Prediction of Anomalies for a Dynamic Data System in an Optimized Approach

Nethravathi K
Research Scholar, Computer Science,
S.N.R Sons College, Coimbatore, India
nethra90.2009@Gmail.Com

Dr.Anna Saro Vijendran
Director, Department Of Computer Application,
S.N.R Sons College, Coimbatore, India
saroviji@Rediffmail.Com

Abstract— Events, as an application specific subset in the target sequence, are often closely related to certain time-ordered structures, called temporal patterns. Discovering such temporal patterns that are characteristic and predictive of the events is critical in many applications. Multivariate Reconstructed Phase Space (MRPS) embedding is proposed in the existing system, to detect multivariate predictive temporal patterns and explore their relationships with defined events in the target data sequence. This embedding creates a new feature space combining all the individual embedding of each variable sequence. This algorithm also provides a discriminative module that utilizes the Gaussian mixture model to score temporal patterns based on the posterior likelihood. In this thesis, proposing a new classifier and an associated objective function that integrate both temporal pattern modeling in MRPS and Bayesian discriminative scoring for temporal pattern classification in multivariate data sequences. Gaussian Mixer Model Classifiers are often affected by over fitting problems and the decision boundaries. To prevent this, proposed a new method which considers labeling errors independently of their distance to the decision boundaries. This is achieved by introducing binary latent variables which indicates a given instance to be an outlier (wrongly labeled instance) or not. The Bayesian inference is used to solve difficulty in learning problem. Model Evidence, Prediction and Outlier Identification are the three methods proposed to improve GMM classifier. In this thesis, includes a dual simplex optimization method to solve the complex event. The proposed objective method is an alternate optimization for the optimization objective function proposed in the existing work. From the experimentation result, the proposed system is well effective than the existing work.

Keywords- *Dual simplex optimization; Dynamic Data System; Gaussian Mixture Model; Multivariate Reconstructed Phase Space; Temporal Pattern.*

*****

I. INTRODUCTION

Data mining is the process of extracting or mining knowledge from large amount of data. It is an analytic process designed to explore large amounts of data in search of consistent patterns and systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. It can be viewed as a result of natural evolution of information in development of functionalities such as data collection, database creation, data management, data analysis. It is the process where intelligent methods are applied in order to extract data patterns from databases, data warehouses, or other information repositories. The data mining concept can be classified into two types: descriptive and predictive. Data mining systems can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used, or the applications adapted. This query language can be designed to support ad hoc and interactive data. The functionalities are concept and class descriptions, associations and correlations, classification and prediction, cluster analysis and outlier analysis. Concise and precise descriptions of a class or a concept are called concept and class description. Frequent patterns are the patterns that occur frequently in data. Mining frequent patterns lead to the discovery of interesting associations and correlations within data. Classification is the process of finding a model that describes and distinguishes data classes or concepts. The process of finding interesting, interpreted, useful and novel data from a large set of data is known as Knowledge Discovery in Databases (KDD). The steps involved in mining the data are as follows: Pre-processing, mine the data and interpret the results.

II. LITERATURE REVIEW

H. Wang, W. Fan, P.S. Yu, and J. Han, [13], proposes Data stream mining with drifts concept is challenging concept for fraud detection and all marketing intrusion[16]. In this, proposing a method for streaming data drifting. The data stream is divided into two parts training dataset and testing data set .the training data set is a model of classification that are the chunks of data. The testing data is used to find the prediction of accuracy of the ensemble classifier. In this paper proposing a method called weighted ensemble classifier for drifting the streaming data.

In the training data stream, the data is not consistent and timely manner. So presenting a refine models to handle drifting called classifier incremental decision tree is used to predict the data stream more accurately without loss in the data stream.

