International Journal of Engineering & Technology, 7 (2.33) (2018) 532-535

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A study on sequential pattern mining on chemical information

S. Sathya [1] * , N. Rajendran [2]

[1] Research Scholar, Bharathiar University, Coimbatore. Assistant Professor, School of Computing Sciences, VISTAS, Chennai
[2] Principal, Vivekanandha College for Women, Namakkal Dist
**Corresponding author E-mail: ssathya.scs@velsuniv.ac.in, vpnraj@gmail.com*

Abstract

Data mining (DM) is used for extracting the useful and non-trivial information from the large amount of data to collect in many and diverse fields. Data mining determines explanation through clustering visualization, association and sequential analysis. Chemical compounds are well-defined structures compressed by a graph representation. Chemical bonding is the association of atoms into molecules, ions, crystals and other stable species which frame the common substances in chemical information. However, large-scale sequential data is a fundamental problem like higher classification time and bonding time in data mining with many applications. In this work, chemical structured index bonding is used for sequential pattern mining. Our research work helps to evaluate the structural patterns of chemical bonding in chemical information data sets.

Keywords: *Data Mining (DM); Chemical Compounds; Chemical Bonding; Sequential Pattern Mining*

1. Introduction

Data mining is the process of determining the patterns from large quantity of data. There are many data mining process like classification, clustering, association rule mining and sequential pattern mining. Sequential pattern mining identifies the group of data items that occur together in many sequences.

Sequential pattern mining removes the common subsequences from sequence database. Sequential pattern mining locates the significant patterns connecting the data where the values are distributed in a series. It is taken that the values are discrete and time series mining is connected. Sequential pattern mining is a type of the structured data mining. There are many traditional computational issues solved in this field. It also comprises the efficient databases and indexes for series information, removing the patterns, evaluating the series for similarity and regaining their missing sequence members.

This paper is organized as follows: Section II discusses chemical structured index bonding for sequential pattern mining, Section III shows the study and analysis of the existing chemical bonding techniques in data mining, Section IV identifies the possible comparison between them and Section V concludes the paper, key areas of research is given to evaluate the structural patterns of chemical bonding in chemical information data sets.

2. Literature review

Temporal Skeletonization approach [2] was introduced to minimize the representation of sequences for discovering the essential and hidden temporal structures in classification. The temporal correlations are reviewed in an undirected graph.

Temporal approach modifies the curse of cardinality in sequential pattern mining and clustering. However, the approach failed to find the relevant temporal structures in sequential data. However, a new strategy has to be designed for sequential pattern mining.

Fusing classifiers was planned in [7] to remove the sample data from classification. It is derived from the applications of probabilistic generative classifiers (CMM) by multinomial distributions. CMM has many components are multivariate normal distribution modelling and multiple multinomial distributions. Though, the fusing approach is not simplified to additional distributions and study is not managed in many prior distributions.

Probabilistic Sequence Translation-Alignment Model (PSTAM) was presented in [4] to collect the feature alignment. The latent variables of alignment and the base sequences are designed for mapping between sequences. A time monotonicity limitation on hidden alignment variables allocates efficient learning of class-specific time-warping and feature alteration. But, time warping changes the classification results. To choose the clustering algorithm, Fast Clustering-Based Feature Selection algorithm (FAST) was established by [8]. Features are partitioned into clusters by graph-theoretic clustering techniques. The clustering-based strategy of FAST has a high probability of creating a subset of functional and self-determining features. However, the algorithm fails todiscover many types of correlation measures and does not includethe properties of low-dimensional feature subspace.

Joint learning framework was designed in [6] depending on the reconstruction error to manage both multi-label and multi-view learning settings in sequential patterns. Semi-Supervised Dimension Reduction for Multi-Label and Multi-view Learning
(SSDR-MML) executes the optimization for dimension reduction and label inference in semisupervised setting. However, the multiple labels and multi-label learning are notadvantageousbut detrimental. Online class imbalance learning method [5] was designed to increase the resampling policy. The learning method is designed with two learning algorithms are over sampling based Online Bagging (OOB) and under sampling- based Online Bagging (UOB). The original OOB and UOB were developed and the sampling rate is reliable. A group of static data streams unreliable in data distributions and imbalance the examined rates. However, dynamic data streams are not designed with concept drifts and learning techniques are not extended in multi-class cases.

3. Chemical structured index bonding for sequential pattern mining

Graphs are playing major role in modelling difficult structures like circuits, images, chemical compounds, protein structures, biological networks, social networks, the web, workflows and XML documents. Several graph search algorithms are designed in chemical informatics, computer vision, video indexing and text retrieval with huge requirement on the examination of large number of structured data.
Chemical compounds are well-defined structures that are easily compressed by graph representation.
Compounds are made up of number of atoms represented as vertices in a graph and a number of bonds between atoms denoted as edges in the graph. Vertices are mentioned with atom element form and edges are labelled with bond form. The edges in the graph are undirected because there is lack of direction associated with chemical bonds.

