity-based retrieval is used in all a fore mentioned task types of time-series analysis. However, the distance between time series needs to be carefully defined in order to reflect the underlying similarity of these specific data which is based on shapes and patterns. However, time-series representation is the reformulation of the original data in a new form, either numerical or symbolic. A fundamental problem that needs to be solved is the representation of the data and the pre-processing that needs to be applied before the actual data mining phase take place (Mörchen & Ultsch 2005; Meghanathan et al. 2012; Chaudhari et al. 2014). The discretization, reduction, and transformation are the three steps, which are part of this phase. The temporal order of values is overlooked by a number of discretization methods by which numeric time series is converted to symbolic time series (Lin et al. 2003; Dash et al. 2011; Han et al. 2011).

Reducing the dimensions of the data has now become the heart of time-series data mining, which is the first step to deal efficiently with the tasks of data mining huge data (Maimon & Rokach 2005; Wilson 2017). The specific point in a very high dimension is known as the high-dimensional time series and it is realized when the time series dimension contains each time point. A point in space dimensions is compatible with the length of the time series. Consequently, developing the representation techniques, through which the dimensionality of time series can be reduced, is the crucial need of time. To demonstrate time series with condensed dimensionality, researchers have recommended a number of techniques (Faloutsos et al. 1994; Chan & Fu 1999; Cai & Ng 2004; Chen et al. 2007; Han et al. 2011). The Piecewise Aggregate Approximation (PAA) is the common technique for piecewise approximation representation. According to Keogh et al. (2001), each sequence needs to be divided into k segments having equal length. Moreover, they suggested that the average value of each segment may be used as a coordinate of a k-dimensional feature vector. In addition, the Symbolic Aggregate approximation (SAX) is applied for symbolic data representation (Lin et al. 2003). The group of researchers describe that a PAA is obtained at the start and then a time series is discretized. Afterwards, predetermined breakpoints are used to plot the PAA coefficients into SAX symbols. To bring significant improvements in the PAA representation, researchers have

suggested various techniques (Hung & Anh 2008; Karamitopoulos & Evangelidis 2009; Li & Guo 2013).

The main tasks relevant time series include: clustering, prediction, novelty and motif detection and pattern discovery (Han et al. 2011). Time series prediction task aims to build a predictive model for the data to predict explicitly modelling such dependencies variable to predict the next few values of the series based on the values observed previously (Shasha & Zhu 2004; Marroquín et al. 2009; Park et al. 2011). Several techniques have been introduced to handle the task of time series prediction (Nanopoulos et al. 2001; Shah 2012; Kuncheva & Rodríguez 2013; Kenabatho et al. 2015; Wang et al. 2016; Zheng et al. 2016). The Naive Bayesian predictor (NB) is one of the generally used prediction techniques. It is an Artificial Intelligence (AI) approach for solving problems where a model of the domain knowledge is not available, and it displays basic machine learning capabilities. The Bayes' theorem is the foundation for the probabilistic model of naive classifiers, and the hypothesis that "the features in a dataset are equally autonomous" result in the adjective naïve (Raschka 2014). In practice, the independence assumption is often violated, but naive Bayes predictors still tend to perform very well under this unrealistic assumption. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives (Rish 2001). Moreover, this approach is relatively robust, precise, speedy and easy to implement. In various fields, researchers have been performing the time series prediction through this approach (Povinelli et al. 2004; Kim et al. 2006; Chen et al. 2009; Rosen et al. 2011; Lines et al. 2012). Some examples include the diagnosis of diseases and making decisions about treatment processes (Kazmierska & Malicki 2008), the classification of rainfall sequences (Rivero et al. 2013; Gupta & Ghose 2015) and wind speed time series forecasting (Baran 2014) in weather studies.

The patterns discovery in data leading to pragmatic future predictions is known as the predictive analytics. It is a challenging yet interesting task to discover the patterns from time series. Many algorithms have been presented for time series pattern discovery (Bayardo Jr 1998; Han et al. 2000; Wang et al. 2003; Esling & Agon 2012). These algorithms is not appropriate for many applications despite the detected

patterns have specific features and often related to their occur might remain unexplored and hidden interesting patterns, for instance, the behaviours of the temperature and humidity during the days, months or years. The flow of river after heavy rainfall would be another example. This study could also consider the rising rainfall amount may affect the river flow. However, to extract this kind of patterns is that each analysis is related with a time. The notion of the pattern detection of time has been proposed by Lee et al. (2001). It basically aims to identify the repeated patterns or abrupt changes that take place over a specified interval, which is not more than the entire database with respect to the first and last events of each pattern.

Pattern discovery problem, purposed to detect a various type of meaningful patterns from a large time series data has been defined in (Agrawal et al. 1993). Provision of frequent patterns among the variables of large time series data is the key objective, through which the researcher is enabled to find out a relationship 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 and use them to find out more information about natural of the weather. The time series analysis is indicative of this problem in a number of applications and this problem pertains to the extraction of frequent patterns. Concerning weather analysis, this extraction pulls out the sets of regular/irregular patterns through which future events are predicted for human life. These pattern sets become ideal information, if the researchers correctly identify them from the statistical features or domain experts.

In time series, the regular occurrence of serial events is described as a pattern. The facts containing important knowledge that needs to be discovered are known as patterns. The experts can have remarkable perceptions by discovering the frequent patterns in various domains. A number of research studies have comprehensively discussed various frequent pattern algorithms. In this regard, algorithms were proposed by Katoh et al. (2007), (2009) and (2013), so that frequent patterns could be discovered. For discovering the patterns from diamond patterns, the support was developed by the algorithm indicating the bacteria replacement. Furthermore, theoretical study reveals that polynomial space data and polynomial delay data are

required to execute these algorithms in total input size. Additionally, a frequent pattern of time series is relatively easy. The sequential course of events is known as patterns. Moreover, the confidence level is about calculating the number of sliding windows where they take place and we can realize it by calculating the self-reliance of an item set over a list of customer. With high confidence, the pattern is accepted by the current algorithms on frequent patterns and the inputs with low confidence series are simply rejected by them (Katoh et al. 2006; Katoh et al. 2007; de Silva et al. 2009; Katoh et al. 2009; Katoh et al. 2013; Shigezumi & Inakoshi 2014).

The Malaysian river flow and rainfall data acquired from Institute of Climate Change at University Kebangsaan Malaysia (UKM) is emphasized in this study. The experiments in this study would employ the two types of time series dataset: (i) First one using public data sets available through UCR Time Series Data Mining Archive (Keogh 2006). Twenty time series data sets collectively form this data, which derive from different domains. (ii) The other one is the real weather datasets, which comprise of the river flow and rainfall time series data for a period of 35 years, i.e. 1975–2009. Regarding rainfall data, the Malaysian climate change experts divide this data into four main events for rainfall data, which are No rain, Light, Moderate and Heavy. Moreover, their initials, such as N, L, M and H are used to recognize these terms. For river flow data divide into three events, l, m and h are the terms used to indicate low, medium and high flow. Besides discovering sequential and frequent patterns in the river flow and rainfall sequences, this study aims to recommend the prediction and pattern discovery algorithms. In addition, the common patterns for different time periods are found by applying the prediction and pattern discovery of mining tasks.

1.2 PROBLEM STATEMENT

Pattern discovery in weather time series is complex and difficult, since the data is collected events of the over time for unusual and unexpected phenomena particularly in weather prediction. The Malaysian meteorologists are of view that the seasonal wind flow is to influence the river flow and rainfall patterns consistent with the local topographic features (MMD 2012). Forecasting the amount of rainfall and discharge of river flow is one of the challenges in river flow and rainfall prediction. Moreover,

for planning and management of water resources, the operational hydrology area is also affected. The purpose is to deal with flash floods (Alias 2011; Abdullah et al. 2016). Therefore, the long term series of rainfall and river flow data stream besides the prediction accuracy are the areas that need urgent improvement (Ashri et al. 2011; Candela et al. 2014; Wang et al. 2015).

The main major challenges in the weather time series analysis is data reduction because time series are observations made in sequence, the relationship between consecutive data items in a time series gives data analysts the opportunity to reduce the size of the data without substantial loss of information (Zhu 2004; Mörchen & Ultsch 2005; Debnath 2012; Garcia et al. 2013; Chaudhari et al. 2014; Wilson 2017). Symbolic time series is generally required by the knowledge discovery in time series. SAX algorithm based on the PAA has introduced by Lin et al. (2003) and (2007), which is PAA algorithm converts the data into particular values by means of a probability distribution function (Keogh et al. 2001). The word sizes and the alphabets are required by the SAX algorithm as an input, which is one of the key drawbacks, since understanding them from a specific time series data set becomes difficult (Suri & Bailis 2017). The PAA is about mean values approximation. In some time series datasets, there is likelihood that the PAA might miss some important patterns because of the large Euclidean distances. Hence, a good tightness of lower bound might not be generated by the SAX representation (Lkhagva et al. 2006; Hung & Anh 2008; Sun et al. 2014; Zan & Yamana 2016). In the study by Karamitopoulos and Evangelidis (2009) and Cai et al. (2015), noted that PAA approach focus to take into consideration only the central tendency and not concern to the present dispersion in each segments.. For enhancing PAA representation, researchers have suggested a number of techniques (Keogh et al. 2001; Hung & Anh 2008; Karamitopoulos & Evangelidis 2009; Li & Guo 2013; Camerra et al. 2014; Bankó and Abonyi 2015; Lei et al. 2017). Therefore, the enhancement of SAX algorithm is called an enhanced SAX plus ($ESAX^{+}$) for symbolic representation. The aim of this part is to adapt PAA with SAX algorithm to find the most suitable aggregated methods for time series approximation.

Another challenge for time series analyses is time series prediction problem (Diwan et al. 2012). In symbol of time series data analysis, this present study will

introduce models of prediction interval-valuable time series. When interval data is collected in an ordered sequence against time, Symbolic interval time series can be observed in various areas (Maia & De Carvalho 2008). For instance, if weather point of view is considered (the maximum and minimum functions in a given month are to measure the rainfall, maximum and minimum besides the relative humidity in a particular place). For the short term wind speed forecasting, a number of ANNs models were developed by Liu et al. (2013). To forecast the monthly rainfall, a modular type SVM was developed by Lu and Wang (2011). Nonetheless, SVMs and ANNs algorithms are numerically executed, due to which, pulling out significant patterns from large data is a challenging task for future pattern prediction. The weather application is one of the time series application for which these algorithms cannot be applied due to their inappropriateness, since equal size of patterns are only eligible to work with ANNs, while different size of patterns are possessed by a river flow and rainfall applications. The Naïve Bayesian predictor (NB) is one of the commonly used prediction techniques, which contains the prediction of rainfall sequences (Rivero et al. 2013), the decision making about treatment processes and the diagnosis of diseases (Kazmierska & Malicki 2008) and wind speed time series forecasting (Baran 2014) in weather studies. Therefore, NB still has some impairment in term of the data discretization and knowledge base of weather data. The study aims to improve the algorithm for weather problem to extract useful rules and patterns from symbolic time series data, to predict future events based on previously observed values with new strategy to prepare the knowledge base of the experts.

The problem faced by the pattern discovery task included pattern detection based on segmentation approach to detect a various type of significant patterns from a large time series data. The task is to find the number of change points first and identify the class of those points as one window (pattern). The problem need to extract a signal with the best fitting lines and return the end points of the segments as change points or sequence of time points known as a window (pattern) (Epifani et al. 2010; Chandola et al. 2012; Miao et al. 2016). (Keogh et al. 2001; Mueen & Keogh 2010; Alshareef et al. 2016). To discover the surprise patterns and change detection in time series, (Dangendorf et al. 2015) use algorithms for adjusting the data with linear segments such as sliding windows, top down and bottom up approaches (Keogh et al.

2004). Sliding windows algorithm using for segmentation problem, which try to extract temporal data to detect the relationship between the non-trivial patterns and identifying the meaningful patterns and events trends in the data. However, The algorithm has difficult to be ability to discover the differences between two variables of weathers has many potential applications, such as finding the unique behavior, regular or irregular patterns of climate changes (Ahmed et al. 2012; Feng et al. 2012). According to the study of Branch et al. (2013) and Gupta et al. (2014), the change point detection in time series is reduced into anomaly detection. These algorithms cannot be directly applied to multivariate data and some real life applications such as a weather application (rainfall or river flow problem) since in this domain the data is typically nonlinear, non-stationary and the segment might contains more than ten time points and the error threshold is dynamic.

In addition to pattern detection in pattern discovery is finding the relationships among the events of weather data is one main problem of frequent pattern mining algorithms, such as find the association rules that of rainfall that affect to river flow related to frequent pattern for no rain, light, moderate and heavy rainfall associated with low, medium and high river flow among several stations during different rime period. However, in weather events may be difficult to extract the interesting patterns between multivariable time series because they do not indicate annual changes in rainfall through river flow. More sophisticated methods are also used on interval sequences that could be obtained from time series in parallel to Hopper's approaches. A algorithm is proposed by H¨oppner (2002) which mines temporal rules uttered with Allen's (Allen 1983) interval logic and a sliding window to hamper the pattern length. The patterns are mined with an *Apriori* algorithm using support and confidence and ranked by an interestingness measure afterwards. Another interval logics for the temporal rules which led to Region Connection Calculus method (RCC) introduce by Randell et al. (1992) and Cohn et al. (1997). Several research studies have comprehensively discussed the frequent pattern problem. According to Katoh et al. (2007), (2009) and (2013), frequent bipartite patterns from an input sequence without repetition in time can be found through the newly developed algorithms. The approaches were quite helpful in discovering the patterns from diamond patterns, which illustrates the bacteria replacement. From an input event sequence without

duplication, the problem was further discussed to identify all of the frequent *k*-partite patterns (Katoh et al. 2009; Shigezumi & Inakoshi 2014). As per the findings of theoretical analysis, polynomial space data and the polynomial delay data in total input size are required to implement these two algorithms.

The sequential pattern discovery is another pattern discovery problem, which has been discussed in a number of research studies. To beat this challenge of multiple duplicated patterns on time series, system analysts have proposed numerous algorithms (Katoh et al. 2006; Katoh et al. 2007; de Silva et al. 2009; Katoh et al. 2009; Katoh et al. 2013; Shigezumi & Inakoshi 2014). The *Apriori* algorithm has been potentially used to develop sequential pattern discovery (Han & Kamber 2006; Aggarwal 2013). However, applying *Apriori* algorithm for weather data has limitation results in sequential pattern discovery due to temporal issues which are less meaningful pattern discovery; the algorithm needs to be enhanced to handle temporal information for weather data. The lack of temporal patterns can be overcome by this research through enhancement of *Apriori* algorithm that will be able to show the weather pattern over location and time, and identify the meaningful and useful sequential pattern with specific time for weather experts.

1.3 RESEARCH QUESTION

For modelling the prediction and pattern discovery of time-series data, this research aims to recommend suitable algorithms. Various other issues are being handled to overcome this problem, so that the performance of data representation, prediction and pattern discovery could be improved. Given below are the particular research issues:

i. What kind of time series data representation is appropriate for data structure that equivalent to identify key patterns from a large number for weather data?

ii. How the suitable strategy to represent data structure and the current algorithm of time series prediction can work ideally for pattern prediction?

iii. How to discover frequent and sequential weather patterns and identify a suitable pattern detection algorithm can work for time series data to summarize important patterns from a huge number of patterns that can be easily interpreted by the domain experts?

1.4 OBJECTIVE OF THE STUDY

This research intends to demonstrate that the univariate and multivariate analysis on high-frequency data can be successful in grouping similar climate changes with the help of machine learning techniques. In addition, the study aims to establish a suitable symbolic representation through which makes it easy to store and analyze the data, with the goal of eventually developing time series representation methods using Malaysian rainfall and river flow datasets for pattern prediction and pattern discovery. In particular, following are the objectives of this study:

i. To propose an improved SAX algorithm for symbolic time series representation of weather data known as $ESAX^{+}$.

ii. To propose an integrated NB predictor algorithm with $ESAX^{+}$ symbolic representation for weather pattern prediction rules that can drive weather experts in better decision making.

iii. To propose a pattern detection algorithm for detecting subsequence of the sequences in weather data (known as ESW^{+}), the algorithm would apply frequent patterns and sequential patterns by an integrated Apriori algorithm process through interval logic of time series rules based on RCC relations for weather pattern discovery algorithms.