X. Feng and H. Huang [29], reports a new temporal pattern identification framework (called New Framework thereafter) significantly improves over the original TSDM framework. Events, as an application specific subset in the target sequence, are often closely related to certain time-

ordered structures, called temporal patterns. Discovering such temporal events are critical in many applications, such as determining the timing of positions of certain financial securities [8], forecasting economic growth [25], detecting a medical anomaly condition [23], and interpreting of the dynamics in the underlying system [14], [20]. Specifically, the evolutional patterns of GDP are used to address the characteristics of the economic growth [22].Although much attention has been on the research of time-varying data streams classification [3], [4], time series matching [2], [11], [12], clustering evolving stream data [27],[28], [29], and symbolization [6], much less research has addressed the problem of characterizing and predicting important events and anomalies in the dynamic data system.[5] The new objective function is the composition of event characteristic function and the membership function of a fuzzy set based temporal pattern vector in the time delay embedding phase space, and is more logical for defining the pattern of temporal events naturally. The new objective function is the composition of event characteristic function and the membership function of a fuzzy set based temporal pattern vector in the time delay embedding phase space, and is more logical for defining the pattern of temporal events naturally.

The center and radius of the temporal event cluster are calculated by the statistical mean and standard deviation of a continuously differentiable Gaussian-shaped membership function. An efficient two-step optimization strategy is proposed to search the optimal temporal pattern cluster in the phase space and make use of the mutual information and false neighbors methods to systematically estimate the time delay and the cluster region. Other advantages of New Framework include the computing efficiency and consistency of the search results. It should be stressed that our New Framework is not aimed to model or to "fit" all data points of the time series (i.e.) the resulting temporal pattern cluster is not targeted to include all temporal patterns. It rather intends to only identify those patterns that are significant in predicting future events with high accuracy and low false positive rates. In the Data Preparation Stage, data must be preprocessed and initial parameters should be selected. This includes: Dividing the entire time series into the training series and the testing series, defining the event characterization function g(Xt) , the objective function f(v; σ), and estimating the time delay τ and embedding dimension Q using nearest false neighbor algorithm and the mutual information algorithm, respectively. In the Training Stage, the following steps are presented. Embedding the training part of time series into the Q-dimensional phase space. Evaluating the event characterization function g(Xt) for each Xt RQ. Performing two-step optimization and filtering strategies to obtain optimal clusters. Evaluating performance of the resulting pattern clusters. During the testing, apply the temporal pattern clusters obtained in the training stage to identify or predict temporal events to the testing time series. An event Xt is predicted whenever a point $\mathbf{X}_t = (x_{t-(Q-1)\tau}, x_{t-(Q-2)\tau}, \cdots, x_t) \in R^Q$ is within the pattern cluster. After evaluating the testing results, a decision will be made whether or not the results are accepted, or the training stage is repeated till the results are satisfied

C.-H. Lee, A. Liu, and W.-S. Chen, "Pattern Discovery of Fuzzy Time Series for Financial Prediction"[7], In this knowledge based method in representing the financial time series and to equip. the knowledge discovery process of the time series [7]. The variance in the stock price is represented in fuzzy candlestick patterns which make the imprecise and huge investment knowledge computable, comprehensive visual and also editable. The rich information is carried by the fuzzy candle stick pattern to increase the efficiency of the knowledge discovery process of financial time series. To implement a system prototype to illustrate the usage of the fuzzy candle stick pattern the pattern construction and the recognition process are introduced. So that the investors can save and share their investment experience and also increase the efficiency of their investing strategies.

The fuzzy candlestick patterns carry rich information and can be used to increase the efficiency of the knowledge discovery process of financial time series. Pattern construction and the recognition process are introduced and implemented in a system prototype to illustrate the usage of the fuzzy candlestick patterns. Moreover, investors can save and share their investment experience. By reusing and modifying the stored candlestick pattern information, the investor can also increase the efficiency of their investing strategies. A system prototype, named Candlestick Tutor (CT) which was proposed, is adapted to realize the idea. The CT includes a graphical user interface (GUI) for displaying candlestick charts, a pattern authoring and acquiring tool for mining .A pattern recognizes the module and a pattern validation tool for checking pattern efficiency. The CPFA is implemented in the pattern authoring and acquiring tool to facilitate the pattern editing process. An information agent, a stock information database, and a fuzzy candlestick pattern database are also designed to support the system. The information agent connects to the Web site which provides the daily stock information for acquiring the stock information. For programming convenience, the system is coded in Java language. The NRC Fuzzy Toolkit provides a set of Java API which enhances the system with the capabilities of handling fuzzy concepts and reasoning.