3.1. Probabilistic sequence translation alignment model for time-series classification

The sequence classification issue is essential because of the series data in many forms like videos, speech signals, biological structures and meteorological records. The sequence classification is demanding and less explored than standard multivariate classification because of the complexity in controlling the variable-length sequences with different changes by possible noise and gaps. The sequence classification is divided into two types. They are: alignment-based and model based. In alignment-based methods, potentially nonequal-length sequences are aligned in time in nonparametric fashion for different time scales in quantity, dissimilar rates of changes and noise or gaps.
A new technique isdesigned to sequence the classification with the advantages fromthe worlds. The main aim is to havethe series in training data associatedwith theadditional base sequence that parameterize as well asfind out the class-specific time-warping and feature transformation regarding the base series. The class-conditional density models are considered with alignment of sequences. The key aim of the approach is the conditional density model for pair of series that collects the translation/ alignment process from one series to another. Through considering the sequences as sentences and features as words or phrases, it allows theanalogy to statistical machine translation in computational linguistics to formthe translation from one language into an additional one.
With probabilistic sentence translation approaches, the designed model plans the hidden alignment variables that instruct the matching of words and find out the feature-to-feature mappings from data with all possible alignments through marginalization. The designed model is termed as the probabilistic sequence translation-alignment model (PSTAM). The model considerablymanages the real-valued multivariate time-series throughcommanding the time-monotonicity limitations in the parameter space. Throughenhancing the hidden base sequence and learning the model parameters in coordinate-ascent fashion during the nested expectation maximization
(EM) algorithm, PSTAM has feasiblekey to series classification. The base sequence reviews the alignment information of all training series while learned model parameters denote the class-specific translation processes.

3.2. Temporal skeletonization on sequential data

Sequential pattern analysis aims on locating thesignificant temporal structures where the values are distributed in a series. The separation of the meaningful and essential temporal structures from large-scale sequential data is a key issue in data mining with many applications like mining the customer purchasing sequences, motion gesture/ video sequence recognitionand biological sequence analysis. When thetechniques effectively used in applications, there are many problemssolved when the irresistible scale and the heterogeneous sequential data occurs. In few applications, it is complex to attain the knowledge of symbols. Many sequential data utilizes an arbitrary coding of events for simplicity or securitycauses. There are many conditions where it is complexto describethe distance betweenthe symbols and clusteringis developing into an impractical. It is uncertain to describe the distance between performances in purchasing process. The key difficulty is the grouping where the techniques are carriedout regardlessof the temporal content. The techniques are not used to discover the relevant temporal structures in sequential data.
A temporal skeletonization approach is designed to minimize the demonstration sequences to describe their hidden temporal structures. The main aim is to design the temporal structures of sequences as accurate and simplified onewhich isvulnerable to the discovery. The key statement is the existenceof symbolic events that leads to theaggregation. By identifying the temporal clusters and mapping all symbols to the correspondingcluster, the cardinality of sequences and their temporal changes are reduced. The hidden temporal structuresare recognized andmade as unclear in the original demonstration. The temporal clustersdetermining from large number of sequencesare demanding one. For determining the temporal clusters, graph-based manifold learningis taken. The main aimis to review the temporal connections in the data inundirected graph. Theskeleton graph is removed through the graphLapponian thatprovided as advanced granularity where hiddentemporal patterns are recognized. Ananalysis of temporal grouping is carried out when theindividual symbols are restored through their cluster labels. The averaged efficiency of all series is increased. The chance significant sequential patterns are identified and also examined. The embedding topology of graphchanges the temporal content of symbolicseries into metric space for the examination and visualization.

