CHAPTER II

REVIEW OF LITERATURE

Similarity search over time series data has been studied for many years. Similarity pattern search in time series was attempted initially by Agarwal et al. (1993) proposed an indexing method for time sequences to process similarity queries by transforming the time series from time domain to frequency domain using discrete Fourier transform. Agarwal et al. (1995) introduced a model of similarity for time sequences that captures the intuitive notion that two sequences should be considered similar if they have enough non-overlapping time ordered pairs of subsequences that are similar. Two subsequences are considered similar if one can be enclosed within an envelope of a specified width drawn around the other. Amplitude scaling and offset adjustments are performed to compare two sequences.

Xiao-Li Dong et al. (2006) presented a shape-based discrete symbolic representation and a novel distance measure that was used to measure the similarity between two time series. Eamonn Keogh et al. (2002) introduced a technique that defines a pattern surprising if the frequency of the pattern differs substantially
from that expected by chance. The proposed technique adapts suffix tree to efficiently encode the frequency of all observed patterns and allows a Markov model to predict the expected frequency of previously unobserved patterns. Yong Shi et al. (2005) proposed a data pre-processing technique called shrinking that optimizes the inherent characteristic of distribution of data. This data reorganization concept can be applied to several fields such as pattern recognition, data clustering, and signal processing. Shrinking approach for multidimensional data analysis consists of three steps: data shrinking, cluster detection, and cluster evaluation.

Laxman and Sastry (2006) carried out a survey of temporal data mining in which they have provided an overview of data mining tasks such as prediction, classification, clustering, search and retrieval, and pattern discovery. However this work mainly focuses on the developments in the research on similarity search in time series databases. Dimensionality reduction plays a vital role in enhancing the speed of operation during similarity search.

Batyrsin et al. (2008) proposed an architecture of perception based decision making system in time series database. This architecture works by integrating perception based time series data mining, and expert knowledge. Francisco et al. (2008) introduced a method for supervised discretization based on interval distances by using a concept of neighborhood in the target’s space. Chan et al. (2003) used the Haar wavelets with and without time warping for efficient similarity search in time series.
Khanh et al. (2008) presented boundedness for improving performance in approximation search and other well-established criteria for dimensionality reduction methods.

Agarwal et al. (1993) proposed a signature based technique for similarity-based queries, where approximate matching was carried out for a whole sequence using key words called signatures. As an alternate method, Keogh and Pazzani (1999) represented the time series as a piecewise linear segments for accurate classification and clustering in a relevance feedback framework. The results were further improved by Scargle et al. (2003) for the signals in multidimensional spaces using adaptive piecewise constant method. Tak-chung Fu et al. (2008) represented a financial time series according to the importance of the data points and proposed a method for evaluating the data point importance. In the concept of data point importance, a tree data structure called Specialized Binary tree (SB-tree) is proposed to represent the time series. Also, based on the order of importance, an access method for retrieving the time series data point from the tree is discussed. To evaluate the Perceptually Important Points (PIP) they introduced three measures namely PIP-Euclidean distance, PIP vertical distance, and PIP perpendicular distance.

Yan-Ping Huang et al. (2006) proposed an approach to find novel pattern in mining architecture with mixed attributes. This architecture combines the extended visualization-induced self organizing map algorithm and the extended naive Bayesian algorithm.
Their approach uses a systematic approach to deal with missing financial database information. Wang and Megalooikonomou (2008) proposed a piecewise vector quantization for dimensionality reduction. The extended version of PCA called Vector Quantization (VQ) can also be used to reduce the dimension of time series. In vector quantization, long time series is effectively represented with a much lower dimensionality where segments are represented by symbols or code words. This symbolic representation potentially allows for the application of text based retrieval techniques into the time series similarity analysis. VQ is widely used in signal compression and coding. Recently, it has been applied in time series forecasting and segmentation. According to Shannon’s theory, coding systems can perform better if they operate on vectors or groups of symbols rather than on individual symbols.