The Research of time-varying data streams classifies [13], [14], time series matching [10], [5], [19],

clustering evolving stream data [17],[21], [18], and symbolization [13], has addressed the less problem of characterizing and predicting important events [5]. In other words, detecting usual temporal patterns in the data sequence, but more interested in identifying the patterns that are characteristic and predictive of important events in the target data sequence and exploring the causal relationships between the events and causing variables in the multivariate data sequence [12], [29],[24]. R.J. Povinelli and X. Feng, [24], introduces a new method for identifying temporal patterns in time series that are significant for characterizing and predicting events, i.e., the important occurrences. The new method called Reconstructed Phase Space method is capable of characterizing temporal patterns of complex time series, which are often no periodic, irregular, and chaotic. Reconstructed Phase Space method embeds the univariate data sequence into a multidimensional time-legged phase space with appropriate time delay and embedding dimension [25], [3], so that the temporal patterns in the data sequence can be well represented in the embedded Reconstructed Phase Space. Furthermore, the RPS method employs a numerical optimization method to search for temporal pattern clusters predictive of the events such that the event functions can be maximized. Since the Reconstructed Phase Space method is capable of representing temporal patterns of nonlinear dynamic sequence data that are typically chaotic and irregular [15], it has been applied in a variety of research fields, such as ECG signal pattern identification of the Arial electrophysiology [25], wilding droplet identification [24], and security index forecasting [1], [8] .First, the method focuses on the identification of the temporal patterns that are characteristic of the events. Second, with the temporal patterns identified, the new method focuses on event prediction rather than complete time series prediction. This allows the prediction of complicated time series events such as the release of metal droplets from a welder. Third, the objective function in the optimization reflects the goal of the time series being examined, i.e., droplet releases, and is problem specific.

Y. Quilfen, A. Bentamy, P. Delecluse, K. Katsaros, and N. Grima, "Prediction of sea level anomalies using ocean circulation model forced by scatter meter wind and validation using Topex/Poseidon data" [31] presents a further comparison of the ocean model results with the TOPEX-Poseidon altimeter measurements. Simulated and measured sea level variability is described over the three tropical oceans. The annual and semi-annual signals, as well as the inter annual variability, partly linked to the El Niño southern oscillation (ENSO) phenomenon, are well simulated by the OPA7 model when the satellite winds are used. Furthermore, it shows that the objective method kriging technique is used to interpolate the mean ERS wind fields, dramatically reduces the effects of the satellite band like sampling. In the last part of this work, focuses on the relationship between the wind stress anomalies and the sea level anomalies in the case of the 1997–1998 El Niño events. It clearly shows that sea level anomalies in the eastern and western parts of the Pacific are strongly linked to wind stress anomalies in the central Pacific. The forthcoming scatterometers aboard the METOP and ADEOS satellites will provide a much better coverage. It will enable the wind variability spatial and temporal scales to be resolved better, in order that wind uncertainties no longer blur the interpretation of ocean circulation numerical models results. This work, analyze ERS surface wind measurements to examine their adequacy to force the Oceanic Global Circulation Model (OGCM) OPA7.

Taking advantage of this wealth of surface wind measurements over a time period covering several annual cycles, mean weekly surface wind and wind stress fields were computed onto a 1 square grid using a statistical interpolation method. This temporal resolution was chosen because the ERS scatter meter coverage is not sufficient to resolve higher scales, but it allows sampling of important features of the tropical atmospheric variability, such as the so-called Madden-Julian oscillations or the wind bursts triggering the oceanic variability in the western Pacific. The 3-D OGCMOPA7was forced by the ERS winds over the time period 1992-1995, and the results are examined by comparison with the Tropical Atmosphere And Ocean (TAO) buoy array in the Pacific Ocean. This paper, presents a detailed description of the wind patterns and the associated simulated and measured sea level patterns, focusing on the tropical areas, where the response to the wind forcing and to the wind forcing errors are easier to quantify. Indeed, the tropical oceans are mainly wind-driven, and since the ocean response to the wind forcing is faster than in mid-latitudes, it is possible to analyze it over the time period (1992–1997) available. It presents a comparison between the scatter meter winds and the TAO winds in three key areas of the tropical Pacific Ocean, to assess the adequacy of these different data sources, making measurements at different spatial and temporal scales. Here analysing the measured and simulated sea level by means of an Empirical Orthogonal Function (EOF) analysis over the three tropical oceans.