3.3. Probabilistic generative classifiers with data mining applications

Inmachine learning applications, the process of taking out the knowledge from the sample data is separated into number ofsubtasks. Probabilistic classifiers presentthe outputs are interpreted as conditional probabilities with conditional distribution of classes given as the input sample.
Generative classifiers objective is to model the processes where the sample data are taken as original data. Probabilistic generative classifiers depend on Bayes' theorem. Initially, the classifiers are utilized in type of ensembles, a design is made for many existing recognitions. For probabilistic classifiers, the outputs are used as the posterior probabilities which are easier. The probabilistic classifiers present the chance to join the classifiers at level of components of the combination models. It is achievedthrough combining all the component sets and renormalizing the combination coefficients. Finally, the components or rules are combined at the level of parameters. It is essential to the parameters of many components in a suitable method if the components are same.
Fusion classifiers are used to combine the parameters of components. Averaging of parameters is taken as simple at initial view but multivariate distributions are required. It is essential to decompose the covariance matrix into two matrices relating scaling and replacement of a multivariate standard normal distribution. For all the parameter, mixture coefficient, centre, covariance matrix are described as hyperdistributions that model the ambiguity with exact value of the parameter. A normal-gamma distribution is required for the second-order distribution over two parameters is centre and variance of normal distribution. A distribution is needed for all
Gaussian component of classifier. The parameters of hyper-distributions are trained from sample data in
Bayesian system. For actual description of classifier, parameters of classifier's components are identified with the probabilities of the second-order distributions.
The main objectives the actual fusion of classifiers that realizedthrough developing the second-order distributions when the classifier is depending on the members of exponential distributions.

4. Comparison of chemical structured index bonding for sequential pattern mining & suggestions

In order to compare the chemical structured index bonding for sequential pattern mining, number of sequential patterns is taken to execute the experiment. Various parameters are used to measure the chemical structured index bonding of data mining techniques.

4.1. Classification accuracy (CA)

Classification accuracy is defined as the ratio of number of exactly identified chemical bonds to the total number of sequential patterns. It is measured in terms of percentage (percentage).

$$CA = \frac{\text{Total number of exactly identified chemical bonds}}{\text{Total number of sequential patterns}}$$

Fig 4.1: explains the comparison between the three methods namely, Temporal Skeletonization Approach, Probabilistic Sequence Translation Alignment Model (PSTAM) and Fusing Classifier. From the figure, Probabilistic Sequence Translation Alignment Model (PSTAM) provides higher classification accuracy in terms of chemical structured index bonding as compared to other methods. The classification accuracy is raised when the number of sequential patterns gets increased. The percentage of Probabilistic Sequence Translation-Alignment Model (PSTAM) improves the classification accuracy by 21.5%when compared to Temporal Skeletonization Approach and also improves by
14.5% when compared to Fusing Classifier.

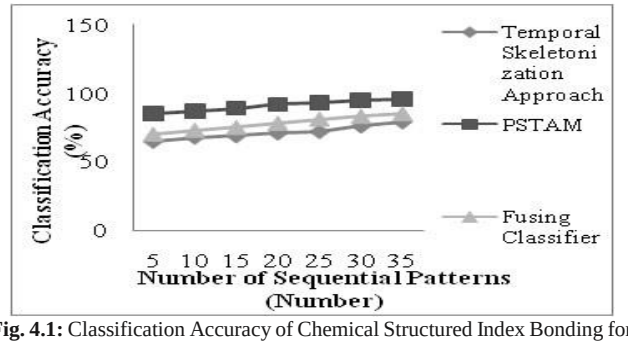

Fig. 4.1: Classification Accuracy of Chemical Structured Index Bonding for Sequential Pattern Mining.

4.2 Density level of chemical bonding

Density level of chemical bonding is defined as the number of chemical information compacted in the space for chemical bonding. It is measured in terms of percentage (percentage).

$$Density\ Level = \frac{\text{Number of Chemical Information exists}}{\text{Memory Space (size) for total Chemical Bondong}}$$

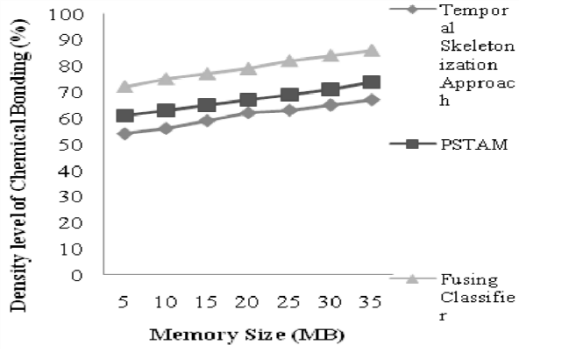

Fig. 4.2: Density Level of Chemical Structured Index Bonding for Sequential Pattern Mining.

In figure 4.2, density level of chemical structured index bonding for sequential pattern mining is described. The existing methods such as Temporal Skeletonization Approach, Probabilistic Sequence Translation-Alignment Model (PSTAM) and Fusing Classifier are compared with each other. From the figure, the density level of fusing classifier approach is comparatively higher than that of the Probabilistic
Sequence Translation-Alignment Model (PSTAM) and Temporal Skeletonization Approach. The density level of fusing classifier is 25.1% and 9.1 % higher than that of Temporal Skeletonization Approach and Probabilistic Sequence Translation-Alignment Model (PSTAM) respectively.