Chim and Deng (2008) proposed a phrase based document similarity to compute the pair-wise similarities of documents based on the suffix tree document model. This can be easily extended to the vector quantized time series. Durga Toshniwal and Joshi (2005) proposed a method for similarity search in time series using time weighted slope. This technique is based on the intuition that similar time sequences will have similar variations in time series. Faloutsos et al.(1994) proposed a frequency based signature technique, where approximate matching was done for a whole sequence using key words called signatures. Signatures are scanned sequentially and matched against the signature of the given query. And also proposed a
similarity measure based on signatures, extracted from the time series.

Hyo-Sang Lim et al. (2008) proposed a similar sequence matching method that efficiently supports variable-length and variable-tolerance continuous query sequences on time-series data stream. In case of variable-length query sequences, a window construction mechanism is used to divide long sequence into smaller windows for indexing and searching. For the variable-tolerance query sequences, notion of interleaved sequences whose individual entries are an interval of real numbers rather than actual real number is proposed. Rawshan Basha et al. (2007) proposed a method for the detection of time series discord subsequence and especially those with periodicity. Further, an entropy based measure is introduced as an alternative to the Euclidean distance measure for identifying discord subsequence. So far the work related to the similarity search in time series data has assumed that both the series have the same length. However, there may be some occasions where the two series may have different lengths.

In case of variable-length query sequences, a window construction mechanism is used to divide long sequence into smaller windows for indexing and searching. For the variable-tolerance query sequences, notion of interleaved sequences whose individual entries are an interval of real numbers rather than actual real number is proposed. Lian and Chen (2008) proposed three approaches namely Polynomial, Probabilistic and DFT to predict the unknown values of
stream data that have not arrived at the system and answer similarity queries based on the predicted data. And also used multidimensional hash index and B+ tree, to facilitate the prediction and similarity search of future time series. Hui Zhang and Tu Bao Ho (2006) proposed an unsupervised feature extraction algorithm using orthogonal wavelet transform for automatically choosing the dimensionality of features. The feature extraction algorithm selects the feature dimensionality by leveraging two conflicting requirements, i.e. lower dimensionality and lower sum of squared errors between the features and the original time series. Time series can also be transformed to text by using vector quantization.

Wan (2007) proposed a document similarity measure based on Earth Mover’s Distance EMD to search similar documents from a set of documents. Rubner et al (2000) computed the Earth Mover’s Distance (EMD) by solving the transportation problem.

Assent et al. (2008) introduced earth mover’s distance which was originally developed by Peleg et al. (1989). In general, the transportation problem deals with finding the least expensive flow of goods from the suppliers to the consumers that satisfies the consumer’s demand. However, in case of EMD, the distance between two histograms is defined as the minimum cost required moving pixels of one bin to other bin of first signature so that the resulting signature is same as second one. Hyunjin Yoon et al. (2005) proposed a family of novel unsupervised methods for Feature Subset Selection (FSS) from Multivariate Time Series (MTS) based on common principal component
analysis. Traditional FSS techniques have been applied to MTS data sets. The patterns in a time series are not always static. In some occasion, users may be in search of objects with specific movement patterns such as going up, going towards south west or a combination of these. Movement pattern queries ask for moving objects which show a given movement pattern in a specific scale.

Povinelli and Xin Feng (2003) proposed a method for analyzing data that employs time-delayed embedding and identifies temporal pattern in the resulting phase spaces. An optimization method is applied to search the phase spaces for optimal heterogeneous temporal pattern clusters that reveal hidden temporal patterns, which are characteristic and predictive of time series events.

Sung-Hyuk et al. (2002) proposed a similarity measure that uses the histogram of the time series. The distance between two histograms can be measured using vector and probabilistic methods. The probabilistic approaches use the probability density function and the distance between two probability density functions is same as Bayes probability. It is similar to Bhattacharyya distance and Kullback Leibler distance. Chan and Fu (1999) proposed a method for matching two time series by transforming the time series into frequency domain using wavelet transformation.

Franky Kin-Pong Chan et al. (2003) applied the Haar wavelets along with time warping for efficient similarity search in time series. Selvakumar and Senthamarai Kannan (2010) introduced for similarity
search for two time series data by adopting piecewise constant approximation method for dimensionality reduction. The k-medoid cluster analysis was carried out for the similarity pattern of the data. Also, the Markov model was proposed for predicting the future values.