Data Stream Clustering and Other Approaches Significant efforts have been devoted in detecting temporal patterns in the literature of sequential data mining and analysis. Among them Aggarwal, et al. [18][6] proposed a clustering method for clustering evolving stream data, while others in [31], [5], [7], [13], worked on a variety of applications using temporal pattern clustering and analysis. Although these methods have been demonstrated effectiveness in their specific application areas, the focuses were mostly on the detection and clustering of temporal patterns. Less attention was on the characterization and

 ISSN: 2321-8169
2711 – 2718

prediction of events in the data sequence. On the other hand, other multivariate approaches in the existing literature include Discrete Fourier Transform (DFT)[5] and Discrete Wavelet Transform (DWT) [19], [9],autoregressive modeling [11], nonlinear classification using neural networks [4], decision trees [16], to name a few. Since frequency domain approaches are based on spectral patterns of the feature space, data sequences with different nonlinear dynamic patterns but similar power spectrum may not be distinguished. Methods using Artificial Neural Network (ANN) may suffer from low generalize ability and explain ability. In [30], multiple patterns are used to detect patterns based on predictability measure validation. There have also been research works on finding patterns for anomaly prediction in large-scale cluster systems. In [28] the authors applied Markov chain models to capture changing patterns of system metrics and used a naive Bayesian classifier to learn system symptoms for predictive anomaly classification.

III. METHODOLOGY MRPS-GMM CLASSIFICATION IN PHASE SPACE

3.1 Data categorization

Consider the multivariate sequence. An (m + 1)-dimensional data vector xi = (x1i; x2i; . . . ; xmi; xei)^T is observed at each time i as a sample instance. The definition of event function depends on specific applications. One commonly used event function is a threshold function which can be defined as follows:

$$g(\mathbf{x}_t) = \begin{cases} +1 & \text{if } \max\{x_{e,}i+1,\ldots,x_{e,}i+k\} > c \text{ and } x_{ei} \le c \\ -1 & \text{if } \max\{x_{e,}i+1,\ldots,x_{e,}i+k\} \le c \text{ and } x_{ei} \le c \\ 0 & \text{if } x_{ei} > c, \end{cases}$$

Where k is the time-step ahead, c is the threshold above which an event occurs, and {xei; i = 1; . . .; N} denotes the target sequence with events that are of interest. The parameter k is problem-specific constant which specifies the maximum forecasting time horizon. For example, k is equal to 7 if we are interested in predicting the electric consumption of a customer over the next week horizon. The pattern vectors that are predictive of future events take positive values of event function, whereas normal state vectors take negative values. Therefore, in training stage, we can categorize multivariate data sequences into event, pattern or normal state at each time i. In general application, it is meaningful to predict events or classify patterns when the underlying system is not in event state. Thus, for detecting predictive patterns, our focus is primarily on classification of two categories of data, which are sequences in pattern and normal state according to the definition of event function g(xi) = {+1; -1} for each time i.

3.2 Multivariate Phase Space Embedding

Consider general multivariate data sequences with m causing variables Xj = {xji; i = 1; . . .; N}, with one target sequence Xe = {xei; i = 1; . . .; N} or an aggregated composite of multiple sequences, where j = 1; . . .; m. Denote xi = (x1i; x2i; . . . ; xmi; xei)^T as an observation at time i, the observation data matrix is represented by

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \ddots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mN} \end{bmatrix}_{m\times N}.$$

For each sequence, the resulting embedding for each sequence becomes

$$\mathbf{x}_{ji} = [x_{ji} \quad x_{j,i-\tau_j} \quad \cdots \quad x_{j,i-(Q_j-1)\tau_j}]$$

Where i = 1; 2; . . .; N; j = 1; 2; . . .; m + 1.The multivariate phase space embedding can then be constructed as

$$\mathbf{X}_i = (\mathbf{x}_{1i}, \ \mathbf{x}_{2i}, \ldots, \mathbf{x}_{ji}, \ldots, \mathbf{x}_{mi}, \mathbf{x}_{e,i})$$

at each time i, where xji represents the phase space embedding for jth variable xj with the time delay τj and dimension Qj at time i. The dimension Q of the multivariate embedding is the sum of each embedding dimension Q_j, $Q = \sum_j Q_j$.