1.5 SIGNIFICANCE OF STUDY AND CONTRIBUTION

The prediction and pattern discovery procedures are the major elements of this study, through which the patterns of events knowledge are effectively detected by the researchers. The aim of this overall process is to improve climate changes decisions

from massive weather data available in the weather patterns. The statistical data is highly focused by most of the weather application. From the weather data, the detection of patterns and events make this research worthwhile. From this perspective, researchers detected and highlighted any exceptional or unforeseen weather reporting as a possible new knowledge. The weather analysis areas were the beneficiaries of this knowledge and its extraction.

To accomplish all the goals of pattern discovery and prediction, a number of methods are integrated by the recommended algorithms. By considering the time series representation based information extraction approach, the most appropriate patterns were extracted, which aimed to reduce the high dimensionality problem of processing weather data. The algorithms for time series representation, known as an enhanced SAX plus ($ESAX^{+}$) were proposed by this research for capturing the significant patterns. The findings reveal that representation accuracy can be improved by the $ESAX^{+}$ due to a better tightness of lower bound than the original PAA, whose entire focus was on the centroid area. Adjusting the PAA with SAX algorithm is the basic goal of this part so that the most suitable aggregated methods could be discovered for time series approximation. An efficient mining process can be ensured by the proposed algorithm whereby the researchers can obtain a better knowledge model with no major loss. According to the outcomes of the experiments, performance of the proposed algorithms is much better than those of compared methods for high dimensionality task.

Prediction algorithms for time series data is another challenge. Moreover, this research is believed to be of great consequence in many applications. In this regard, the Malaysian weather data has been focused using rainfall and river flow time series data sets, through which future events can be forecasted in an efficient and effective way. When accuracy is to be balanced with efficiency, this becomes a challenge because it is computationally expensive to implement the highly accurate mining techniques. The segment rainfall and river flow data would observe the main algorithm together with its implementation. The developers have improved the segmentation algorithm with respect to threshold value in this study. Instead of a fixed value, this is assumed to be dynamic and in terms of number of time points in a segment, the level has been raised, which is now more than two points in a segment

that went up to 20 points in one segment. The rainfall and river flow data sets are utilized to evaluate the success of the algorithms. Therefore the pattern detection methods are faster in terms of performance besides having least level of complication. Consequently, the large datasets are found applicable for implementation of the proposed methods.

Moreover, using prediction methods and patterns discovery techniques, applying the symbolic processed rainfall and river flow data to the sequential patterns algorithm and the frequent patterns algorithm is the key aim of mining tasks. Initially, applying the proposed algorithm for generating a number of patterns besides creating the length of each pattern is the main contribution in this particular phase. A high frequency of repeated patterns could be detected by the frequent patterns algorithm. Secondly, the relationship between river flow and the rainfall at different stations in different occasions is extracted by applying the sequential pattern algorithm. Valuable patterns might be demonstrated by those algorithms, which could be précised as rules for the experts of Institute of Climate Change, Malaysian. The application of such algorithms is not restricted in other fields, although the researchers implement them in the weather domain.

The researchers would be enabled to analyze weather data as a result of this study. As far as most computer science problems are concerned, the effective and efficient solutions can be delivered with the ideal representation of the time series. The symbolic representation of time series is corroborated by the research. An efficient mining process can be resultantly ensured, whereby we can acquire a better knowledge with no major loss. The identification and extraction of patterns are emphasized in the intentional approach, which basically influence weather time series due to which a flexible pattern of time-series matching is eventually created to extract the distinct patterns, by which the researchers are likely to realize the benefits of the patterns discovery and prediction. The rainfall and rives flow data sets are used to measure the success of the algorithms. The storage, computation and transmission of the data are efficiently manipulated by the algorithms. Moreover, they develop useful patterns, through which a researcher is enabled to execute it in the field of weather prediction.

1.6 SCOPE OF STUDY

The scope of the proposed research can be summarised in the following points:

i. The proposed UCR time series classification is benchmarked data comprises of 20 data sets (Keogh 2006; Li & Guo 2013), and Malaysian weather data are basically the rainfall and river flow represent as two variables of time series data, used in this research study . For a period of 35 years, i.e. 1975–2009, a collection of data sets from Institute of Climate Change, University Kebangsaan Malaysia, Malaysia. Overall, different stations are to generate the rainfall and river flow pattern. Since, pattern discovery in time series data is the prime focus of the study, two variables would be used to select the weather data. Other data could include the water level, temperature data and humidity data which usually exist in weather data, which are very close relations between those data.

ii. This research focuses on the time series representation of weather data by employing the main high dimensionality of Malaysian rainfall and river flow variables. The proposed $ESAX^{+}$ algorithm is to adapt SAX algorithm to find the most suitable aggregated methods for time series approximation. The proposed algorithms are smoothed average assigns three average locations to the more recent data. Thus, it is a top mean, mean and bottom mean. The approaches are also tested using the benchmarked data.

iii. This research also focuses on pattern prediction based on rainfall and river flow variables using the NB predictor algorithms. NB predictor is used to improve weather pattern prediction rules bases on the $ESAX^{+}$ task.

iv. The research address the issue of segmentation of multivariate weather data in Malaysia rainfall and river flow data, and therefore the pattern discovery phase is out of the scope of this research.

v. The research also focuses frequent patterns algorithm to extract the patterns from rainfall and river flow sequences by integrated RCC relations approach for interval logics for the temporal rules to produce the width of the patterns, and the frequent pattern algorithm is applies to extract the maximum frequent patterns in different time periods.

vi. The sequential pattern algorithm has been suggested in this research to discover the interesting patterns of associated rainfall and river flow variables, locations based on regular and irregular pattern discovery.

1.7 RESEARCH DESIGN

The stated earlier in this chapter, key goal of this research is to implement a pattern discovery algorithm for weather problem. The data representation, pattern prediction, pattern detection and pattern discovery are handled by this approach. The researchers implement the high-quality dimensions based on data discretization approach could be symbolized the time series data. In time series sequence, the most frequent pattern is found in the pattern prediction on the basis of probability method through which the best prediction is generated in future. Furthermore, to discover hidden and useful patterns, researchers tend to apply the pattern detection, frequent pattern and sequential pattern algorithms on time series data. The segmentation algorithm is applied in this part for weather data so that the number of change points in time series could be identified besides discovering the class of those points on the basis of change point detection. To achieve the objectives of this thesis and to answer the research questions, this study adopts an experimental for objective, the pattern discovery algorithms is developed and evaluated using experimental Malaysia rainfall and river flow datasets with suitable performance criteria. The methodology of this study includes four main phases as shown in Figure 1.1; theoretical study, research architecture, experiments design, and experimental results and analysis.

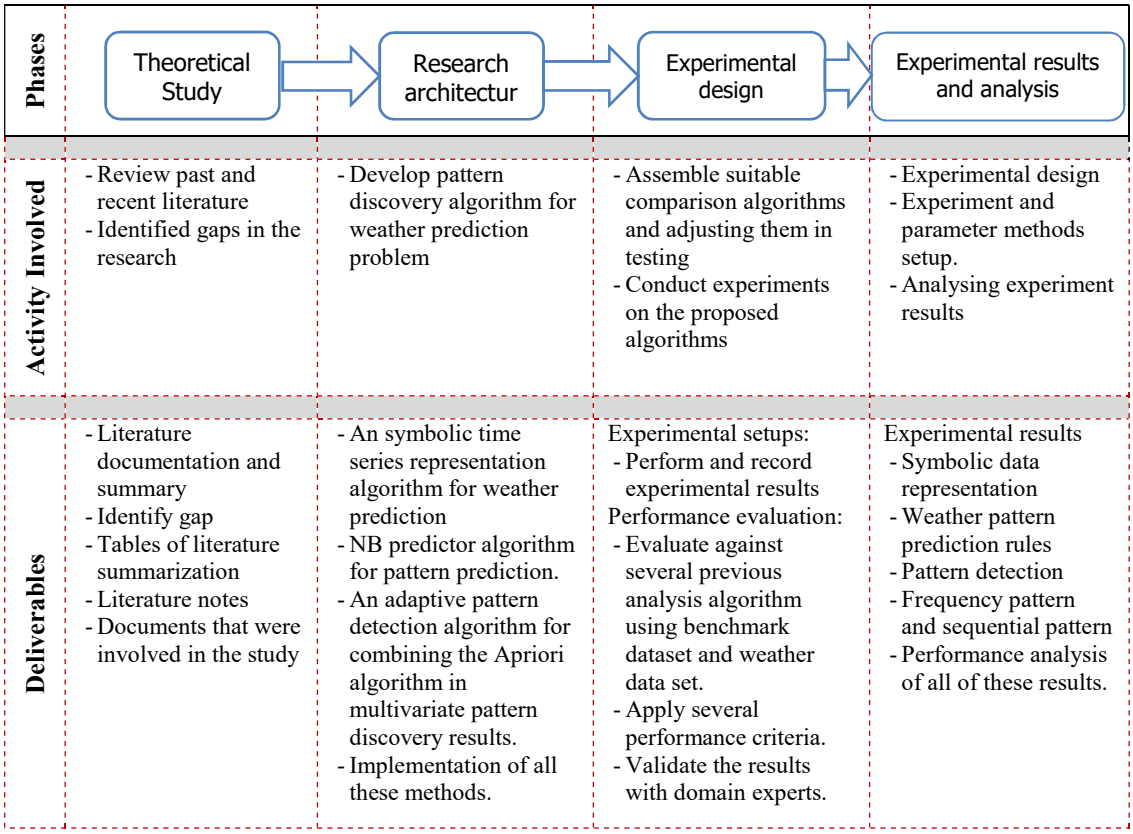

Figure 1.1 Summary of the research design of the study

The theoretical study phase identifies the thesis objectives and scope by examining the most relevant state-of-the-art research problems and challenges that require further improvement. The literature on the identified problems was reviewed and analysed. Based on this analysis, three main subjects were reviewed further as discussed in Chapter II. The first subject is time series analysis and data representation to identify the shortcomings and criteria for improvements, researchers would review the related work in the weather area of symbolic representation, the second is concerned in pattern prediction and the third is concerned with current and past studies in the field of pattern discovery algorithms. The findings of this phase are tabulated and summarised to easily identify the gap that exists in time series prediction algorithms. The details of this preliminary study are discussed in Chapter II. The next phase gives the research architecture, design and implementation.

Based on the outcomes of the theoretical study, the next phase looks at the research architecture, which is concerned with the design and implementation of the

solutions to the previously determined issues of the study. The development process starts with proposing an ESAX$^+$ algorithm for symbolic representation based on the main tissue components that appear in the time series data, as discussed in Chapter IV. Another proposed scheme is a pattern prediction algorithm that addresses the univariate problem in weather data is proposed whereby the NB predictor algorithm is integrated with the ESAX$^+$ algorithm filter algorithm for the symbolic domain, as presented in Chapter V. Finally, a pattern discovery algorithm is implemented based on pattern detection algorithm using time series segmentation approach. The frequent pattern algorithm is proposed to extract the maximum frequent patterns in different time periods by integrated RCC relations for interval logics for the temporal rules to generate the width of the patterns. Then, the *Apriori* algorithm takes into consideration complementarity of the sequential pattern discovery along with the regular and irregular patterns of associated rainfall and river flow variables, locations and time. These patterns discovery algorithms will be deeply examined in Chapter VI.

In the third phase, numerous experiments using the developed weather pattern prediction and pattern discovery algorithms and other suitable comparator algorithms were designed and executed to assess the performance of the proposed methods through this experiment. Prior to start the experiments, the group of researchers defined an experimental setup. The UCR time series data, i.e., benchmark data, was selected to perform the experiment (Keogh 2006). The 20 time series data sets originating from different domains collectively form the data. In addition, weather datasets pertain to river flow and rainfall data acquired, for a period of 35 years, i.e. 1975–2009, from Institute of Climate Change, University Kebangsaan Malaysia (UKM), Malaysia.

The final phase of the research design is the experimental results and analysis phase, whereby the outcome of the experiments is evaluated based on the benchmark time series data and self-collected Malaysian weather datasets. The standard metrics of evaluation in this field of study have been used to evaluate the proposed algorithms. The performance of the proposed time series representation strategy, pattern prediction approach and pattern discovery algorithm are evaluated based on the impact of the proposed algorithms on the performance of the information loss,

accuracy, confidence, support and experts validation. The performance of the proposed algorithms is benchmarked against suitable comparator algorithms selected from the most relevant literature.

1.8 THESIS ORGANIZATION

This thesis contains seven chapters, including the existing chapter containing the introduction. The overview of the research and the background information is given in the "Introduction part". Other than this, there are six more chapters in this thesis, which are given as follows:

In time series data mining, the literature review of the connected studies is given in Chapter II. Before indicating important definitions of time series mining methods, a brief review of the literature on knowledge discovery and data mining is to begin this chapter. The time series data representation methods, pattern discovery methods and prediction methods are highly focused by the discussion.

Besides the requisite techniques, the variety of concepts and the methodology of the research studies are demonstrated in Chapter III. Moreover, the gap in the previous work and the unresolved issues discussed in Chapter II were highlighted to bring improvement in the existing study. With no major loss of data, the data reduction is included among these issues. Moreover, according to selection of the SAX parameters for the data representation phase, the subsequence of the time series sequence realizes the benefits, for example, in the river flow and rainfall applications, river flow data can be applied to find the time, when the sequence of 'high' flow or 'low' or 'heavy' change occurs. In addition, rainfall data is used to detect the 'no rain' scenario and it is taken as a pattern detection problem. Afterwards, a proposal is presented by the Chapter III for the solution of above-mentioned issues. Moreover, certain procedures were also suggested, and the research objectives were accomplished through the execution of those procedures. The authors presented the hypothetical study together with the evaluation measurement and the experimental design, which are then chosen and applied so that the proposed methods could be evaluated.

The first phase of the experimental design is described in Chapter IV containing the details of the proposed, ESAX^{+} algorithm for symbolic representation and the data representation of time series. Afterwards, the research group discussed and analyzed the results.

The proposed time series prediction algorithm is introduced in Chapter V. Keeping in view the produced symbolic representation data sets, the author presented the proposed method in this chapter. The definitions, concepts and NB predictor are to start this chapter. The NB predictor is an essential element in the prediction algorithms offered in this study. Subsequently, the author introduced the fundamental concept of the similarity method. Afterwards, the algorithms for implementing NB were discussed by the authors and these would work as the time series prediction algorithm. This Chapter illustrates an example, which demonstrates how a similar sequence is detected by the NB from the given symbolic representation. Subsequently, the chapter explains how the adjustment was made in prediction algorithm so that the symbolic representation for comparison purposes could be recognized. Afterwards, the author presented the findings of the comparison and finally a summary concluded the chapter.

For pattern detection, frequent and sequential patterns from weather data sets, the algorithms used in segmentation approach are presented by Chapter VI.

Eventually, the findings of this research are given in Chapter VII. Moreover, the research illustrates the main contributions and its significance to the data mining community. Future suggestions are also described by this chapter. Moreover, weather community for future investigation is summarized with respect to pattern discovery and weather prediction.

The background required for the existing research is presented by this chapter. Moreover, the objectives, problem statement, and scope of the research work are also described by this chapter. Besides the brief summary of thesis outline, this study explains an illustration of research significance, adopted methodology and major contributions.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

A considerable interest in weather mining is being demonstrated during the past decades. Gaining an insight with the dynamics of weather is a crucial task for establishing weather system. The researcher would thus be enabled to choose and evaluate the desirable technique for onwards procedure. A review of the related work and background of the study was introduced by this chapter. Besides offering representation methods for pattern discovery and prediction in Malaysian weather data, different time series data and analysis was provided by the literature. Because of the time series representation study, the problem is illustrated in symbolic representation and the methodology applied by other researchers. Subsequently, the study addressed the patterns detection time series method and the associated problems. Afterwards, the author discussed the sequential pattern and frequency pattern the frequency pattern algorithms. Next to it, the chapter discussed the research issues and problems. The constraints and the identified summary would conclude the chapter. Afterwards, the feasible solutions of the issue are presented.