Fu-Lai Chung et al. (2004) proposed an evolutionary time series segmentation algorithm that allows a sizeable set of pattern templates to be generated for mining. Fabian Morchen, (2008) adapted a unified view of temporal concepts and data models to categorize existing approaches for unsupervised pattern mining from symbolic data. And also applied distinguishing time point based methods and interval based methods. Xiaoyan Liu et al. (2008) proposed a segmentation criterion and developed two online piecewise linear segmentation methods based on it. They address the two basic requirements such as representation quality and computing efficiency of the piecewise representation of the time series by introducing the following criterions:

(i) Minimizing the representation error owing to approximation by $K$ segments.

(ii) Minimizing the number of segments such that the representation error of any segment is less than $\text{max}_\text{error}$, and

(iii) Minimizing the number of segments such that the total representation error of all segments does not exceed $\text{max}_\text{total_error}$, where $K$, $\text{max}_\text{error}$, and $\text{max}_\text{total_error}$ are predefined parameters.
In addition they introduce two online piecewise linear segmentation methods viz. feasible state window method and stepwise feasible space window method.

Piecewise Constant Approximation (PCA) also known as Piecewise Aggregate Approximation (PAA) is an approach for efficient dimensionality reduction in time series. This approach split the given time series into $m$ segments of equal length and the mean value of the segment is used to represent the series. There is always a trade-off between the accuracy of the result and the selection of the value of $m$. Therefore, the value $m$ has an impact directly on the accuracy. PCA transformation has several advantages. The resultant signature of the PCA can be used with arbitrary $L_p$ norms and the index can be built in linear time.

Gullo et al. (2009) adapted the Derivative time series Segment Approximation (DSA) representation model. It originally features derivative estimation, segmentation, and segment approximation. Segment approximation provides high sensitivity in capturing the main trends of time series and data compression. Selvakumar and Senthamarai Kannan (2010) introduced derivative segment approximation and Von Neumann ratio test for reducing the dimensions and similarity can be found by measuring the distance between the time series. Selvakumar et al. (2010) proposed the similarity distance between two images is found using nonparametric tests. Also composite similarity measure that comprises of earth mover’s distance is proposed. Balamurugan et al. (2011) proposed
Piecewise Constant Approximation Autocorrelation (PCAA) and k-means clustering which are used to approximate the time sequence and earth mover’s distance was proposed for measuring the distance between two time series.

Periodicity mining is a tool that helps in predicting the behavior of the time series data. Christos Berberidis et al. (2002) proposed a candidate periods featured in a time series that satisfy a minimum confidence threshold by using the autocorrelation function and Fast Fourier Transform (FFT). Elfeky et al. (2005) defined two types of periodicities namely segment periodicity and symbol periodicity.

Berberidis et al. (2002) introduced weak periodic signals in time series databases when no period length is known in advance. Lavinia Egidi and Paolo Terenziani (2003) proposed high level symbolic language for representing user defined periodicity which seems more human oriented than mathematical ones. And also a methodology for designing symbolic languages based on a preliminary analysis of the required expressiveness.

mining asynchronous periodic patterns. Winarko et al. (2008) proposed a signature based indexing method to optimize the storage and retrieval of a large collection of relative temporal patterns.

Kuo–Yu Huang and Chia-Hui Chang (2005) proposed a general model of asynchronous periodic patterns from a sequence of symbol sets where a time slot can contain multiple events. Three parameters, namely min-rep, max-dis and global-rep are employed to specify the minimum number of repetitions required for a valid segment of non-disrupted pattern occurrences, the maximum allowed disturbance between two successive valid segments, and the total repetitions required for a valid sequence. Aref et al.(2004) proposed online adaptation of the thresholds in order to produce interactive mining of partial periodic patterns.

Selvakumar and Senthamarai Kannan (2011) proposed piecewise aggregate approximation and autocorrelation techniques used for reducing the dimensionality of the data. K-means cluster technique for transforming the data. Also convolution algorithm was proposed for detecting periodicity patterns of the time series data.