3.3 Reconstructed Phase Space with Transformation

In many applications, due to the time-varying nature of the system, similar temporal patterns may appear at different time instances and therefore have different starting values or local trends. Since structural similar patterns represent the same dynamics, they should be considered as the same category of patterns. For traditional RPS embedding methods, different starting values or local trends can cause separation of structural similar temporal patterns into different regions of phase space. To address this problem, we consider applying a linear transform on phase space embeddings. As a result, the new phase space gives a detruded representation of the data sequence. According to the theorem of filtered delay embedding prevalence in Sauer et al., given a linear constant transformation, the resulting filtered delay mapping also gives a valid embedding of the underlying dynamic system. The Euclidean distance in the new phase space, therefore, can better measure the structural similarity between temporal patterns than in a traditional phase space.

$$\mathbf{x} = \{x_{t-(Q-1)\tau}, x_{t-(Q-2)\tau}, \ldots, x_t\},$$
$$\mathbf{y} = \{y_{t-(Q-1)\tau}, y_{t-(Q-2)\tau}, \ldots, y_t\},$$

And the similarity of temporal structures between two sequences can be measured by

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{Q} (x_{t-(Q-i)\tau} - y_{t-(Q-i)\tau} - d_0)^2$$

Where $d_0 = (x_{t-(Q-1)\tau} - y_{t-(Q-1)\tau})$ is the initial difference of two sequences.

3.4 Gaussian Mixture Model Classification

As mentioned in previous, the data sequences are separated by three states: normal, pattern, and event state.

Local temporal structures and statistical property of explaining variables can be viewed as two sets of features in classifying patterns in complex dynamic data sequence. Previous work under the RPS framework employed the clustering method to identify temporal patterns. However, few discussions were made to apply discriminative approach characterizing patterns based on statistical correlations for a multivariate data system. The approach proposed here applies a multivariate Gaussian mixture model to exploit the discriminative information that can be incorporated in the design of classifier, as well as in the optimization. A data instance xt can be considered belong to pattern state, normal state or event state according to event function. Hence, data sequence can be considered as mixture of three classes of variables, which represent recurring states. If all the data sequences including event sequence are treated as observations, the resulting multivariate vector at each time t can then be represented by xt = {x1t; x2t; . . . ; xmt; xet} with dimension m + 1.

3.5 Test process

In the testing stage, we apply the predictive pattern classifier obtained in the training stage to predict the events in the target sequence. At each time t, based on the classification decision, a forecast will be made whether or not an event will occur.

3.6 Bayesian inference

Bayesian inference is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is acquired. Bayesian updating is an important technique throughout statistics, and especially in mathematical. Bayesian inference is closely related to discussions of subjective probability often called "Bayesian probability.Bayesian inference computes the posterior probability according to Bayes rule:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

- Where denotes a conditional probability; more specifically, it means *given*. Stands for any *hypothesis* whose probability may be affected by data (called *evidence* below). Often there are competing hypotheses, from which one chooses the most probable.
- The *evidence* E corresponds to new data that were not used in computing the prior probability. $P(H)$, the *prior probability*, is the probability of H *before* E is observed. This indicates one's previous estimate of the probability that a hypothesis is true, before gaining the current evidence.
- $P(H \mid E)$, the *posterior probability*, is the probability of H *given* E, i.e., *after* E is observed. This tells us what we want to know: the probability of a hypothesis *given* the observed evidence.
- $P(E \mid H)$ Is the probability of observing E *given* H. As a function of E with H fixed, this is the *likelihood*. The likelihood function should **not** be confused with $P(H \mid E)$ as a function of H rather than of E. It indicates the compatibility of the evidence with the given hypothesis.
- $P(E)$ is sometimes termed the marginal likelihood or "model evidence".