2.2 TIME SERIES DATA

A sequence of event values which take place during a period of time can be referred to as a time series. A value is associated with each event occurring at each time point and this value is recorded as well. A single variable (for instance, a river flow or rainfall signal over a time period) is represented by the collection of all these values. Hence, a sequence of recorded observations of an interesting event is included in a time series of a single variable. In this thesis, both the univariate and multivariate time series pattern discovery and prediction problem are highlighted by this research. The time

series pattern discovery and prediction have a number of terms and concepts which are used to define them (Baydogan 2012); however they should be defined internationally.

A time series is a set of observations measured consecutively through time is known as a time series (Chatfield 2001). A time-series containing single observations recorded sequentially through time is referred to as the Univariate time-series (UTS), e.g. the monthly rainfall, humidity or temperature. An ordered set of T values is given as UTS, $c^n = (c_1^n, c_2^n, ..., c_T^n)$. It is supposed that the time series through index 't' are measured at equally-spaced time points. For n = 1, 2, …, N and $a^n \in \{0,1,\ 2,\ \ldots,\ B-1\}$, a class a^n is associated with each time series. Figure 2.1 demonstrates four time series of each class from a four-class time series classification problem, i.e., $(T = 90, B = 3, q^n \in \{0,\ 1,\ 2\})$. The frequency of events, similarity distance for certain intervals or number of peaks may describe the time series from class q^n.

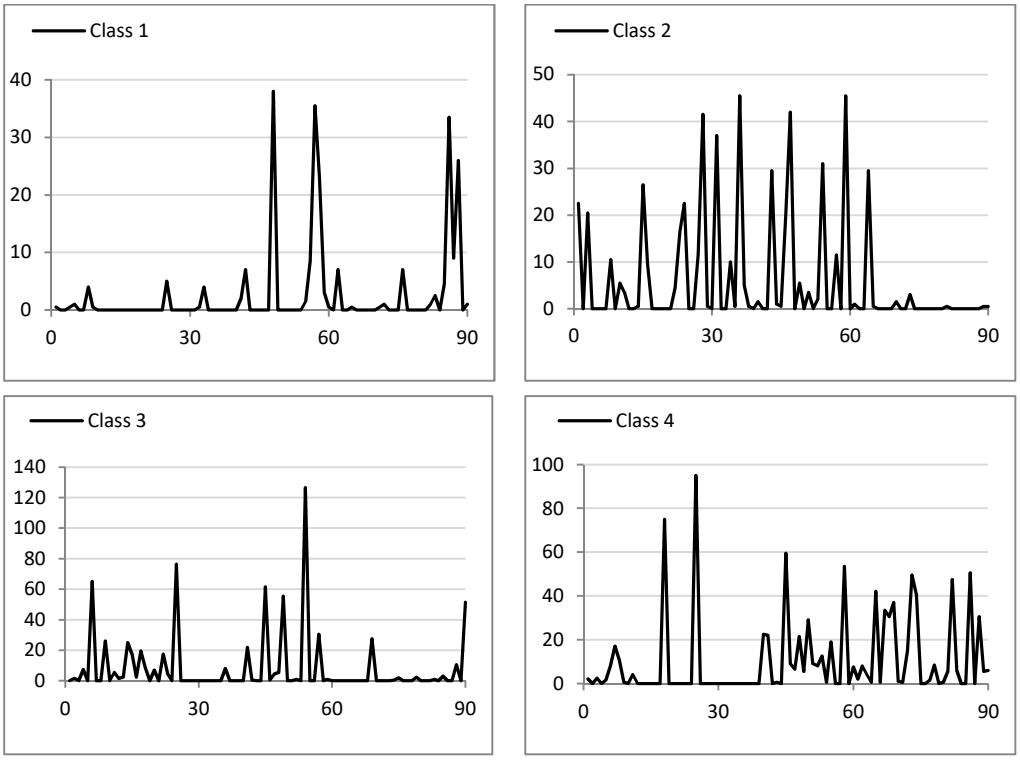

Figure 2.1 The four classes comprising of the time series data in weather application.

Multivariate time series (MTS) C^n contains M univariate time series, whereby, each one has T observations where $c_m^n(t)$ depicts the observation at time t from variable m of MTS n. The '$T \times M$' matrix represents the MTS C^n, which is given as:

$$c^n = [c_1^n, c_2^n, \ldots, c_m^n, \ldots, c_M^n]$$

Where

$$c^n = [c_1^n(1), c_2^n(2), \ldots, c_M^n(T)]$$

We notice the N training MTS, each of which is linked with a class a^n, for n = 1, 2, . . . ,N and $a^n \in \{0,1, 2, \ldots, B-1\}$. A special case of MTS is known as the univariate time series, where value of M is one.

There are several important challenges which have to be considered when dealing with time series analysis: the first pre-processing transformation, where time series contain some distortions that is misrepresentation, which could be consequences of bad measurements or just a property of underlying process which generated time series. The presence of distortions can seriously deteriorate the indexing problem because the distance between two time series could be very large although their overall shape is very similar. The task of the pre-processing transformations is to remove different kinds of distortions. Some of the most common pre-processing tasks are: offset translation, amplitude scaling, removing linear trend, removing noise etc. (Keogh & Pazzani 1999). Pre-processing transformations can greatly improve the performance of time-series applications by removing different kinds of distortions. The second time-series representation, which is generally high-dimensional data and a direct dealing with such data in its raw format is very time and memory consuming. Therefore, it is highly desirable to develop representation techniques that can reduce the dimensionality of time series. In addition, time series representation transform the time series to another dimensionality reduced vector $\overline{C}_i = (\overline{c_1}, \overline{c_2}, \ldots, \overline{c_x})$ where $x<T$. Moreover, if two series within the original space are similar, then their representations

in the transformation space should also be similar. The third consideration is similarity or distance measure, which similarity-based retrieval is used in all a fore mentioned task types of time-series analysis. However, the distance between time series needs to be carefully defined in order to reflect the underlying similarity of these specific data which is based on shapes and patterns. However, time series distance defines the distance between two time series across all time points, then the equation: $dist(C_i, C_j) = \sum_{n=1}^{T} dist(c_{in}, c_{jn})$ is the summation of the distance between individual points.

There are an explosion of interest within data mining tasks relevant time series include query by content (indexing) to find the most similar time series in a database for a given query time series (Faloutsos et al. 1994; Aref et al. 2004; Chen et al. 2007), finding groups of the time series in a database through clustering in such a way that there is similarity between the time series of the same group while the time series of different groups show distinction from one another (Kalpakis et al. 2001; Keogh & Lin 2005; Da & De 2012), classification to set a time series given to a group pre-defined in a way that is more similar to other time series of the same group than it is for the time series of other groups (Bakshi & Stephanopoulos 1994; Wei & Keogh 2006; Xi et al. 2006; Lotte et al. 2007), anticipating goal for the prediction of clearly modeling; with the help of these dependencies variable the next few values of the series can also be predicted (Chatfield & Weigend 1994; Shasha & Zhu 2004; Shah 2012; Goerg 2013; Kattan et al. 2015), time series of sections showing behaviors different than the expectations in relation to some base model can be detected through novelty (Weiss 2004; Izakian & Pedrycz 2014), detection of unknown repeated patterns in the time-series database by motif discovery (Lin et al. 2004; Li et al. 2012; Begum & Keogh 2014), creation of an accurate approximation of time series through segmentation aims by reduction in dimensions while the basic features are kept as they were (Chung et al. 2001; Keogh et al. 2004; Ogras & Ferhatosmanoglu 2006; Dobos & Abonyi 2012; Guo et al. 2015) and rule discovery is employed for inference, according to this rule a few time series which describe the behavior may be used to predict the behavior of similar time series at a specific point in time (Das et al. 1998; Wang & Chan 2006; Lai et al. 2009).

2.3 TIME SERIES REPRESENTATION TECHNIQUES

Time series representation of the data has now become the primary step of time series analysis, which deals with huge data for efficiently mining tasks. the analysis of such intriguing data might reveal patterns deserving further attention (Montañés et al. 2011). Dimensionality reduction approach refer to contractive property of their transform and feature-space distance choice defined by lower bounding explains that the distance in the feature space essentially underestimates the real distance between time series. Several methods have been proposed that focus on overcoming dimensionality reduction techniques challenges that can be employed for time series representation (Bakar et al. 2010). In (Faloutsos et al. 1994; Chan & Fu 1999; Keogh et al. 2001; Chakrabarti et al. 2002; Lin et al. 2003; Lkhagva et al. 2006; Chen et al. 2007; Shieh & Keogh 2008) have been proposed for dimensionality reduction to represent time series while maintaining the basic features for similarity computation and mining operations. Since time series conclusions are made in sequence, the association among sequential data items in a time series enables data analysts to minimize the dimensions of the data, without significant loss of information (Zhu 2004). In the following subsections present review the main dimensionality reduction techniques that preserve the lower bounding property.

2.3.1 Piecewise Approximation

One of the most significant methods used in several effective similarity researches on time series depends is Piecewise Aggregate Approximation (PAA), which was introduced by Yi and Faloutsos (2000) and Keogh et al. (2001). In The main concept of PAA is to partition every sequence into k segments of similar size and to utilize the average of every section to represent the latter. PAA proposed to divide each sequence into k segments of equal length and to use the average of each segment as a coordinate of a k-dimensional feature vector. In addition, the straight lines are used to approximate the time series instead of constant e.g. Piecewise Linear Approximation (PLA), the PLA tends to closely align the endpoint of consecutive segments, giving the piecewise approximation with connected line (Shatkay & Zdonik 1996; Chen et al. 2007).

The following mechanism can be used to understand the concept of dimensionality reduction. Suppose a time series question where, $C = c_1, c_2, ..., c_n$, consider N as the variable of dimensionality of the transformed area that needs to be indexed ($1 \leq N \leq n$). It will be easier for us to understand the phenomenon if we suppose that N is a factor of n. A time series X of length n is denoted in N space by a vector $\overline{C} = \overline{c_1}, \overline{c_2}, ..., \overline{c_N}$, where $\overline{C}$ is calculated using the equation given below:

$$\overline{c}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} c_j \tag{2.1}$$

PAA method, irrespective of present dispersion in for each segment, considers the central tendency only (Karamitopoulos & Evangelidis 2009). As a result, Keogh et al. (2001) suggested the use of an adaptive PCA, which unlike PAA has elements of varying lengths and yields a highly operational compression. There are two major differences between the PAA and APCA techniques. The major difference is that in PAA, all segments have equal lengths, whilst in APCA, the segment lengths are allowed to vary. The other difference is that the APCA method uses the discrete wavelet transformation to segment the time series, while in the PAA method, the time series is directly partitioned into equal-length segments. The APCA was teste and compared against other methods and the results declared its high quality and various advantages that are discussed below:

- The comprehension and application of this process is extremely easy. The resultant distant measures (such as the weighted Euclidean queries and the index that can be fashioned in linear time) of this approach are flexible.
- It offers effective solutions for random length queries and unceasing time inclusions and exclusions. It enables the linear time construction of the index.
- A weighted Euclidean distance can be used for the representation, where each single section of the series possesses different weight.

- The methods offers remarkable dimensionality reduction, while PAA minimizes dimensionality by the mean values of equal sized frames. The resultant representation of the mean value may result in omission of important patterns in certain datasets of time series (Lkhagva et al. 2006; Hung & Anh 2008).

Hung and Anh (2008), devised a technique for dimensionality reduction, which is known as Piecewise Linear Aggregate Approximation or PLAA. This technique is known for yielding most effective representation of time series data. In order to yield matching scaled segments, this technique uses the mean value in addition to another value that happens to be the slope of the most fitting straight line and equals the data points in the section. When we look at the findings of three different experiments that used PLAA and PAA improved trimming capacity, implementation systems and higher stability of lower bound. The statistical evidence on this technique highlights that PLAA is a much better choice than PAA. The PLAA shows higher stability of reduced bound the resultant representation of various datasets is extremely accurate.

Dispersion based PAA (DPAA) is another type of extended PAA representation and a measure of distance used for lower bounds. This technique was proposed by Karamitopoulos & Evangelidis (2009). This technique divides the time series into sequences of same-length segments. The resulting mean and standard deviation, which are also similar, are then noted. Two variables are involved in DPAA has two variables. Variable 1, Weight is denoted by w and variable 2, the number of segments into which the original time series is divided is denoted by *k*. The test values range from 0 to 1 and can grow up to 0.1. The analyses of DPAA implement by 1-NN classification on 20 real world and synthetic datasets. According to the findings of the analyses the resultant representation superseded the general representations that are a product of massive datasets. The central propensity acts as the main difference between both the approaches.

The ASCC technique acronymic of average, slope, curvature and rate of change of the curvature is driven by the concept of four shape features. The ASCC technique was also proposed by Li & Guo (2013). The major shape features of the

time series play an essential role in both stages of the model, where time series data is divided into same-length sequences. The technique possesses exceptional tightness and cutting power of the lower bound. As far as the original time series is concerned the feature sequences offer extensive information hence the chances of fitting errors are reduced. When the number of segmented subsequences is fixed, the resultant fit to the original time series yields minimal error. As far as data mining is concerned, the outcomes are precise and there is consistent rise in the time complexity. The unique functions of distance measure offer lower bound on the Euclidean distance.

One of the most recent additions to these techniques is the Symbolic Aggregate Approximation SAX method, which was devised by Lin et al. (2003). This technique is governed by the Piecewise Aggregate Approximation or PAA guidelines where dimensionality reduction is crucial. This technique uses alphabets and word sizes as inputs, which are difficult to locate and differentiate in certain time series data sets. Hence it's not that effective. The discretization is a special method that uses advanced representation of unique time series and the symbolic patterns. The procedure is carried out by converting the data into the PAA representation, which are then shifted to PAA representation in unique patterns. The PAA values are converted into distinctive values using a probability distribution function. The SAX algorithm needs the alphabet and word sizes as inputs, which is the major drawback, since it is not apparent to figure out them from a specific time series data set (Keogh et al. 2001; Yi and Faloutsos 2000).

2.3.2 Other Methods

The first approach of time series representation presented by Agrawal et al. (1993) employed Discrete Fourier Transformation (DFT) for dimensionality reduction for similarity search. The fundamental idea of spectral decomposition is that each signal, each wave is represented by a single complex number known as a Fourier coefficient. DFT is utilized to convert a series from the time domain to a point in the frequency domain. Selecting the *N* first frequencies and later representing each series as a point in the *N*-dimensional space accomplishes this goal. DFT has the attractive property that the amplitude of the Fourier coefficients is invariant under shifts, which allows

extending the method to find similar sequences ignoring shifts (Fu et al. 2008). Furthermore, it is a good ability to compress signals and it is the most natural, fast indexing algorithm.

The Discrete Wavelet Transform (DWT) is a technique of linear transformation that breaks down the unique sequence into different frequency components, while retaining the information regarding the element's sudden occurrence. The resultant sequence is deemed as the wavelet coefficients that represent features of the event. The sequence can be represented by using a few coefficients. This aspect of time series representation increased its manageability and also improved its application to conjoint data mining operations. A contemporary trend is to use DWT for retranslating the sequence from the time domain into the frequency or time domain (Chan & Fu 1999). However, the Fourier transform is good for a spectral analysis or which frequency components occurred in signal, it will not give information about at which time it happens. Therefore, the wavelet transform is suitable for the time-frequency analysis. It is also good for signal demonising, but of course it has some disadvantages.