Outlier detection is a primary step in many data-mining applications. Irad Ben-Gal (2005) presented several methods for outlier detection, while distinguishing between univariate against multivariate techniques and parametric against nonparametric procedures. Outlier detection for data mining is often based on distance measures, clustering and spatial methods. The outlier
identification problem is then the problem of identifying those observations that lie in a region so-called outlier.

Lian et al. (2009) presented a new definition for outliers, namely cluster-based outlier, which is intuitive and meaningful. This approach is based on a key observation that many abnormal events have both temporal and spatial locality, so a small cluster might also be an outlier. Also a new definition for outliers: cluster-based outlier was presented, which is meaningful and provides importance to the local data behavior, and how to detect outliers by the clustering algorithm. As the collection of moving object data become much easier, event-based outlier detection such as congestion in trajectory data are becoming increasingly attractive to data mining community. Most of the existing methods only perform the trajectory outlier detection on the spatial information.

Shurya Nakai et al. (2002) introduced unsupervised classifications for principal component similarity (PCS) and cluster analysis compared for outlier detection. The principal component similarity is new as it is different from the currently most popular technique for identifying the outliers. For verifying the outliers thus obtained, random-centroid optimization was applied for selecting the best samples by each cluster. This combination of principle component similarity and random-centroid optimization may be useful for consumers in transportation problem by creating food products to correctly respond to the demands of different consumer groups, which can assign each new event to a cluster to determine its type.
Xia Ying et al. (2009) framework for congestion outlier detection with clustering method was proposed. Trajectory data are analyzed according to both temporal and spatial factors by introducing the concept of minimal bounding boxes (MBBs), and super dense clusters are regarded as congestion outliers. George Kollios et al. (2003) investigated the use of biased sampling according to the density of the data set to speed up the operation of general data mining tasks such as clustering and outlier detection in large multidimensional data sets.

Peng Yang (2005) proposed a K-Nearest Neighbour (KNN) based outlier detection algorithm, it partitions the dataset into several clusters and then in each cluster, it calculates the K nearest neighbourhood for object to find outliers. The new method excludes outlier points by giving them extremely small membership values in existing clusters while fuzzy c-means algorithm tends give them outsized membership values. This algorithm also incorporates the positive aspects of k-means algorithm in calculating the new cluster centers in a more efficient approach than the c-means method.

Binu Thomas and Raju (2009) introduced outlier detection can lead to discovering unexpected and interesting knowledge, which is critically important to some areas such as monitoring of criminal activities in electronic commerce, credit card fraud, and the like. Also, proposed an efficient outlier detection method with clusters as by-product, which works efficiently for large datasets. Fahim et al. (2008) introduced a competent procedure to overcome
this problem. The proposed method is based on shifting the center of the large cluster toward the small cluster, and re computing the membership of small cluster points as considered as outliers.

Fawaz et al. (2008) introduced the distance of a point from its k nearest points (or neighbours) is calculated. If the neighbouring points are relatively close, then the point is considered normal; however, if the neighboring points are far away, then the point is considered outlier. Belal and Zoubi (2009) proposed method based on clustering approaches for outlier detection is presented. Initially, Partition Around Medoid (PAM) clustering algorithm was performed. Small clusters are then determined and considered as outlier clusters. Data clustering is an important data exploration technique with many applications in data mining. The k-means algorithm is well known for its efficiency in clustering large data sets. However, this algorithm is suitable for spherical shaped clusters of similar sizes and densities. The quality of the resulting clusters decreases when the data set contains spherical shaped with large variance in sizes.

Henley et al. (1996) introduced the k-nearest-neighbour (k-NN) method; this technique is the standard technique in pattern recognition and nonparametric statistics to credit the scoring problem. And also proposed an adjusted version of the Euclidean distance metric which attempts to incorporate knowledge of class separation contained that data. K-NN methodology is applied to a real data set and also discussed the selection of optimal values of the parameters k
and D included in this method. Ajay challagella et al (2010), proposed a technique for detecting outliers while preserving privacy, using hierarchical clustering methods. Outliers generally represent anomalous behavior. Outliers are rare occurrences and hence represent a small portion of the data. Manzoor Elahi et al. (2008), introduce a clustering based approach, which divide the stream in chunks and cluster each chunk using k-mean in fixed number of clusters.