3.8 Performance evaluation

Accuracy rate

Accuracy is defined as the overall accuracy rate or classification accuracy and is calculated as

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

True Positive rate

True Positive rate (TP rate), also called sensitivity or recall, is the proportion of actual positives which are predicted to be positive and is calculated as

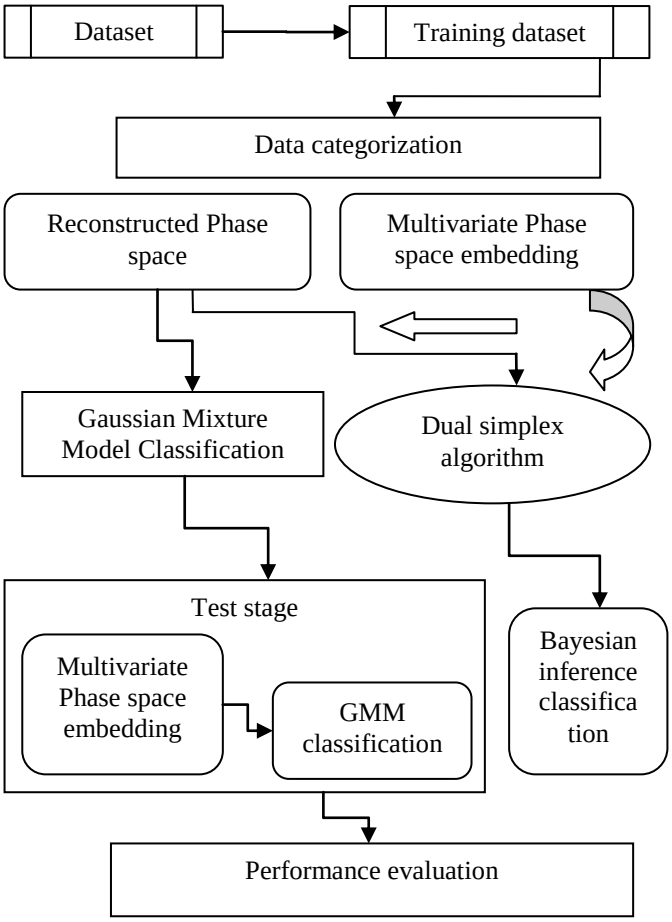
$$\text{TP} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

True Negative rate

True Negative rate (TN rate), or specificity, is the proportion of actual negatives which are predicted to be negative and is calculated as

$$\text{TN} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

IV. 4.ARCHITECTURE DIAGRAM



 ISSN: 2321-8169
2711 – 2718

V. RESULTS AND DISCUSSION

In this section, we are comparing the performance of the existing system such as MRPS-GMM classification with proposed system i.e., Bayesian inference in terms of classification accuracy, True Positive rate (TP rate) and True Negative rate (TN rate). To assess efficiency, we measured these comparison parameters for proposed system. From the end of this experimentation section, we can say that the proposed system has higher efficiency than the other techniques.

Accuracy rate

Accuracy is defined as the overall accuracy rate or classification accuracy and is calculated as

$$\text{Accuracy} = \frac{\text{True positive} + \text{True nega tive}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

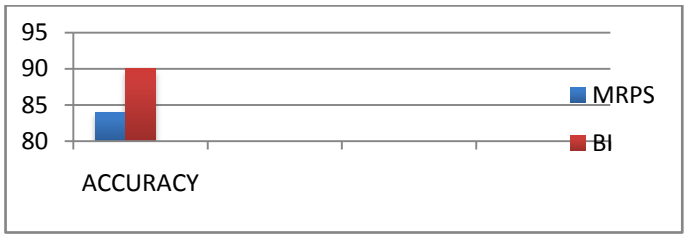


Fig.1. showed that comparison of the accuracy parameter between the existing system such that MRPS-GMM classification with proposed system i.e., Bayesian inference.

Accuracy rate is mathematically calculated by using formula. As usual in the graph X-axis will be methods (existing and proposed system) and Y-axis will be accuracy rate. From view of this accuracy comparison graph we obtain conclude as the proposed system has more effective in accuracy performance comparatively.

True Positive Rate

True Positive rate (TP rate), also called sensitivity or recall, is the proportion of actual positives which are predicted to be positive and is calculated as TP =

$$\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

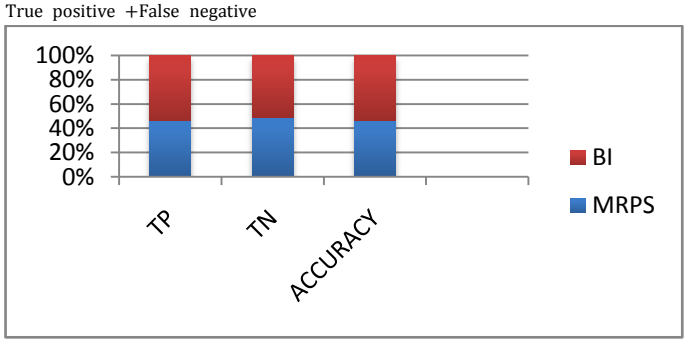


Fig.2. showed that comparison of TP parameter between the existing system such that MRPS-GMM classification with proposed system i.e., Bayesian inference.