Singular Value Decomposition (SVD) was first effectively used indexing images and other media objects, and has recently been proposed for time series indexing (Korn et al. 1997). SVD is similar to DFT and DWT, accounting for the way in terms of a linear combination of the base (in this case, they called their own waves). However, the SVD is different from the DFT and DWT in a very important way, that is, DFT and DWT are local, but a data object at a time to investigate and apply a transformation. These transformations are completely independent of other data. However, the SVD is a global transformation. In SVD, the data is analysed and rotated so that the first axis of variation is possible, these second axis is perpendicular to the maximum possible variance first, and the third axis is perpendicular to the maximum possible variance for the first two. This global transformation is both a weakness and strength in terms of indexing.

Another technique works on the principle of bit level approximation of the data and is named Clipped representation (Clipping) introduced by Bagnall et al.

(2006). This approach of clipped series representation has many advantages: it allows raw data to be directly compared to the reduced representation, while still guaranteeing lower bounds to either Euclidean Distance or DTW. The Clipping representation can outperform by a few orders of magnitude. The new technique can improve the compression ratio by a wide margin, while being able to maintain or increase the tightness of its lower bound, which allows even faster nearest neighbour queries, especially in ones that require Dynamic Time Warping distance measure. Other than producing, faster exact algorithms for similarity search. The new clipped series representation approach can support time series clustering and scale to much larger datasets. These results demonstrate unequivocally that clipping is a useful transformation for time series data mining based on similarity in structure.

Another approach for time series representation is developed by Fu et al. (2008). The representation idea that a time series is constructed by a sequence of data points and the amplitude of a data point has different extent of influence on the shape of the time series, each data point has its own importance to the time series, it may contribute on the overall shape of the time series or it may have little influence on the time series or may even be discarded. The data point with importance calculation is named as perceptually important point (PIP). Then, they introduced three methods for evaluating the importance of the PIPs in a time series, they are: Euclidean distance (PIP-ED), perpendicular distance (PIP-PD) and vertical distance (PIP-VD). Experiments show that PIP-VD is a preferable method for evaluating the data point importance in most of the cases in financial domain. Then, a novel of time series representation used tree structure is called specialized binary (SB) tree and has been developed based on binary tree structure. The SB-Tree supports a fast lookup of time series starting from the most important data point. A new updating method and two dimensionality reduction approaches are proposed by Park et al. (2010), they presented representation and clustering of time series by means of segmentation based on PIPs detection. PIP detection technique has been proposed to represent the movement curve of time series using a small number of PIPs called the inflection points of time series (Chung et al. 2001).

2.4 SYMBOLIC REPRESENTATION TECHNIQUES

One important problem of time series analysis is time series discretization algorithms that divide into two main categories: unsupervised, which discretize attributes without taking into account class information, and supervised, which discretize attributes while using interdependence between the known class labels and the attribute values (Meghanathan et al. 2012). Discretization is an important pre-processing step in the knowledge discovery of making raw time series data applies to a symbol of data mining algorithms. To improve the under stainability of the mined results, or to assist the induction step of the mining algorithms, in discretized, it is natural to prefer the discrete levels that can be mapped into easy characters. Knowledge Discovery in time series usually requires symbolic time series. Many discretization methods that convert numeric time series to symbolic time series ignore the temporal order of values (Mörchen & Ultsch 2005).

Discretization is concerned with the process of mapping variables with continuous values into discrete values. This process has widely used to compress data to facilitate computation in terms of space and time. The discretization divided into two main sections. The first one is to find the number of discrete groups to do the mapping from continuous to discrete. The second one is to define the range or limits of each interval in the continuous domain (Chen et al. 2005). Two common methods used in most of the applications are Equal Width Discretization (EWD) and Equal Frequency Discretization (EFD) (Han et al. 2011). Other than these are *k*-means Clustering (Dash et al. 2011; Han et al. 2011), SAX (Lin et al. 2003) and Frequency Dynamic Interval (FDIC) (Ahmed et al. 2009) for discrete time series are required to state a set of parameters for processing.

The Symbolic Aggregate Approximation (SAX) method was introduced by Lin et al. (2007) based on Lin et al. (2003), and this particular method is determined by PAA, a dimensionality reduction algorithm. The algorithm involves an exclusive process of discretization that uses advanced representation between the raw time series and the symbolic patterns. Discretization involves data conversion into the PAA representation, and then shows a distinct pattern of that PAA representation, thus

creating a symbolic representation of the time series. Ultimately, the method allows for distance measurements to be less than the ones defined in the symbolic space, in that a common time series representation (a sequence of data points presented by a line) is converted into a symbolic representation.

The SAX algorithm implements discretization in two phases as show in Figure 2.2 (Lin et al. 2003). First, a time series is divided into equal-sized segments called *N*, the values of which are estimated and then substituted by a single coordinate. These *N* coordinates collectively form the standard method of the strings (PAA) representation of time series. The second phase involves conversion of the PAA coordinates to representations. It is achieved by measuring the cut-off points, dividing the distribution space in an area into equal parts, where *a* is the size of alphabet defined by the user. The cut-off points make sure that a segment in any area is approximately equal. If the symbols of the representation are unequal, then some channels will be more likely than others (which is why, probabilistic tendency is included in the process). PAA is applied to the values obtained, which are then further converted into specific values using a probability distribution function. An important downside to the SAX method is the need to have alphabet and word sizes as inputs, which is not simply derived from a specific time series data set.

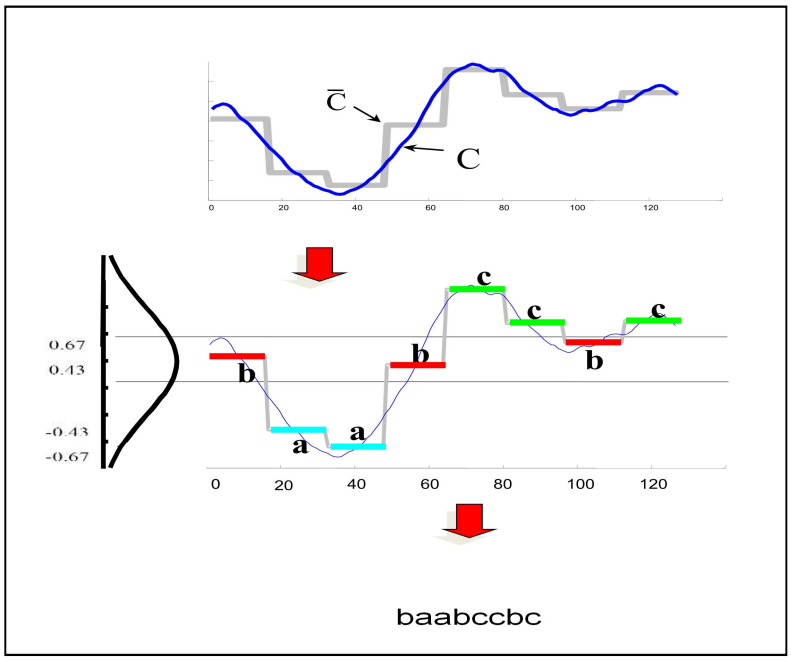

Figure 2.2 Time series transformed using SAX symbols, as described in the SAX method

The SAX involves reduction of a time series of random length n to a string of random length N (where $N < n$), and the alphabet size is defined by a random integer a (where $a > 2$). Table 2.1 summarizes the main notations used in this section and later.

Table 2.1 Summarization of the notation used in SAX

Notation	Definition
C	A time series $C=c_1,...,c_n$
$\overline{C}$	A Piecewise Aggregate Approximation of a time series $\bar{C}= \bar{c}_1,...,\bar{c}_N$
$\widehat{C}$	A symbol representation of a time series $\hat{C}=\hat{c}_{1,...,}\hat{c}_N$
w	The number of PAA segments representing time series C
a	Alphabet size

The SAX discretization process is unique because it uses an intermediate representation between the raw time series and symbolic chains. SAX first converts the data to approach the total sections (PAA) representation and the symbolic representation followed by the PAA in a discrete chain. There are two advantages to this method. First, dimensionality reduction is well defined and documented in PAA (Keogh et al. 2001), and the representation is passed automatically to the symbolic representation. Second, the lower limit that demonstrates the measure of the distance between two symbolic strings, i.e., the lower limits between the actual distance and the original time series is not trivial. The key observation that allows us to prove lower bounds is to focus on facts that the symbolic distance to measure the lower limit of the measurement of the distance from the PAA. Then the desired result by transitivity, pointed to the existing data on the PAA representation can be obtained.

A symbolic approximation SAX of time series employs a discretization technique that transforms the numerical values of the time series into a sequence of symbols from a discrete alphabet. The discretization process allows researchers to apply algorithms from text processing and bioinformatics disciplines (Lin et al. 2003). SAX has become an important tool in the time series data mining, and has been used for several applications such as time series classification, events detection (Zoumboulakis & Roussos 2007; Onishi & Watanabe 2011), and anomaly detection (Keogh et al. 2005). It enables using the Euclidian distance of the discretized subsequences (Hung & Anh 2007) and allows both dimensionality reduction and

lower bounding of L_p norms (Keogh et al. 2005).

Although the above mentioned advantages, SAX suffers from some limitations. It does not pay enough attention to the directions of the time subsequences and may produce similar strings for completely different time series. To overcome this problem we propose the Trend-based approximations representation (TVA) and value-based approximations which extend SAX by adding new string symbols in order to represent the trends of time series. Kontaki et al. (2005) proposed using PLA to transform the time series to a vector of symbols (U and D) denoting the trend of the series. Keogh & Pazzani (1998) suggested a representation that consists of piecewise linear segments to represent a shape; and a weight vector that contains the relative importance of each individual linear segment.

In addition, an extended method based on SAX was put forward by Lkhagva et al. (2006) called Extended SAX (ESAX), which proposes two new unique points to wholly represent time series data, i.e., max and min points of each segment. The method is specifically designed for financial use which requires identification of important patterns effectively and with precision; and is determined by PAA representation in order to decrease dimensionality using mean values of equally scaled frames. Thus, three values for each segment are used for time series data representation: the original mean value, and max and min points. Once PAA is determined, the even-sized pieces and their mean values are established and consequently, the max and min values of each section. The ESAX method is a more accurate representation for a variety of datasets especially the higher frequency datasets. Furthermore, ESAX refines the dimensionality to show time series more efficiently, yet there is still need for further researches to reduce dimensionality without compromising the efficiency.

Taking into account this problem as one part of the search; the algorithms in the literature generate high possibility to miss some important patterns in some time series datasets, because some of the Euclidean Distances (ED) are quite large. SAX algorithm focuses to take into consideration only the central tendency and not concern to the present dispersion in each segment. In addition, PAA algorithm converts the

data into particular values by means of a probability distribution function (Keogh et al. 2001). The SAX algorithm needs the alphabet and word sizes as inputs, which is the major drawback, since it is not apparent to figure out them from a specific time series data set. In (Hung & Anh 2007) have combined SAX and PLA to improve similarity search on time series. However, since the SAX algorithm suffers some defects. The PAA is based on mean values approximation, it has high possibility to miss some important patterns in some time series datasets, because some of the ED are quite large. Therefore SAX may not produce a good tightness of lower bound (Lkhagva et al. 2006; Hung & Anh 2008). In (Karamitopoulos & Evangelidis 2009) noted that PAA approach focus to take into consideration only the central tendency and not concern to the present dispersion in each segments. Figure 2.3 shows some important patterns that are missing while SAX representation has been used on a sample financial time series data (Lkhagva et al. 2006). Therefore, the enhancement of SAX representation aims of this part is to adapt PAA with SAX algorithm to find the most suitable aggregated methods for time series. Many techniques have been proposed for enhancing PAA representation, (Keogh et al. 2001; Hung & Anh 2008; Karamitopoulos & Evangelidis 2009; Li & Guo 2013). Therefore, the ESAX^{+} representation aims to adapt PAA with SAX algorithm to find the most suitable aggregated methods for time series.

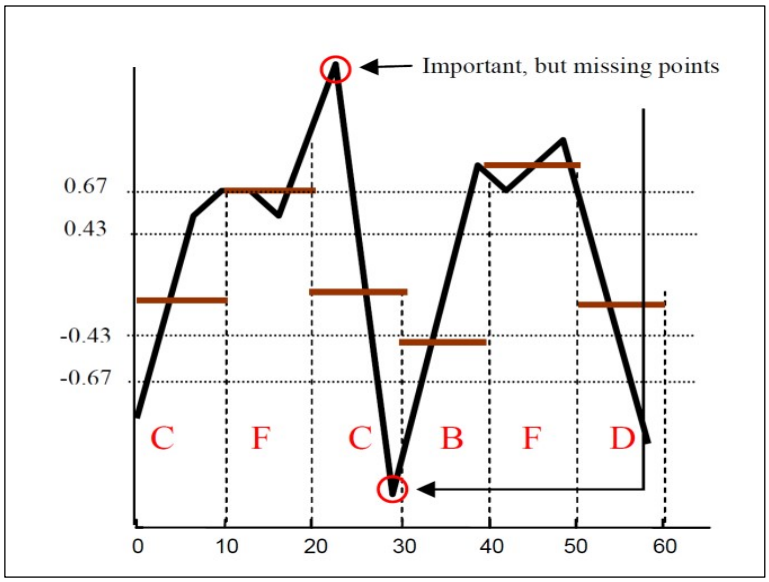

Figure 2.3 Some important points (shown in red) are missing

Despite the undeniable importance not much work is in progress or the expansion of the SAX representation. SAX provides solutions or most of issues at hand. Some of SAX extensions proposed are discussed below. A representation algorithm (HSAX) was proposed to work in combination with Harmony Search algorithm (HS) for finding the optimum word size (*Ws*) and alphabet size (*a*), these features hold importance in the SAX process for differentiating and symbolizing the time series data sets (Ahmed et al. 2011; Alshareef et al. 2016). The HSAX algorithm has been applied on some conventional time series data sets for the evaluation of the HSAX with other meta-heuristic GENEBLA (García-López & Acosta-Mesa 2009) and unique SAX algorithms. It is deduced from the results that HSAX gives a superior error-rate in comparison to that by SAX tests, when the dimensions of the alphabets and the size of the word of the actors used is fixed. The error rate is pretty much the same however, but HSAX has better word and alphabet dimensions in comparison to GENEBLA. The order of search also supports maintenance of substantial volume of information. The HSAX has the ability to work without a priori, as the variables are calculated by HS algorithm, which is another merit of this algorithm.

Shieh & Keogh (2008) proposed another extended method based on SAX for indexing and mining terabyte sized time series called the indexable Symbolic Aggregate approximation (iSAX). Thereby establishing itself as a multiresolution symbolic representation to index datasets that are numerous orders of magnitude larger than anything else ever recorded, and the inherent characteristics of iSAX makes it a fast and quantifiable method of indexing without the need for any particular databases or file managers. An iSAX word, representing a set of time series, forms the main part of any iSAX index. This index can be further split into two mutually exclusive subsets by increasing the cardinality along one or more dimensions. The iSAX aims to achieve a combination of both the rapid precise search and ultra-fast estimate search as sub-routines in data mining algorithms, and thus enable accurate mining of particularly large datasets, that include a huge number of time series, taking up to a terabyte of disk space. The iSAX representation boasts an important advantage, despite not being able to get quicker times: it's representational, enabling the use of data components and algorithms that are not sufficiently defined for real-valued data, for example Markov, hashing, suffix trees models etc. (Faloutsos et al.

1994; Lonardi & Patel 2002; Kwapisz et al. 2011).

Owing to the drawback of the SAX method, several studies have suggested ways to refine it, indicating the need for further development of the representation, and thus improving the SAX method. Pham et al. (2011) proposed an adaptive symbolic approach, i.e. the adaptive Symbolic Aggregate approximation (aSAX), that can be further extended to the indexable multiresolution symbolic representation called the indexable aSAX (iaSAX). The original SAX inspired these representations but the adaptive vector is of "breakpoints" that are determined by a pre-processing phase. This phase consists of two factors as inputs; word size that controls the number of approximating elements and alphabet size that controls the granularity of each approximating element. Applied with real world time series datasets, the experiments established the two proposed algorithms' theoretical analyses and their efficiency in terms of producing tighter lower bound, greater pruning power, and less random disk accesses which implied higher performance in query by content.