Instead of keeping only the summary information, which often used in case of clustering data stream, we keep the candidate outliers and mean value of every cluster for the next fixed number of steam chunks, to make sure that the detected candidate outliers are the real outliers. By employing the mean value of the clusters of previous chunk with mean values of the current chunk of stream.

To overcome the limitation of fixed number of clusters $k$ as in k-mean algorithm for outlier detection, we have proposed a k-median algorithm that gives us a range of k-values in the intermediate steps that provides more stability to our algorithm than the k-means. The k-mean objective function is to minimize the largest assignment distance, whereas in k-median the sum of assignment distances (replaced by their squares) and the total cost is to be minimized by Parneeta Dhaliwal (2010). Michael Leblanc and John Crowley (1993) developed a recursive partitioning procedure based on maximizing the dissimilarity in the survival distributions of patients between different regions of the covariate space. The overall performance of the tree
structure representing the recursive partition is the sum of the two-sample test statistics between nodes in the binary tree.

Support vector machines (SVM) are widely applied to various classification problems. Shao et al. (2012) improved the projection twin support vector machine (PTSVM), a simple and fast algorithm by using least squares projection twin support vector machine (LSPT SVM) for generating binary classifier, which this gives the better generalization ability for optimization problems.

Yao et al. (2005) compared the performances of SVM and Radial Basis Function Neural Network (RBFNN) with other classification methods using two data sets. Their result shows that SVM is better than RBFNN. Carolina et al. (2007) proposed SVM based classification approach, based on the use of the support vector data description (SVDD) applied for the training data and obtained high classification accuracies.

Jair Cervantesa et al. (2008) presented a novel SVM classification approach for large data sets by using minimum enclosing ball clustering, were used for the first time SVM classification. Carl Gold and Peter Sollich (2003) proposed probabilistic methods to the case of quadratic slack penalties and also developed a baseline algorithm which can be used to find in principle exact maxima of the evidence.

Huang and Chang (2007) proposed evolutionary approach in designing SVM-based classifier by simultaneous optimization of automatic feature selection combined with $k$-fold cross validation
regarded as an estimator of generalization ability. Yang et al. (2010) proposed Probability Least Squares Support Vector Classification Machine (PLSSVCM) gives the classification results both a qualitative explanation and a quantitative evaluation. Fabian Laue and Gerard Bloch (2008) proposed an efficient evolutionary classification method by hybridizing the advantages of intelligent genetic algorithm (IGA) and support vector machine (SVM). The proposed method concluded the results for Support Vector Machine (SVM) is better than IGA for accuracy.

Jiang (2011) proposed an effective semi-supervised learning algorithm, called Instance Weighted Naive Bayes (IWNB) algorithm. Initially, this algorithm makes use of the naive Bayes trained data for labeling and then weights the instances from the unlabeled data set. At a final stage, a naive Bayes is trained again using both the originally labeled data and the (newly labeled and weighted) unlabeled data. Zhang and Su (2008) investigated the ranking performance of naive Bayes from both empirical and theoretical approaches. When compared to other study, naive Bayes performs well in ranking and classification. Zengchang Qin (2006) proposed two hybrid models by combining Naive Bayes classifier and probability estimation trees in order to build a model with good performance without losing too much transparency. The first model uses Naive Bayes estimation given a probability estimation trees and the second model uses a group of small-sized probability estimation trees as Naïve Bayes estimators.
Ratanamahatana and Gunopulos (2003) described a Selective Bayesian classifier (SBC) that simply uses the features in its decision tree, when learning a small example of a training set and a combination of the two different natures of classifiers. It is concluded that SBC performs markedly better than Naive Bayesian Classifier (NBC) on all domains. Also, Augmented Bayesian classifier (ABC) is also tested on the similar data, and it is evident that SBC perform well as ABC. Zhang and Lian (2003) proposed three layer geometric learnability of naive Bayes classification. Vaughn, and Qui Wang (2008) Classified an unknown observation into several populations by using tree-structured allocation rules analysis.