TP rate is mathematically calculated by using formula. As usual in the graph X-axis will be methods (existing and proposed system) and Y-axis will be TP rate. From view of this TP comparison graph we obtain conclude as the proposed system has more effective Min TP performance. .

True Negative Rate

True Negative rate (TN rate), or specificity, is the proportion of actual negatives which are predicted to be negative and is calculated as TN =

$$\frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

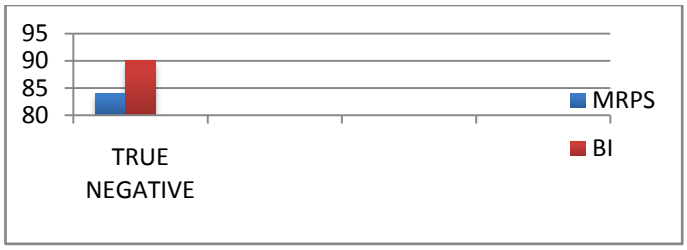


Fig.3. Showed that comparison of TN parameter between the existing system such that MRPS-GMM classification with proposed system i.e., Bayesian inference.

TN rate is mathematically calculated by using formula. As usual in the graph X-axis will be methods (existing and proposed system) and Y-axis will be TN rate. From view of this TN comparison graph we obtain conclude as the proposed system has more effective in TP performance comparatively.

TABLE 1: PERFORMANCE TABLE

METHOD	TRUE POSITIVE	TRUE NEGATIVE	ACCURACY
MPRS	83	87	82
BI	92	93	95

VI. CONCLUSION

A novel Multivariate Reconstructed Phase Space-Gaussian Mixture Model method is proposed for identifying temporal patterns predictive of events in a multivariate data system. However, in this existing work many disadvantages and limitations are there. Some of them are Gaussian Mixer Model Classifiers are often affected by over fitting problems while labeling and errors occur far from the decision boundaries. To prevent this, a new method is considered for labeling errors independently to the decision boundaries. This is achieved by introducing binary latent variables that

indicate a given instance to be an outlier (wrongly labeled instance) or not. The Bayesian inference is used to solve difficulty in learning problem. In proposing algorithm a new classifier and an associated objective function that integrate both temporal pattern modeling in MRPS and Bayesian discriminative scoring for temporal pattern classification in multivariate data sequences. The experimentation result shows that the proposed system has higher classification accuracy compared to the existing system.

VII. FUTURE WORK

Future scope of this thesis is to enhance more complex event function for different applications such as credit card fraud detection, health insurance fraud detection. One of fraud detection can be seen in all insurance types including health insurance. It is estimated that approximately $700 billion is lost due to fraud, waste, and abuse in the US healthcare system. Medicaid has been particularly susceptible target for fraud in recent years, with a distributed management model, limited cross program communications, difficult-to-track patient's population of low-income adults, their children, and people with certain disabilities.

REFERENCE

[1] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Soc., Series B, vol. 39, no. 1, pp. 1-38, 1977.

[2] A.G. Capodaglio, H. Jones, V. Novotny, and X. Feng, "Sludge Bulking Analysis and Forecasting: Application of System Identification and Artificial Neural Computing Technologies," Water Research, vol. 25, no. 10, pp. 1217-1224, 1991.

[3] A.M. Fraser and H.L. Swinney, "Independent Coordinates for Strange Attractors from Mutual Information," Physical Rev. vol. 33, no. 2, pp. 1134-1146, 1986.

[4] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, "Feature-Based Classification of Time-Series Data," Int'l J. Computer Research, pp. 49-61, 2001.

[5] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 419-429, 1994.

[6] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A Framework for Clustering Evolving Data Streams," Proc. 29th Int'l Conf. Very Large Data Bases, pp. 81-92, 2003.

[7] C.-H. Lee, A. Liu, and W.-S. Chen, "Pattern Discovery of Fuzzy Time Series for Financial Prediction," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 5, pp. 613-625, May 2006.