In addition to the methods outlined for time series representation, a symbolic technique was presented by Li et al. (2012) called Trend-based symbolic approximation (TSX) and applied to real fiscal time-series dataset to affirm the practicality of this approach. This original symbolic representation manages to include the segment average value feature as well as trend feature and information with good resolution, and additional support time-series data mining tasks. Trend features of time series is obtained through TSX using Most Peak point (MP) and Most Dip point (MD), and to design multi-resolution discretization algorithm the trend function is then transformed into symbols. A time series representation method called Random Shifting based SAX (rSAX) was designed by (Bai et al. 2013) to drastically improve the tightness of lower bounds of representations without increasing the granularity of respective representations as was the case in SAX.

2.5 TIME SERIES PREDICTION

Prediction denotes a method of extracting data that is most common and easily recognized, and can be viewed as clustering or classification. Rather than predict a

current state, prediction is anticipation of a future state (Ralanamahatana et al. 2005), and its precision and accuracy is therefore of utmost importance considering its role in various decision processes. Thus, there is an ongoing need for improvement of its efficiency. It involves anticipation of future values of time series of time-series variables using variables obtained in the past and is proved to be one of the most intricate tasks (Han et al. 2011; Kleist 2015). Prediction methods are classified into three categories: short-term, medium-term and long-term predictions. Short-term prediction, as the name suggests, is used to predict events in short intervals in the future such as days, weeks or months. Additionally, medium-term will help predict events in longer periods of time such as a year or two, and long-term will be utilized to predict events in even longer periods of time such as several years in the future.

Time series prediction has garnered widespread recognition amongst researchers owing to its importance in various fields in the recent years and consequently, is one of the most applied time-series tasks that is designed to predict visible dependencies variables to further predict the next several values of the series. Various time series prediction applications can be appreciated in different fields (Niennattrakul et al. 2010) and the technique commonly utilizes regression analysis. These applications range from prediction of natural disasters (for getting advance warnings such as temperature, flooding, hurricane, snowstorm, epidemics, stock crashes, etc.) to forecasting of electricity, that has become a significant part of the competitive energy markets for planning system energy efficiency and operation, and volume of sales of cell phone accessories.

Various researchers have put forward several different time series prediction techniques, and extensive literature on time prediction is present in the field of statistics. Some of the more common methods for time series prediction have been proven to be effective only in certain experimental conditions, such as exponential smoothing (Gelper et al., 2010), ARMA (Huang and Shih 2003; Taylor 2010), ARIMA model (Chatfield 2002; Kang 2003; Kim 2003). The exponential smoothing model and ARIMA (Autoregressive Integrated Moving Average) model are only useful for capturing linear features of time series. A new model based on ANFIS (Adaptive Neuro Fuzzy Inference System) and ARIMA was proposed by Faulina et al.

(2012) to predict monthly rainfall for Pujon and Wagir, two different areas of Indonesia. The results showed that AFNIS gives excellent forecast in monthly Pujon's data yet in Wagir's data ARIMA model gives a perfect forecast. Moreover, ARIMA method is also popular and extensively used for forecasting climatic changes. ARIMA uses the data collected from weather stations located in Vancouver Island and applies temperature as a indictor (Gao et al. 2016).

2.5.1 Naive Bayesian Approach

Bayesian approach is made use of by other fascinating time series prediction methods. A strong Naïve independence assumption is taken into account by a probabilistic classifier based on the Bayes' theorem, namely Naive Bayes predictor. Every feature individually contributes to the probability of a specific decision, as considered by a Naive Bayes. The training of the Naive Bayes can be done in an efficacious way in a supervised learning setting, performing much better in different complicated real-life situations, particularly in the discipline of the computer-aided diagnosis, keeping in mind the nature of the underlying probability model (Belciug 2008; Gorunescu 2011).

Problems in which there is no model of the domain knowledge can be solved by an Artificial Intelligence approach, namely Naive Bayesian (NB), and primary machine learning capabilities are also demonstrated by it. Text classification can be done by one of the most shared supervised classification techniques which is NB (McCallum & Nigam 1998; Medhat et al. 2014). Bayes' rule is the basis of NB, which is based on a simple probabilistic classifier. Given a possible value taken by the output attribute, the naive Bayes technique learns the conditional probabilities of every input attribute so that it can build a probabilistic model. Afterwards, the prediction of an output value is performed, when a set of inputs is provided, by this model. This takes place by the application of Bayes' rule on the conditional probability regarding noticing a potential output value provided that the attribute values in the given instance are noticed to be together. The Bayes' rule is defined prior to defining the technique (Langley et al. 1992). There is a requirement of every variable in the case to be discrete in the NB technique. Prior to being used, continuous valued variables must be discretized. Due to the fact that missing values for a variable can cause difficulties

at the time of calculating the probability values for that variable, they are not allowed. As a solution, they can be replaced with a default value for that variable.

The utilization of the NB approach is in various areas alongside good outcomes, there has been utilization of relatively robust, easy to implement, fast, and accurate, naive Bayes classifiers for mining tasks in a number of different disciplines (Kim et al. 2006; Chen et al. 2009; Rosen et al. 2011; Lines et al. 2012). E-mail clients based on spam filtering (Sahami et al. 1998; Guo et al. 2014; SathyaBama et al. 2016), making decisions regarding treatment methods and the diseases diagnosis (Kazmierska & Malicki 2008; Ali et al. 2015; Zhou et al. 2015) and the rRNA sequences classification in taxonomic studies (Wang et al. 2007; Cole et al. 2013) are some of its examples. Nevertheless, Bayes classifiers will perform in a really poor way because of the non-linear classification problems and strong violations of the independence assumptions. There has to be consideration that the classification model which we want to select is dictated by the kind of problem to be solved and the type of data. Pragmatically, there is recommendation to hold comparison between various classification models on the particular dataset and the prediction performances alongside computational efficiency are to be taken into account.

For managing a time series prediction utilizing NB predictor, a number of applications have been provided. A Naïve Bayesian method for cumulative rainfall time series prediction was suggested by Rivero et al. (2013). The proposed model is provided with the time series rainfall data sets acquired from some geographical points of Cordoba, Argentina. The given data set accepts the prediction outcomes acquired from the NB model which ANN filter applies. Identification of the experiments in weather prediction has been done by Nikam and Meshram (2013). Weather data can be the name given to the meteorological data with useful information. For the research, the data gathered from Indian Meteorological Department (IMD) is considered. The rainfall is analyzed and predicted by the utilization of the Bayesian data mining techniques. The training data set is utilized to train the model and present test data is used to test for prediction accuracy. Weather prediction model works on high performance computing and supercomputing power used by the meteorological centres. There is indication regarding the prediction

accuracy being superior with moderate computing resources for rainfall prediction by the result making use of the Naïve Bayesian approach. On the basis of Regression tree (CART), naïve Bayes, K-nearest Neighbour and Neural Network, rainfall forecasting models were developed by Gupta and Ghose (2015). The dataset of 2245 samples of New Delhi rainfall records from June to September (the annual rainfall period) from 1996 to 2014 has been used including the features dew point temperature, mean temperature, sea pressure humidity and wind speed. Neural Network performed the best with this data with 82.1% accuracy, second best is KNN with 80.7%, Regression Tree (CART) scored 80.3% while Naive Bayes provides 78.9% accuracy.

For predicting the ECG abnormalities without former knowledge, NB predictor for multivariate maximal time series motif was used by (Padmavathi & Ramanujam 2015). Steps like pre-processing, feature extraction, and prediction are present in ECG signal prediction, in this task. With the utilization of Baseline wander removal, symbolic discretization utilizing R-R interval approach, the original ECG signal should be pre-processed. The Multivariate Maximal Time Series Motif is mined by using the symbolic discretized signal, in Feature extraction stage. Ultimately, the classification of the basis of Multivariate Maximal Motif of Time Series Signal is predicted by the application of NB classifier. The comparison between the precision of existing and different Feature extraction Techniques proves the efficacy of the proposed work.

Froelich (2015) proposed a method regarding the representation of predictive model in a dynamic Gaussian Bayesian network (DGBN). For being controlled for the predictive of a hydrological time series, the research was introduced. Reaching the most superior potential prediction precision of daily urban water demand is the goal of DGBN technique. The models which have their bases on automatic learning of network structures are not the most efficacious, according to the findings of a large number of comparative experiments. Models having manually designed structures are better in performance than them. For this reason, other simple DGBNs are investigated and compared with by the study. The simple DGBN model, which is better in performance than other selected state-of-the-art prediction methods, is superior, as proved by the experiments.

For evaluating uncertainty in real-time flood prediction making use of river discharge problem linked to every most probable reason of error, a Bayesian Forecasting System (BFS) was introduced by Biondi and De Luca (2013). The various hypotheses for which the relative influence on the BFS outcomes has been assessed are following: (i) UD, which works with the utilization of a perfect hydrological model; (ii) PD, which is the same as former distribution when, for predicting the river discharge, the PQPF and the model simulation are non-informative, for this reason a normal reference for evaluating skills of the predictive distributions is provided; (iii) HD, which works with the utilization of an ideal input forecast (due to the true future rainfall values being unknown, they are not appropriate for online forecasting); and (iv) TTD, when both causes of uncertainty are considered, it is the total predictive distribution. Sufficient verification tools, appropriate for probabilistic forecasts of continuous variables as stream flow, were used for the performance of the BFS. A large number of variables of the prediction quality of the all-time-varying predictive distributions: accuracy, sharpness, and calibration, have been evaluated by the utilization of scalar metrics and graphical tools. For hydrological typical, a non-informative rainfall for forecasting stream flow and a perfect input and output prediction, the BFS properties was really appropriate.

Detection of heart disease prediction was done utilizing Cleveland dataset as NB technique was employed by Medhekar et al. (2013), in another task. Health data is differentiated into five categories which are no, low, average, high and very high, by the system, the class label of the case will be predicted by the system when an unknown case arrives. Training for classification and testing for prediction were the two basic functions which were run. For health students, the system can be helped as training tool and doctors will also use it as their serving hand. The method dealing with the enhancement of the accuracy with respect to proficiency and precise prediction in the health field is the centralization of the research. Moreover, Kharya and Soni (2016) made use of a predictive approach in general framework. Weighted Naive Bayes predictor was the basis of the approach regarding breast cancer detection. For predicting whether there is breast cancer, building of the model making use of UCI datasets is done. Training mechanism which is easy to implement, modifiable, readable, efficacious is present in the model and it does not depend on training set

size, engagement of training with high dimensionality is trouble-free and the doctors can be assisted in the diagnosis of different diseases by it.

The assumption regarding the values of every input variable being conditionally independent provided the value of the output variable is utilized by the naive Bayes technique. However, there is probability in appropriate predictions not being produced when assuming conditional independence of every given input. The different approach given by Li (2010) is that for differentiating the information content of forward looking statements, a Naive Bayesian machine learning approach is used. With respect to the correlations amongst the predicted tones of forward looking statements and actual earnings, the naïve Bayesian technique is better as compared to a traditional dictionary based approach, as claimed by the researcher. The concept regarding the application of conditional independence on a certain number of inputs instead of on each of them is introduced by Bayesian Networks (Bayes Nets). The worldwide assumption regarding conditional independence is avoided and simultaneously, some amount of conditional independence amongst the inputs is maintained by this idea (Goldszmidt 2010; Bielza & Larrañaga 2014). For a set of variables, the joint probability distribution is provided by Bayes Nets. The network contains all the variables in the case in the form of nodes.

A symbolic approximation SAX of time series employs a discretization technique that transforms the numerical values of the time series into a sequence of symbols from a discrete alphabet. The discretization process allows researchers to apply algorithms from text processing and bioinformatics disciplines (Lin et al. 2003). SAX has become an important tool in the time series data mining, and has been used for several applications such as time series classification, events detection (Zoumboulakis & Roussos 2007; Onishi & Watanabe 2011), and anomaly detection (Keogh et al. 2005). It enables using the Euclidian distance of the discretized subsequences (Hung & Anh 2007) and allows both dimensionality reduction and lower bounding of L_p norms (Keogh et al. 2005).

A representation of multivariate time series which combines trend-based and value-based approximations (TVA) proposed by Esmael et al. (2012). It produces a

compact representation of the time series which consists of symbolic strings that represent the trends and the values of each variable in the series. The TVA representation improves both the accuracy and the running time of the classification process by extracting a set of informative features suitable for Naïve Bayes, SVM, RI, K-NN and DT classifier algorithms. The inputs of the proposed classifier are the TVA features computed from the current segment, as well as the predicted class of the previous segment. The approach enables highly accurate and fast classification of multivariate time series. In addition, Bat-Erdene et al. (2017) presented a method for identifying unknown packing algorithms for malware detection, and classify them based on entropy approach by using four similarity measurements and naive Bayes (NB) and a support vector machine (SVM) classification techniques. The proposed method converts the entropy values of a particular location of memory into symbolic representations based on SAX algorithm, which is known to be effective for large data conversions. The proposed system was tested on a large dataset that includes 324 benign packed files and more than 326 packed malware types. The method can identify packing algorithms of given executables with a high accuracy ranging from 95.0 to 99.9%. The method confirms that the combination with strong pattern-recognition algorithms through an entropy analysis, produces a highly accurate packer classification system for real data.

The approach for the detection of complex events in wireless sensor networks presented by Zoumboulakis and Roussos (2007). Complex events are sets of data points that correspond to interesting or unusual patterns in the underlying phenomenon that the network monitors. The approach transforms a stream of real-valued sensor readings into a SAX symbolic representation, and then performed using distance metrics for string comparison to detect events that are difficult or even impossible to describe using SQL-like languages and thresholds. The approach implemented for the TinyOS and Contiki operating systems, for the sky mote platform, and temperature, relative humidity, light and voltage and ECG data, the detection was accurate in both cases and the proposed algorithm using SAX managed to detect the exact points of change. The performance approach appears to match very well the unique resource constraints of sensor networks. Furthermore, Event detection using smart house sensor data based on SAX representation that allows fast retrieval

of archived sensor data for a smart house. SAX converts data from a time series into a string (Onishi & Watanabe 2011). The approach achieved fast similarity event detections using a suffix tree and performed experiments for event detections such as the opening and closing of doors. The approach needs to perform similarity retrieval because sequences assigned to identical events have slightly different patterns depending on the person triggering the event. The error-tolerant recognition algorithm is to realize the fast similarity retrieval, which deals with the edit distance retrieval method with a tree

Niu & Zhang (2015) presented a rainfall prediction model and the data was taken from China Meteorological Administration (CMA) and the different data mining Algorithms such as Naive Bayes (NB), Support Vector Machine and Back Propagation Neural Network were used. In this model only seven attributes are taken as input are pressure, temperature, evaporation, humidity, wind speed ,sunshine, surface temperature and the comparison has been made between overall-data -rate (RO) and rainfall data-rate (RR). SVM gives high prediction accuracy by 90% accuracy, but it is highly complex and it needs extensive memory requirements for classification. Furthermore, Nikam and Meshram (2013) presented a Naive Bayes algorithm for a rainfall prediction model. NB is simple to implement. Classification efficiency is high by 80% accuracy. It predicts accurate results for most of the classification and prediction problem. The model only 6 attributes are taken as input are mean sea level pressure, relative humidity, station level pressure, temperature, vapour pressure, and wind speed. NB needs large number of records to obtain good results. The precision of algorithm decreases if the amount of data is less.

Climate change prediction analyses the behaviours of weather for a specific time in rainfall forecasting was presented by Zainudin et al. (2016), where specific features such as humidity and wind used to predict rainfall in specific locations. Different data dining techniques lead to different performances depending on rainfall data representation including representation for monthly and daily patterns. The study used multiple classifiers such as Naïve Bayes, Support Vector Machine, Decision Tree, Neural Network and Random Forest for rainfall prediction using Malaysian data. The dataset has been collected from multiple stations in Selangor, Malaysia.

Several pre-processing tasks have been applied in order to resolve missing values and eliminating noise. The results show that with small training data (10%) from 1581 instances Random Forest correctly classified 1043 instances. This result put Random Forest in the forefront of the five techniques we have been used.

2.5.2 Other Methods

A Support Vector Machine (SVM) is usually used for learning classification, and prediction. SVM is based on statistical theory and structural risk minimization principal and have the purpose of defining the location of decision boundaries known as hyperplane which produce the optimum splitting of labels (Cortes & Vapnik 1995; Burges 1998; Han et al. 2011). Maximizing the margin and producing the largest potential distance among the splitting hyperplane and the instances on both side of it has been confirmed to decrease an upper bound on the estimated generalization error. SVM is a set of related supervised learning methods which are typically used for prediction problem. A training is prepared with two pre-defined category i.e. normal and deviation. SVM operates by building a model that maps these training data into a high-dimensional vector space separating the normal and deviating data instances (Campbell & Ying 2011). The model is then used to predict the category of the new unlabelled data instance to either belong to normal or deviation category. SVM is a preferred a huge data for prediction because it can handle large feature space and sparse instances SVM was used by various researchers to detect deviations in huge data such as weather data (Lu & Wang 2011; Ma & Guo 2014; Chen et al. 2015; Kisi et al. 2015).

SVMs approach has been used in various other researches for time series prediction such as in (Mukherjee et al. 1997; Miranian & Abdollahzade 2013), where they applied the SVM on the same data base of chaotic time series that was implemented to compare their performances. The different approximation techniques that were compared were polynomial and rational approximation, local polynomial techniques, Radial Basis Functions, and Neural Networks. The study established SVM's performance better than the others. SVMs have also been used for load forecasting by Chen et al. (2015) creating models on relative information, like climate

and previous load demand data. Continuous temperature prediction has not been proven to be accurate and therefore time-series modelling schemes are better suited for temperature approximation. Essentially, in a 30-day period forecast, temperature will not vary much and to integrate it into the model is pointless. Thus, as proven by the experiments, time-series modelling schemes prove to be a better method to forecast load demand more efficiently.

A modular type SVM was proposed by Qing et al. (2012) further improved the efficiency of the time series analysis. The model comprises of various steps such as screening of environmental factors in a nonlinear fashion by SVM, then estimation of the order using Controlled Autoregressive (CAR) and ultimately, using the prediction model to affirm the SVM-CAR. In conclusion, the results showed that this particular model is significantly accurate and proves to be important in drought and flood prediction. Furthermore, Kisi et al. (2015) surveyed water level fluctuation predicting in Urmia Lake using support vector machine coupled with fire fly algorithm (SVM-FA). The study uses daily water-level data were used to train. The findings indicated that the developed SVM–FA models can be used with confidence of predictive strategy for lake level prediction.

Artificial Neural Network (ANN) is a collection of interconnecting artificial neurons or programming constructs. ANN is also defined as a mathematical or computational model for information processing (Fausett 1994; Maren et al. 2014). A typical architecture of an ANN consists of input nodes, hidden nodes and output nodes. Each node has value and is connected with weighted connections. The values are passed from input to output where in each pass the value goes through transformation with the multiplication of weights in the connection and the summation of multiplied values from different nodes going into a single node.

The previously mentioned Neural Network is another learning technique that is particularly useful for prediction in pattern recognition applications (Nanopoulos et al. 2001; Zhang 2012; Wang et al. 2016; Zheng et al. 2016), and has additionally been used in general regression analysis i.e. the estimation of a function by fitting a curve to a set of data points. Nonlinear modelling can be achieved by Artificial Neural

Networks (ANNs) in a multiregional short-term load forecasting system for an electricity utility was developed by Hernandez et al. (2013) using ANNs, the control area of which covers a large geographical area. ANNs, SVM, and decision forests predictors for prediction of event-related potentials (ERP) were compared in EEG data analysis, which affirmed each method's validity but the validity depends greatly on the dataset applied (Kuncheva & Rodríguez 2013).

Two different prediction methods are investigated for short term wind power prediction of a wind farm in Peng et al. (2013). The adopted strategies are individual artificial neural network (ANN) and hybrid strategy based on the physical and the statistical methods. The performance of two prediction methods is comprehensively compared. The calculated results show that the individual ANN prediction method can yield the prediction results quickly. The prediction accuracy is low and the root mean squared error (RMSE) is 10.67%. By contrast the hybrid prediction method operates costly and slowly. However, the prediction accuracy is high and the RMSE is 2.01%, less than 1/5 of that by individual ANN method. Meanwhile, it is found that the errors of the prediction have some relation with the wind speeds. The prediction errors are small when the wind speeds lower than 5 m/s or higher than 15 m/s. The reasons for such phenomena are also investigated. A process of careful analysis having dependency on the properties of the issue is suggested by selecting the suitable model utilizing ANN; as a really good accuracy, which the Electric Utility Control Centre will make use of in Oaxaca for the energy supply, was demonstrated by short term wind speed forecasting. For wind speed prediction modelling, Wavelets, Neuro-Fuzzy and Radial Basis Functions techniques were made use of by Liu et al. (2013).

A data mining methodology, Decision Tree technique (DT) (Quinlan 1986), serves as a powerful solution to prediction problem in different real world applications, in some other work (Kantardzic 2011). In DT, the identification of the local region is done in a sequence of recursive splits via decision nodes with test and it is a hierarchical model (Hayat et al. 2012; Hayat & Khan 2013; Joseph et al. 2013). Moreover, for analyzing and considering multiple variables and for the weather prediction utilizing rainfall data set, decision tree technique (D3) can serve as a perfect technique, as shown by Geetha & Nasira (2014). 80.67% is the success rate of the

rainfall prediction for 2014, as shown by the results and enhancing the accuracy rate is still being worked for. In terms of rainfall prediction, D3 methods seem to hold great significance. In addition, there are many decision tree based techniques like ID3, C4.5, C5.0, CART etc. These techniques have the merits of high classifying speed, strong learning ability and simple construction (Brijain et al. 2014). Decision tree can be applied in weather forecasting process which deals with predicting whether the weather is sunny, overcast or rainy and the amount of humidity if it is sunny. This tree model can be applied to determine whether the atmosphere is suitable to play the tennis or not. So, a person can easily find the present climate and based on that decision can be made whether match can be possible or not (Jadhav & Channe 2016).

Aydin et al. (2009) suggested a prediction technique making use of time series data mining on the basis of fuzzy logic. With the utilization of investigating method at first step back, a synthetic earthquake time series has assisted in earthquake prediction. With the utilization of nonlinear time series analysis, transformation of time series has taken place to phase space. Afterwards, for predicting optimal values of significant parameters because of which the time series events are characterized, fuzzy logic has been utilized. Application results have provided the proof for truth of prediction technique based fuzzy logic. The ability of detecting various events is an advantage of this approach. Variations take place in membership functions and fuzzy rules on the basis of the issue. The determination of temporal pattern clusters without dependent of applications takes place by using Gaussian-shaped fuzzy membership function. For this reason, the definition of the temporal pattern can be done at a smooth region. Local minima and yield to heavy computing load is collected by genetic search technique. Temporal patters are illustrated more pragmatically in the time delay embedding by the fuzzy temporal pattern cluster.

For the prediction task for non-stationary time series, the utilization of fuzzy based methods was suggested by Shah (2012). For capturing the trend which is prevalent in the series, the main concept makes use of IF-THEN based fuzzy rules and is centralized on based on unequal length fuzzy sets. The value is predicted and the transition points are identified, wherein the series is ready to include subject expert's opinion to forecast and may alter its shape, by the method. Three unique kinds of

financial data have been used to test the mode, having outliers and no outliers. The superiority of fuzzy based expert system is established over conventional statistical methods and a realistic universe of discourse is provided by the utilization of the outlier. An enhanced prediction, having lesser MAPE error in terms of every series tested, is provided by the suggested model.

Long-term prediction has proven to be more complex since it has the ability to use its own predicted values as inputs to predict further future values (recursive prediction). And to solve this dilemma, Herrera et al., (2007) proposed using two different methods, the TaSe fuzzy TSK model and the least-squares SVMs model. They presented the models by utilizing it in two different time series examples, the recognized artificial time series Mackey-Glass and a river flow. Both of the models improved their recursive long term prediction for both the applied time series. The TaSe models proved to be highly accurate for function approximation and time series prediction problems as it took into account the intrinsic aspects of the TSK fuzzy systems, while the LS-SVM technique displayed perfect approximation results. Yet another methodology was proposed by Sorjamaa et al., (2007) to overcome the long-term prediction dilemma, where the direct prediction strategy and refined input selection criteria, i.e., *k*-nearest neighbours method (*k*-NN), mutual information (MI) and nonparametric noise estimation (NNE), is combined. These criteria greatly improve the selection of inputs, making it efficient, and was needed to tackle the increase in computational load. This global input strategy that combines forward selection, the backward elimination and the forward-backward selection is better compared to the extensive search that suffers from a massive computational load. Amongst the three criteria, the *k*-NN is the fastest since hyper parameters selection is unnecessary, making it 10 times faster than MI and 20 times faster than NNE.

2.6 TIME SERIES PATTERN DISCOVERY

A vast variety of applications make use of data mining. Nevertheless, in the case of temporal data mining, the groups into which the probable aims of data mining, which are frequently known as the task of data mining or mining operations (Han et al. 2011), can be classified are: (i) prediction, (ii) clustering, (iii) classification, (iv)

search and retrieval and (v) pattern discovery. Thorough investigation of the first four groups has been done in pattern recognition and traditional time series. There is a relation between the tasks of time series prediction and forecasting implying that the basis of finding future values of the time series is on prior data samples. The literature tells that the prediction of time series has been worked upon a lot (Keogh et al. 2002). (Aggarwal et al. 2009; Aggarwal & Han 2014; Aggarwal 2015) focus has also been on the utilization of patterns discovery on time series. There is a connection between discover patterns of time series and finding frequent patterns and sequential patterns of time series on the basis of their similarity. Relevance can be found in the time series patterns discovery activity of a number of applications. In terms of patterns discovery on time series, different techniques are present (Last et al. 2001; Patel et al. 2002; Muthukrishnan et al. 2004; Shi et al. 2011; Maimon & Last 2013).

2.6.1 Pattern Detection

Pattern detection from time series is a really important task in data mining. In the discipline of data mining, attention is being given again to pattern detection in time series. Finding the amount of change points initially and identifying the class of those points as one window is the key concept. For the extraction of a signal having the most superior fitting lines and returning the segments' end points as change points or sequence of time points called a pattern, different applications exhibit the pattern detection problem (Epifani et al. 2010; Pimentel et al. 2014; Miao et al. 2016). Nevertheless, there is a requirement regarding time series being passed through two different processes having to be applied segmentation and pattern discovery for multivariate time series, in the pattern detection phase. For changing points and windows size recurrently with the assumption that nonlinear models can be suitable for the data and detecting the outlier points, discrete wavelet for animal movement were used by Sur et al. (2014). Furthermore, for fitting the data with linear segments so that patterns in time series could be found, online algorithms were utilized by Keogh et al. (2004). The capturing of the change points detection from multivariate data can be done by a window-based detection method suggested by Hu et al. (2014). Hybrid distance functions in recurrence plots generation, through which the difference every pair of multivariate time stamps have with each other could be described more

efficiently, quickly and automatically, were introduced by the researchers.

Segmentation approach in time series is discussed in the literature in different contexts, and therefore the segmentation problem referred as a pre-processing phase and essential task for diversity of data mining tasks, as a trend analysis approach, as a discretisation problem in function of time series representation, as a component in data mining applications in different fields, etc. For described discussions regarding to the segmentation approaches in these contexts interested studies are denoted to the references (Shatkay & Zdonik 1996; Fu et al. 2001; Park et al. 2001; Chung et al. 2004; Gionis & Mannila 2005; Lovrić et al. 2014), which cover excellent and extensive theoretical surveys. For the purpose of pattern detection and analysis of segmentation in time series, the excellent reviews of segmentation algorithms, presented by Keogh et al.(2004), Bingham et al. (2006) and Chundi and Rosenkrantz (2009), have been analysed and theorizes. In significant literature, many algorithms proposed for time series representation into their segmented styles, and resolution of sufficient number of alternatively heterogeneous segments (Hiisilä 2007). Furthermore, many methods in the literature proposed to create time series representations could be found (Mörchen 2006; Lin et al. 2007).

Through reducing the dimensions and maintain the primary features simultaneously, creating an accurate approximation of time series is the purpose of time series segmentation approach (Keogh et al. 2001; Lin et al. 2003). The approximation of time series in the style of linear is the essence of Piecewise Linear Approximation representation (PLA). Fast similarity search, new clustering and classification algorithms, novel distance measures for time series and change point detection can be supported with PLA which is the most frequently utilized representation. The three segmentation algorithms are as follows: Top-Down algorithm, in which a time series is recursively separated until some error measure takes place, Sliding Window algorithm (SW), in which a segment is merged until specific error criterion takes place and Bottom-Up algorithm, which starts with small segments that are grown until some error criterion takes place are the types of segmentation algorithm (Keogh et al. 2004). The measurement of error criterion is usually done with Least Squares Error (LSE). Frequent weak outcomes are provided

by sliding window algorithms even though they are given really fast time series analysis in different applications. In addition, in terms of accuracy, sliding window techniques are overcome by top-down and bottom-up algorithms; however their application cannot usually take place for online applications. Lately, for the purpose of applying the benefits of the concerned methods, sliding window and bottom-up (SWAB) were combined by Keogh et al. (2001) and (2004) to introduce a new algorithm. Issues based on the efficiency and properties of the quite well segmentation algorithm are discussed and analysed in Terzi and Tsaparas (2006) and Lemire (2007).

For detecting the relationship the non-trivial patterns in the time series data have with each other because of which identifying the patterns and events trend in the data is allowed, temporal data is tried to be extracted by sliding windows algorithm (SWA). The differences between two sets of weathers can be discovered by SWA and it has a number of possible applications, like finding the unique behaviour, normal or abnormal patterns of climate changes. Based on whether a discriminative pattern can be detected, the classification of pattern detection can be done (Ahmed et al. 2012; Feng et al. 2012). For the requirement of thorough human expert explanations and difficult to comprehend nature of climate changes, the supervision of this learning approach has been done in a weaker way. A procedure that is building detailed labels for template data and takes too much time, and the frequent subjective biases take them.

The issue regarding sliding windows algorithm (SWA) is taken into account by the study Keogh et al. (2001) and Alshareef et al. (2016). In SWA, till the time an error occurs, the segments are grown and on the basis of the way the point is estimated to one another for creating segments, the progress repeats to next point of time series and merge print to other point. Furthermore, alongside segment boundaries also referred to as segmentation points, PLA of the time series within a segment, in which the time series are represented as segment by interpolation or regression methods, are defined by different segmentation algorithms. Due to the fact that in this domain, more than ten time points might be present in the segment and the error threshold is dynamic, the application of this algorithm cannot be directly to multivariate data and some real life applications like a weather application (river flow problem or rainfall). On the basis of change point detection of dependent variables, a new strategy is used

by the study for enhancing SWA for multivariate data problem and real life applications.

With the utilization of SW approach for collecting and storing temporal data on significant information, significant amounts of temporal patterns are produced by various studies and efficacious applications. To discover merging trends from time series data via a SW concept, the Dual Support Apriori for Temporal data (DSAT) algorithm was suggested by Khan et al. (2010). For finding recently frequent patterns over a data stream, a SW approach was suggested by Chang & Lee (2003). The window size gives the definition of the significant recent range of a data. The recent change of knowledge in a data stream can be monitored by using the approach. With a transaction-sensitive sliding window, the set of frequent item sets efficaciously over data streams in another study presented for mining. A fixed number of transactions are present in the suggested approach known as MFI-TransSW (Li & Lee 2009). Large-scale data can be handled by Segmenting with sliding window algorithm as introduced by Palpanas et al. (2004). Due to the fact that its implementation can be done easily as an online algorithm, this method seems to gain attention. The performances of some present slide window based algorithm are parameter dependent even though they work well. There are complications in finding a general set of parameters for different time-series data types like electrocardiogram (ECG), water level, and stock market as they have really unique properties (Xu et al. 2013).