[8] C.M. Bishop, Pattern Recognition and Machine Learning. Springer, 2007.

[9] D.B. Percival and A.T. Walden, Wavelet Methods for Time Series Analysis. Cambridge Univ. Press, 2000.

[10] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "An Online Algorithm for Segmenting Time Series," Proc. IEEE Int'l Conf. Data Mining, pp. 289-296, 2001.

[11] G. Box and G. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day, 1976.

[12] H. Kantz and T. Schreiber, Nonlinear Time Series Analysis. Cambridge Univ. Press, 1997.

[13] H. Wang, W. Fan, P.S. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 226- 235, 2003.

[14] H. Wang, J. Yin, J. Pei, P.S. Yu, and J.X. Yu, "Suppressing Model Overfitting in Mining Concept-Drifting Data Streams," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 736-741, 2006.

[15] J. Lin, E.J. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a Novel Symbolic Representation of Time Series," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 107-144, 2007.

[16] J.J. Rodriguez and C.J. Alonso, "Interval and Dynamic Time Warping-Based Decision Trees," Proc. ACM Symp. Applied Computing, pp. 548-552, 2004.

[17] J. Yang and J. Leskovec, "Patterns of Temporal Variation in Online Media," Proc. Fourth ACM WSDM Int'l Conf. Web Search and Data Mining, pp. 177-186, 2011.

[18] J. Gama, Knowledge Discovery from Data Streams. Chapman & Hall/ CRC, 2010.

[19] K. Chan and A. Fu, "Efficient Time Series Matching by Wavelets," Proc. IEEE Int'l Conf. Data Eng., pp. 126-133, 1999.

[20] K. Sternickel, "Automatic Pattern Recognition in ECG Time Series," Computer Methods and Programs in Biomedicine, vol. 68, no. 2, pp. 109-115, 2002.

[21] K. Sim, A.K. Poernomo, and V. Gopalkrishnan, "Mining Actionable Subspace Clusters in Sequential Data," Proc. 10th SIAM Int'l Conf. Data Mining, pp. 442-453, 2010.

[22] L. Pritchett, "Understanding Patterns of Economic Growth: Searching for Hills among Plateaus, Mountains, and Plains," World Bank Economic Rev., vol. 14, no. 2, pp. 221-250, 2000.

[23] M.B. Kennel, R. Brown, and H.D. Abarbanel, "Determining Embedding Dimension for Phase-Space Reconstruction Using a Geometrical Construction," Physical Rev. vol. 45, no. 6 pp. 3403- 3411, 1992.

[24] R.J. Povinelli and X. Feng, "A New Temporal Pattern Identification Method for Characterization and Prediction of Complex Time Series Events," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 339-352, Feb. 2003.

[25] R.J. Povinelli, M.T. Johnson, A.C. Lindgren, and J. Ye, "Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 6, pp. 779-783, June 2004.

[26] T.K. Moon, "The Expectation-Maximization Algorithm," IEEE Signal Processing Magazine, vol. 13, no. 6, pp. 47-59, Nov. 1996.

[27] W. Zhang, X. Feng, and N. Bansal, "Detecting Temporal Patterns Using RPS and SVM in the Dynamic Data

Systems," Proc. IEEE Int'l Conf. Information and Automation, pp. 209-214, 2011.

[28] X. Gu and H. Wang, "Online Anomaly Prediction for Robust Cluster Systems," Proc. IEEE Int'l Conf. Data Eng., pp. 1000-1011, 2009.

[29] X. Feng and H. Huang, "A Fuzzy-Set-Based Reconstruction Phase Space Method for Identification of Temporal Patterns in Complex Time Series," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 601-613, May 2005.

[30] X. Feng and O. Senyana, "Mining Multiple Temporal Patterns of Complex Dynamic Data Systems," Proc. IEEE Symp. Computational Intelligence and Data Mining, pp. 411-417, 2009.

[31] Y. Quilfen, A. Bentamy, P. Delecluse, K. Katsaros, and N. Grima, "Prediction of Sea Level Anomalies Using Ocean Circulation Model Forced by Scatterometer Wind and Validation Using TOPEX/Poseidon Data," IEEE Trans. Geoscience and Remote Sensing, vol. 38, no. 4, pp. 1871-1884, July 2000.