As of late, with the utilization of a SW techniques known as MSWTP algorithm, mining top-k frequent patterns from data online are taken into account by mining, as per the consideration of Chen (2014). By using sliding window method, efficient algorithm for frequent patterns mining over time series have been introduced by another study and data structure and corresponding frequent closed patterns are used by the algorithm for storing transactions of the window (Nori et al. 2013). An algorithm, on the basis of weighted maximal frequent pattern mining through which operations centralizing on recently accumulated parts and recent frequent patterns over data streams can be extracted, was suggested by Chang (2014) and Lee et al. (2014) by utilizing SW approach.

Fu et al. (2008) have created another new segmentation approach. The degree of effect inflicted on the shape of the time series by the concept that a sequence of data points and the amplitude of a data point construct a time series can be different. The time series considers every data point important, the overall shape of the time series can be contributed by it or the time series can be little influenced or it can also be eradicated. Perceptually important point (PIP) is the name given to the data point having importance calculation (Chung et al. 2001). In almost all of the cases in financial domain, for the evaluation of the data point importance, PIP is a preferable method, as shown by experiments. Afterwards, the specialized binary tree (SB-Tree), through which a fast lookup of time series starting from the most important data point is supported, is a time series representation used tree structure. Via segmentation on the basis of PIPs detection, representation and clustering of time series were presented by Park et al. (2010). For representing the movement curve of time series utilizing the inflection points of time series, which are a small number of PIPs, PIP has been suggested. Turning Points (IPs), whose extraction use from the maximum or minimum points of the time series, are suggested by Yin et al. (2011). There is generation of segments at various levels of details by this technique. Top-down analysis on the time series are allowed by this type of segmentation, wherein first there is identification of highly visible trends and afterwards, utilization of more detailed segments in following stages. For preserving higher number of trends in comparison to a present approach of segmentation, acceptable outcomes are acquired by the method.

2.6.2 Frequent Patterns

One of the major issues with the frequent patterns algorithm involves the detection of relationships amongst the database items. The issue of discovering the association patterns, as this was seen to be closely associated with the frequent patterns. Basically, the association patterns are considered to be the "second-stage" outputs, which can be derived from the frequent patterns. The following equation is used for clarifying it: For database D covering the transactions $T_i \ldots T_N$, for defining the patterns P that present in at least a small fraction of all the transactions, called as a minimum support, where the parameter, s is expressed in the form of the absolute number or in the form of a fraction of the overall transactions in a database. Every transaction T is considered to

be the sparse binary vector or a group of discrete values which represent the identifiers of the binary attributes which are instantiated up to the value 1. This issue was discussed with respect to the market basket data, which is to detect frequent sets of items which have been bought at the same time (Agrawal et al., 1993; Han et al. 2011). Using this scenario every attribute is seen to match some item in the superstore, whereas the bin values indicate the presence absence of these items in the transactions. Furthermore, this problem was applied in various different applications with regards to the sequential pattern mining, web log mining, and in analysing the software bugs.

The database cannot be searched by pattern discovery as it does not have specific query available. The extraction of every pattern of interest is the primary goal. Data mining contains the origins of the pattern discovery task. The origins of algorithms for pattern discovery in large databases are more recent and the discussion of these algorithms is mostly in the data mining discipline. A pattern which can be found various times in the data can be considered as a frequent patterns (Mörchen 2007). There is a connection between the development of efficacious algorithms and the formulation of useful pattern structures to discover each pattern which can be found many times in the data and most of the data mining literature (Cara et al. 2000; Yang et al. 2000; Hochheiser & Shneiderman 2002; Keogh et al. 2002; Alonso et al. 2003; Chiu et al. 2003; Weng & Zhu 2004; Papadimitriou et al. 2005; Tanaka et al. 2005; Jo\ et al. 2006; Papadimitriou & Yu 2006; Costa da Silva & Klusch 2007; Han et al. 2007; Yankov et al. 2007; Du et al. 2009; Florez & Lim 2009; Mohammad & Nishida 2009; Ye & Keogh 2009; Koopman et al. 2010; Li & Lin 2010; Mueen & Keogh 2010; Ratanamahatana et al. 2010; Sarangi & Murthy 2010; Narang & Bhattcherjee 2011; Shi et al. 2011; Wang et al. 2011).

In parallel to Hopper's approaches, interval sequences that could be acquired from time series make use of more sophisticated methods. Containments of intervals are looked for by Villafane et al. (1999) and Moskovitch and Shahar (2015). The intervals construct a restraint lattice and for reducing the storage space necessary for the naive algorithm, mining of the rules is done with the alleged Growing Snake Traversal. The formulation of patterns was done by Kam and Fu (2000) utilizing Allen's interval operators. Alleged patterns in which concatenation of operators is

only allowed on the right side are the limit of the rules. An Apriori algorithm is used to mine the patterns with. Certain techniques are present which could only be used on symbol sequences potentially acquired from time series.

With the utilization of an Apriori style algorithm, frequent episodes are discovered by Akhmetova (2006) and Leemans and van der Aalst (2014). For the production of a sequence of cluster labels, a clustering of short time series segments extracted via a sliding window was applied by Keogh et al. (2002). In terms of association rules present in a time window called if-then rules, the symbol sequence is mined. Othman & Eljadi (2011) proposed Fuzzy Apriori, FP-growth and Apriori association rule data mining to detect normal and malicious packets from network traffic. The authors collected data from of UKM University at different time intervals by using Wireshark tool. They developed a new tool for intrusion detection called as Nasser tool. The Nasser tool consists of four components namely, preprocessing, option setting, association techniques and analysis components. The preprocessing component is used to transform and select most significant features from network data set. The option setting use for setting time interval for which the analysis shall be carried out. Association rule component includes different algorithms namely Fuzzy Apriori, FP-growth and Apriori to determine normal and malicious class. The analysis components used to analyze the result. A comparative analysis between Fuzzy Apriori, FP-growth and Apriori association rule approach is presented. They shared that the Fuzzy Apriori approach is more accurate while FP-growth approach is faster. Further, they concluded that the tool achieve better performance for intrusion detection.

He & Hu (2009) introduced multivariate time series association rule data mining to analyze network traffic for intrusion detection. They collected data set from real network (intenet2 backbone network) over 20 universities server. The proportion-based analysis method used to. They used PAA to represent time series on TCP flag and convert to discrete symbolic elements. Furthermore, they applied SAX method for the analysis of discrete symbols. The authors presented multivariate time series association rule data mining to determine the normal and malicious packets. They concluded that their approach achieved good performance for intrusion detection.

Temporal rules uttered with Allen's (Allen 1983) interval logic and a sliding window are mined so that the pattern length are hampered by H¨oppner (2002)'s suggested method. With the utilization of an Apriori algorithm which utilizes support and confidence, mixing of the patterns is done and an interestingness measure ranks them later on. Min. support, min. confidence and window width are the requirements of it in the pattern discovery stage; there is no complication in its exact settings. The examination of rules according to the optimal interval relationships is one of the straight-forward enhancements which could be made. Simple relationships in Allen's interval logic (here: X before Y or X after Y) can compose arbitrary relations such as 'X and Y do not intersect'. As discussed in terms of quantitative refinement, an optimal combination can be found similarly as calculation of the support for the basic relationships have been done.

The introduction of Region Connection Calculus method (RCC) took place by Randell et al. (1992) and Cohn et al. (1997) as an interval logic for the temporal rules attempted at Qualitative Reasoning within the QR community for AI. The QR has reasoned about scalar quantities. The utilization of interval logic can be done to reason about space, as described by RCC. In terms of this type of issue wherein there is no requirement of thorough knowledge regarding the considered set relations, tractability is proven and sufficiency of path-consistency is demonstrated by RCC. RCC-8 sets having every base relation are recognized by the method. DC (DisConnected), EC (Externally Connected), PO (Partial Overlap), EQ (EQual), TPP (Tangential Proper Part), NTPP (Non-Tangential Proper Part) are the base relations and TPPi and NTPPi are their converse relations. The mixing of two or more basic relations or the special empty relation makes each non-basic relation. Every potential subset of the set of basic relations corresponds with the set of RCC-8 relations. The RCC method was utilized by Lee et al. (2012) and afterwards, by association rules mining, implicit patterns were revealed. On the basis of short-long clustering regarding spatial peculiarities utilizing RCC, a multi-level clustering method was suggested by Lee et al. (2003). Furthermore, in terms of a fragment of propositional spatiotemporal logic PST so that first-order frequent patterns in environmental and medical data could be mined and frequent patterns can be utilized as features for classification task in an accuracy increment, RCC-8 was introduced by Popelınský and Blaták (2005).

Additionally, reasoning about space and time was involved by RCC method applied for spatio-temporal reasoning and context awareness presented for interpreting human behaviour. An instance of this can be that it is normal to prepare food in the kitchen at noon, though a behaviour which needs special attention would be the one in which someone is doing the same activity at 3 in the morning in the garage (Guesgen & Marsland 2010).

Lately, RCC relations, utilizing for qualitative spatial reasoning techniques so that soccer data having implicit relational information between players and the dynamics of the game or soccer pass prediction can be analysed, have been suggested by Vercruyssen et al. (2016). For dealing with uncertainty in spatial data for weather prediction, Rough Sets Theory RST has been presented by Tripathy (2018). The classification task through which spatial regions having unclear boundaries can be modelled with is RST. The identification of vague region boundaries can be done with mixing RCC relations and the approximation concepts of RST.

Li and Lin (2010) also pondered over the issue regarding the identification of frequently occurring patterns. There is a requirement that the length of the patterns should be known in advance for most of the present work on finding time series motifs. In terms of approximate variable-length time series pattern discovery, there was suggestion of a novel approach on the basis of grammar induction. The discovery of hierarchical structure, regularity and grammar from the data could be possible by the algorithm. The identification of frequent patterns and hierarchical structure in data can be done automatically and capably by the algorithm and this is its central idea. Grammar-based methods regarding feature extraction, classification and forecasting of time series have been of more interest. More encouragement can be gained from the outcomes of Li and Lin (2010). Specifically, the identification of some vital patterns in time series can be done by the grammar-based approach, as demonstrated by them.

2.6.3 Sequential Patterns

One of the main problems is with respect to the sequential pattern mining where in the transactions are completed with a certain order (Agrawal & Srikant 1995). Majority, it

is realized that the temporal order is actual natural in several situations like the customer's purchasing behaviour, as the several items are bought at certain time periods, and are realized to come after a natural form of temporal order. However, the problem is related to the sequential pattern mining, as it is necessary to discover the relevant and the frequent item sequences. Essentially, data that are correlated to businesses and climate change as examples grow speedily, making data mining an actuality and a crucial field in the world, and has improved the attention in the database system. In data mining field, Apriori algorithm (Agrawal et al. 1993; Agrawal et al. 1996), Apriori-Total algorithm (Apriori-T) (Goulbourne et al. 2000; Coenen et al 2004) and frequent Pattern Growth algorithm (FP-growth) as the main algorithms in association rule mining approaches that is applied to obtain sequential patterns for large items in a database (Han 2001; Han et al. 2004).

There are several techniques proposed for the sequential pattern mining (Agrawal & Srikant 1994; Srivastava et al. 2000; Agarwal et al. 2001), after the first paper on the frequent pattern mining was published (Han et al. 2011). An algorithm, which makes use of the Info-Fizzy Network (IFN) so that association rules on adjacent intervals could be mined, was suggested by Maimon & Last (2013). Fuzzy theory assists in reducing the rule set. A general approach to knowledge discover in time series databases was also introduced. Data pre-processing, feature extraction, dimensionality reduction, prediction, and rule extraction are the stages of knowledge discovery data KDD process present in the approach. The extraction of a reduced set of linguistic association rules concerned with that behaviour and the forecasting of the future behaviour of time series are the primary aims of the process. Stocks data and weather data are the two real-world databases on which the approach is demonstrated. The identification of the most important features to be extracted from the raw data is enabled by the method; the information-theoretic method of data mining is applied to the pre-processed data set; and the bases of the cleaning and pre-processing of time series are signal processing techniques, are the certain features of the approach.

The detection of surprising patterns in linear space and time on the basis of a time series data base was made possible by the sequential patterns method suggested by Zolhavarieh et al. (2014). Domain independence is the basis of the method and

provided that more frequency is seen differ vastly from that expected given prior experience; flag patterns are going to be surprising. For pattern discovery in temporal data sequences, the self-organizing map (SOM) was employed which is a clustering approach for pattern discovery from time series. Through this special clustering algorithm, a topological structure is imposed on the data. With the utilization of a continuous sliding window, segmentation of data sequences was done from the numerical time series so that the SOM algorithm could be prepared. Afterwards, with the utilization of SOM, grouping of similar temporal patterns was done into clusters, through which different structures of the data or temporal patterns could be represented later on. The issue regarding representation of patterns in a mulit resolution manner has been tried to overcome. Exponential increase in the time required for the discover process takes place when the number of data points in the patterns (the length of patterns) is increased.

A more flexible model of asynchronous periodic patterns, whose presence may only be within a subsequence and whose occurrences may be shifted because of disturbance, was suggested by Getta & Zimniak (2015). The specification of the maximum allowed disturbance between any two successive valid segments and the minimum number of repetitions which is needed within each segment of non-disrupted pattern occurrences was done by the utilization of two parameters max-dis and min-rep. The longest valid subsequence of a pattern is returns prior to the satisfaction of these two conditions. For generating potential periods by distance-based pruning and afterwards for deriving and validating candidate patterns and locating the longest valid subsequence by an iterative procedure, a two-phase algorithm is devised. The ability of the suggested mining algorithm to discover every periodic pattern without taking the period length into account is one of its innovations. Mining of every pattern (1) by whose periods, a wide range can be covered and periods are not called a priori; (2) which is there within only a subsequence; and (3) because of inserting some random disturbance, whose occurrences may be misaligned, can be allowed by a more flexible model of asynchronous periodic patterns which has been suggested.

For finding the mining dependency between different time series, a technique was suggested by Ykhlef & Al-Reshoud (2009) which makes use of a genetic

algorithm and discretization. Real-life time series could react to the same conditions in a dependent way, though they could be totally dissimilar. Combination of different time series mining techniques, like discretization, classification of shapes, extracting association rules and genetic algorithm, was done. The determination of the dependency between two-time series was the key objective. A number of segments for both time series can be produced by using the discretization via this approach. Afterwards, classification of these segments is done into seven pre-defined segment classes. The generation of rules is done by a genetic algorithm. These rules can be taken as condition/action rules in which a segment from the second time series represents the action clause and a segment from the first time series represents the condition clause. Computation of the confidence regarding these rules is carried out and the only rules taken into consideration are the ones which meet minimum dependency support and confidence given by a user. Application of two-time series was done. A dependency was discovered between those time series as a set of rules between the submitted time series within each data set were mined by the suggested method, as demonstrated by the outcome of the experiments.

Pradhan and Prabhakaran (2009) pondered over the issue regarding mining association rules. An association is formed between the numerous instances of multiple time series data and different quantitative attributes, resulting in the formation of a multiple dimensional framework, when correlations or dependencies are found by time series pattern mining (TSPM) in the same series or in multiple time series. Real-life time series data of muscular activities of human participants acquired out of multiple Electromyogram (EMG) sensors were taken into account by Pradhan and Prabhakaran (2009). The discovery of patterns in these EMG data streams was attempted by them. There is a connection between every EMG data stream and quantitative attributes and onset time, whose requirement is there to be mined alongside EMG time series patterns. In addition, a two-stage approach was suggested for fulfilling the following aims: initially, the discovery of frequent patterns in multiple time series through sequential mining across time slices is the centralization. Secondly, quantitative attributes of only those time series which are there in the patterns discovered initially are focused. Large sets of time series data from multiple EMG sensors were used for the evaluation, because of which, it could be seen that it

scales up linearly in terms of the number of time series involved and the process of finding association rules in a multidimensional environment is made faster by the two-stage approach in comparison to other methods. Every multiple time series data set format can utilize this generic approach.

Pattern discovery in time series was the focus of Mohammad et al. (2012) and discussion regarding present algorithms took place. There cannot be utilization of domain knowledge in whatever manner through which quadratic or at least super-linear time and space complexity result, by almost all of the available algorithms. While searching, the definition of the Constrained Motif Discovery problem took place because of which utilization of domain knowledge into the motif discovery process can be enabled. To efficaciously solve the confined motif discovery issue, (MCFull and MCInc) suggested two algorithms. With the utilization of a change point detection algorithm, the conversion of most unconstrained motif discovery problems can be done into constrained ones, as shown by them. Mohammad and Nishida (2015) Afterwards, there was introduction of the Robust Singular Spectrum Transform (RSST), which is a novel change detection and comparison was done between it and conventional Singular Spectrum Transform making use of two time series data sets. Higher specificity is achieved by RSST and it is more sufficient to find constraints to the conversion of unconstrained motif discovery problems to constrained ones whose solution can be provided by MCFull and MCInc, There was a four to tenfold increase in speed in comparison to the unconstrained motif discovery algorithms studied, devoid of losing any accuracy, by the suggested algorithms, as showed by the results. Later on, for enabling the robot to learn free hand gestures, actions, and their associations by observing humans and other robots communicating, there was utilization of RSST + MCFull in a real-world human–robot interaction experiment.

However, in the data mining, frequent pattern algorithms are thoroughly considered problem with respect of computational and algorithmic development. Though in latest years many algorithms were suggested for settling the frequent pattern and its variants, however, the basic problems are still persistent (Lee et al. 2007; Zhang & Wang 2008). The frequent patterns algorithm in the huge datasets produces apply of the Apriori algorithm and it use in many applications in data mining

tasks such as climate change pattern discovery (Alshareef et al. 2016; Jaafar et al. 2016; Yalcin et al. 2016), loss-leader analysis cross-marketing, basket data analysis, catalogue design, sale campaign analysis, Web log analysis (Raval 2012), and DNA sequence analysis in (Alipanahi et al. 2015; Sanchez-Mut et al. 2016).

With respect to the pattern discovery models, many different models have been proposed for the frequent patterns approach. Therewith, the most popular one contains the support-based model. One of the main advantages of the support-based model is its probability to detect the frequency for the item sets having a frequency above the specified threshold. However this model suffers from a limitation that these item sets sometimes do not represent the interesting correlations present between the items as they cannot normalise the absolute frequencies of all the items. Hence, there are alternate measures for the interestingness which is defined in the studies (Srivastava et al. 2000; Aggarwal and Philip 2008; Aggarwal et al. 2010).

In the algorithms described previously, the candidate patterns have been produced from the earlier produced frequency patterns. Afterwards, the transaction database is applied for defining the candidate frequent patterns. The computational efficiency problems appear with regards to the discovery of the candidate patterns in the carefully-designed, orderly pattern, pruning the irrelevant, the repetitive candidates and picking out the appropriate tips for reducing the complicated work in the candidate calculating. The later algorithms applied the joins approach for the candidate generation procedure, that was the similar applied by the basic Apriori algorithm (Kamber and Han 2006; Aggarwal 2013).

2.7 DISCUSSION

This issue in different ways application can be addressed by various representing data techniques and methods, in mining time series data. Representations are important to reduce dimensionality and generate useful similarity measures. High-level representations such as Fourier transforms, wavelets, piecewise approximation techniques etc., were tried for the purpose of representing data in the finest form, and for easily dealing with unique data mining algorithms, like the clustering,

classification and discover patterns, transferring the time domain to the frequency domain. Nevertheless, the actual fields do not have the application of most of the representation techniques in the mining time series data, as suggested by modern studies and research. Furthermore, there still are limitations and requirements regarding better performance according to time-series data. The extraction of the very complex continuous data (time series data) is the motivation research. Mining time series are not appropriate for working with most of the classic machine learning algorithms. The challenge interesting research is the description of time series data by cases of high-dimensional, very high feature association and large amounts of noise correlation. Additionally, there is still a requirement of similarity measures, prediction and patterns discovery methods of time series to be enhanced.

In various disciplines, there has been suggestion of different techniques dealing with dimensionality reduction. With the utilization of techniques on the basis of linear reduction with numerical representation and linear reduction with symbolical representation, classification of the techniques can be done. The issue regarding the determination of parameter periods, the size of alphabet and the length of time series, is common in all of these techniques and the weakness of measuring the distance, a high dimensional reduction and representation are some of the restrictions present in parts of these techniques. SAX, having more benefits in comparison to other techniques; special mining in pattern discovery and prediction, is the best technique of symbolic representation present in the literature. The benefits and drawbacks of the suggested ways for representing the time series can be seen in Table 2.2.

Table 2.2 Summary of time series data representation technique

Time series representation techniques	**Authors and year**	**Advantages**	**Disadvantages**
DFT	Agrawal et al., (1993)	• The most natural algorithm • A continues manner. • Good for a spectral analysis • Fast indexing and signals • Good compression.	• Require parameters. • The parameters were not evident. • Apply clustering before discretizing. • Weak scale in massive data. • Very low to find amplitudes. • Not optimal for all series. • Compute the average each coefficient.

To be continued…

… Continuation

DWT	Chan and Fu (1999)	• A discrete manner. • Time and frequency analysis. • An efficacious reduction. • Need a few coefficients • An easier application.	• Apply for integral power of two is its length. • May not represent in denoting or thresholding. • Not optimal for all series.
SVD	Faloutsos (1996)	• Good compressed Signals • The most natural. • Fast calculate and indexing.	• Weakness and strength in terms of indexing point of view. • Complex manner in huge data. • Difficult to interpret.
PAA	Yi and Faloutsos (2001)	• Approximating manner • Prove lower bounds. • Well-defined and well-documented. • Efficient reduction. • Computationally efficient. • Support by distance measures.	• Query variable length. • Mean values-based of equal sized frames. • Consideration only the central tendency. • Not concern to dispersion section. • The scaling is larger, information loss increases. • Need optimal parameters • Weakness in huge data.
Clipping	Bagnall et al., (2006)	• Less memory to store TSs data. • High compression ratio • Increase tightness and lower bound. • Fast and significant reduction. • Efficiently in outliers issue • Good accuracy in clustering	• Dynamic compression ratio; user has no choice to make. • Difficult to interpret.
PIP	Fu and Chung (2008)	• Support by the SB-Tree. • Fast indexing and signals • Handle the incremental updating problem. • Easy reduction and preserve the shape of TSs • longest common use to compute the similarity.	• Time represent invariant to amplitude scaling and time warping. • Not optimal for all series • Complexity in computation and storage. • Weakness in huge data
ASCC	Li and Guo (2013)	• 4 features, average, slope, curvature and rate of change of the curvature. • Provide a fit reduction with lower error. • Lower bound on the ED.	• Excellent performance with no increase time complexity. • Equal length of subsequences.

For sequence of time series representation, there has been selection of SAX algorithm in this study. For measuring the similarity of sequence on the basis of symbolic representation, which that lower bound corresponding distance measures defined on the original sequence, SAX is distinctive representation of high dimensional reduction permits. The algorithms is especially ambitious for allowing

the application of data mining tasks on the basis of symbolic representation efficaciously and the sequence of real value was converted into symbolic sequence representation with an infinitesimal time and space by the algorithm. There is a connection between the original sequence representation and the produce representation of SAX sequence.

The enhancement of SAX representation method has been suggested by many studies. Further research can be conducted on the symbolic representation and making the efficacious performance of representation requires new enhancements, as demonstrated by (Lkhagva et al. 2006; Shieh & Keogh 2008; Ahmed et al. 2011; Li et al. 2012). For the representation method and alphabet and word size optimization, the suggested SAX algorithms can be found in Table 2.3.

Table 2.3 Summary of symbolic representation techniques based on piecewise approximation

Symbolic representation techniques	**Authors and year**	**Advantages**	**Disadvantages**
SAX	Lin et al., (2003)	• Stored as bits, give space-saving. • Higher dimensionality than others. • Use less or the same space. • 1 time the dimensionality.	• Fixing parameters. • Miss important patterns. • Pattern matching-based. • Distance is small, cannot find candidate subsequences easily. • Distance is large, many candidates with incorrect. • Need optimal parameters.
ESAX	Lkhagva et al., (2006)	• Less lose information. • Prefect with huge data. • A meaningful representation.	• Fixing parameters. • 3 times the dimensionality. • Noise presence in parameters. • Increase time complexity. • Need optimal parameters • Complexity in computation and storage.
iSAX	Shieh and Keogh (2011)	• Fast approximate search. • Efficiently in real-valued data. • Less storage.	• Low performance in I/O and CPU cost. • Complexity in computation. • Lack in distribution with large data. • Need optimal parameters • Height-unbalanced. • Difficult to interpret.

To be continued…

… Continuation

TSX	Li et al., (2012)	• Good resolution. • Prefect with huge data. • A meaningful representation. • Support in mining tasks. •	• Fixing parameters. • Need optimal parameters • A 4-tuple symbolic. • Ignore slope, depended on size segment. • Complexity in computation and storage.
HSAX	Alshareef et al. (2016)	• Harmony Search (HS) for finding the optimum word size and alphabet size of SAX Algorithm. • Optimal parameters supported.	• Difficult to interpret. • Fixing parameters. • Miss important patterns in a huge data. • Complexity in computation using optimization algorithm

The designing of these suggested algorithms is done in a way through which SAX representation of time series in reducing the full original set of time series to smaller data set series time could be enhanced. An instance of this is acquiring the word size through segment intervals of a time series. On the other hand, the amount of distinct values in time series can be reduced by alphabet size. SAX effort on the ways of reducing the time series has been attempted to be enhanced by some method, as demonstrated in Table 2.3. A high reduction in the time series data that avoids losing hidden important information is resulted by this approach. Therefore, for acquiring an efficacious representation of the data, there is requirement of a time-series data analysis. A part of the study centralizing on the ways of designing an algorithm for discovering most suitable SAX representation based aggregated methods for contained data sets is a problem. The discussion regarding these considerations is still to be done for proving suggested algorithms; frequently, basic information and hidden patterns could get destroyed and vital information can be lost as the researchers only have to enhance classification accuracy tries and compress data sets.

Therefore, for an effective distinct representation of the given data sets, time series data is required to be studied properly. Our study investigates this problem though this is not the primary concern of the research. Our research is primarily focused on techniques through which such an algorithm can be developed that can minimize the dimensionality of time series with little to no loss of information. Such concerns are needed to be examined in the proof of suggested algorithms. Some of the significant sensitive information is lost when the researchers compress data sets and

enhance the precision of classification. For instance, removal of such significant facts from a weather and financial application can lead to the obliteration of essential information and concealed patterns.

Many prediction techniques have proposed to handle processing of huge volume of data. It can predict categorical labels and predictors data based on model built by using training set and correlated labels and then can be used for predicting newly existing test data. Hence, it is defined as an integral unit of data analysis and is gaining further popularity. Prediction uses supervised learning approach. In supervised learning, a training dataset of records is available with associated labels (Meenakshi & Geetika 2014). Prediction techniques regularly find a rule or set of rules to represent data into classes (Zaki et al. 2014). Climate change institutions require rule(s) for making decision to predict time (day, month or year) into good or bad weather risks. Based on this decision weather can be assumed to specific time. The standard and well accepted algorithm for prediction task is initiation of Decision trees (DT) (Quinlan 1986; Zaki et al. 2014). Decision trees applies different measures such as Entropy, Gini index, Information gain etc.to find best split attribute. DT is very simple and fast to yield the accurate result. However DT has limitations such as it can have significantly more complex representation for some concepts due to replication problem (Bhavsar & Ganatra 2012), it requires large amount of memory to store tree (Nikam 2015) and it has a problem of over fitting (Brijain et al. 2014).

Other techniques for prediction task has the potential to significantly improve the conventional techniques in weather applications are SVMs and ANNs (Tiwary 2014). The SVM model using a sigmoid kernel function is equivalent to a two-layer, ANN. SVM models are a close cousin to classical multilayer perceptron neural networks. applying a kernel function, SVMs are an alternative training set for polynomial, radial basis function and multi-layer perceptron classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training (Lin & Wang 2002; Zhang et al. 2004). However, SVMs have a well performance, but it require speed and size either in training and further in testing. ANN is a mathematical model or computational model

that is inspired by the structure or functional aspects of biological neural networks. The model consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation (Fausett 1994). The disadvantages of ANN are: i) Requires extensive memory and high complexity to predict in many cases. ii) Involves high processing time in large data. iii) Difficult to identify number of layers and neurons are required. iv) Slow Learning (Nikam 2015).

Another popular and strong preformed technique which is depending in possibility theory is Naïve Bayes techniques. Naive Bayes technique is the Statistical Bayesian Predictor (Ro & Pe 1973). It is named Naive as it assumes that all variables contribute towards prediction and are mutually correlated. This assumption is called label conditional independence, this label may be considered as pattern feature (Bielza & Larrañaga 2014). It is also called Simple Bayes, Idiot's Bayes, and Independence Bayes. They possible to predict label membership probabilities, for example the probability that a given data item regards to a specific class label. A Naive Bayes Predictor considers that the presence (absence) of a specific feature of a class is unrelated to the presence (absence) of any other feature when the label variable is assumed. The Naive Bayes predictor technique is depending on Bayesian Theorem and it is used when the dimensionality of the feedbacks is high (Duda et al. 2012; Nikam 2015). The Naïve Bayes technique has been investigated for their performance based on different symbolic time series representation for use in weather applications.

Naïve Bayesian technique is suggested in this research, in order to find solution to a particular problem in a weather (rainfall and river flow data) application so that fascinating patterns in different year periods can be identified. We intend to become accustomed with the Naive Bayesian technique so that it can be used as a more precise detection technique than the prediction technique regarding river flow sequence and rainfall sequence. The significance of employing Naïve Bayesian for identifying problems in various applications has been mentioned in prior researches.

In spite of their naive strategy and apparently over-simplified traditions, Naive Bayes predictors have applied efficiently and superior in many complicated real-world states. Analysis of the Bayesian prediction issue has exposed that there are some

theoretic reasons for the apparently unreasonable effectiveness of Naive Bayes predictors (Tiwary 2014). An advantages and superiority of the naive Bayes classifier are that: i) It only requires a small amount of training data to evaluate the parameters necessary for prediction. Because independent variables are assumed, only the variances of the variables for each lebal need to be specified and not all the covariance matrix, i.e., it involves short computational time for training. ii) Great Computational efficiency and classification rate· iii) It enhanced the prediction performance by take away the irrelevant features. iv) The Naive Bayes predictor involves a very huge number of records to achieve good results, e.g., weather data. v) It has quite well performance (Jadhav & Channe 2016).

In the previous decades various pattern discovery algorithms have been identified. The frequent occurrence of serial events in time series is known as pattern. Patterns are a general type of data consisting of significant knowledge that is required to be identified. The primary concern is to effectively discover unidentified frequent patterns. These patterns can be used for developing constructive policies. Moreover, the fundamental features of a domain could be described using these patterns. Identification of frequent patterns from time series is an important undertaking in the process of data mining. Frequent patterns may offer much better understanding to experts in different fields. Such patterns have been very useful especially for alarm log analysis, stock trend relationship analysis and financial patterns (Verkamo 1995). Various studies have examined and discussed this problem.

This section analyzed the problem of pattern discovery which has two parts i.e. pattern detection and discovery of the change points in time series (segmentation). Through the literature, it is depicted that a few approaches attempt to segment the time series depending on any changes that take place with equal segment size, or else a few approaches focus on the change distribution of the time series, which is, the technique cuts the time series and begins with a new point as the distribution changes. This analysis revealed that more attention has given to the segmentation of time series. Various algorithms have been suggested based on Piecewise Linear Representation (PLR) such as Top-Down, Bottom-Up, and Sliding Window algorithms (Keogh et al. 2004). Based on the empirical research and testing of the performances of these