4.3. Bond indexed sequential time

Bond indexed sequential time is the difference of the starting and ending time of the sequential index bonding. It is measured in terms of milliseconds (ms).

Bond Indexed Sequential Time = Ending time –

Starting time of sequential index bonding
Fig 4.3: portrays the bond indexed sequential time comparison of existing methods such as Temporal Skeletonization Approach, Probabilistic Sequence Translation-Alignment Model (PSTAM) and Fusing Classifier. From the comparison, bond indexed time sequential time of Temporal Skeletonization Approach is comparatively lesser than that of Probabilistic Sequence Translation Alignment Model (PSTAM) and Fusing Classifier.

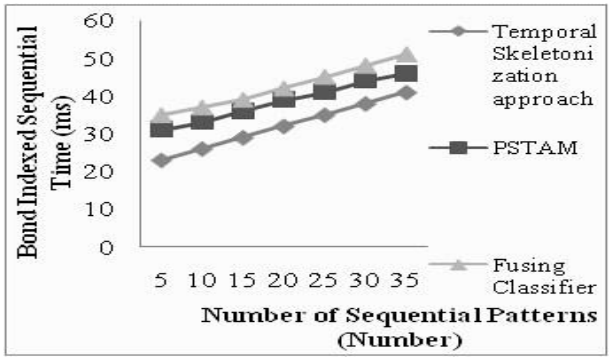

Fig. 4.3: Bond Indexed Sequential Time of Chemical Structured Index Bonding for Sequential Pattern Mining.

Bond indexed sequential time of Temporal Skeletonization Approach is 10.4% lesser when compared to the Probabilistic Sequence Translation Alignment Model (PSTAM). In addition, Bond indexed sequential time of Temporal Skeletonization Approach is 30.4% lesser than Fusing Classifier.

5. Discussion on limitation of chemical structured index bonding for sequential pattern mining

Temporal Skeletonization approach minimizes the sequences to reveal the hidden temporal structures. Temporal approach changes the curse of cardinality in sequential pattern mining and clustering. The temporal Correlations are reviewed in undirected graph. The estimation in Business to-Business (B2B) marketing applications determines the paths from noisy customer event data. However, this approach failed to locate relevant temporal structures in sequential data. However, it needs to develop a new vision and plan for sequential pattern mining.
Fusing classifiers is depending on the utilization of probabilistic generative classifiers (CMM) by multinomial distributions. CMM has many components such as multivariate normal distribution modelling and multiple multinomial distributions. Dirichlet and normal-Wishart distributions are conjugating the prior distributions

of the multinomial and normal distributions. However, the approach is not generalized to other distributions and the investigation is not handled to different prior distributions. Probabilistic model captures both feature alignment and mapping between sequences by introducing the latent variables of alignment and the base sequence. A time monotonicity limitations on hidden alignment variables that allocates learning of class-specific time-warping and feature transformation. The low-dimensional modelling of latent base sequence creates the intrinsic manifold structure. Alignment-based methods estimate the distance measures between non-equal-length sequences through aligned features. However, time warping affects the classification results significantly.

5.1. Related works

Booster in [1] categorizes the features chosen in sequences pattern. The Booster, boost the results of an FS algorithm with synthetic data and microarray data sets. Booster increases the prediction accuracy and calculates the complexity of a data set for classification. However, an FS algorithm itself is not efficient and booster may not be able to obtain high performance. Multivariate Reconstructed Phase Space- Gaussian mixture model (MRPSGMM) algorithm [9] was planned to predictive pattern classification when the event samples are small. Original univariate reconstructed phase space framework is designed depending on the fuzzy unsupervised clustering technique. The algorithm has higher results in dataset with large percentage of heterogeneous patterns. An alternative clustering techniques or distributions are required for the Gaussian mixture model to recognize the classification pattern. To identify the unexplained sequences, an efficient Top-K algorithm was designed in [3]. The algorithms manage the theorems to increase the searching speed for totally/partially unexplained series. Though, it allocates the activity occurrences to violate the temporal limitations in stochastic automata-based activity model. However, it needs specialized data structures to enhance the scalability of algorithm.

5.2. Future work

The future direction of chemical structured index bonding for sequential pattern mining can be carried out to evaluate the structural patterns of chemical bonding in chemical information data sets. In addition, structural patterns can be analysed for the sequences where chemical bonds are organized. Furthermore, Principal Components Analysis (PCA) can be used to identify the exact match of test sample data pattern of chemical bonds to the trained BIS pattern.

6. Conclusion

A comparison of many chemical structured index bonding for sequential pattern mining techniques is surveyed. In the present environment, this approach failed to locate relevant temporal structures in sequential data and also needs to develop a new vision and plan for sequential pattern mining. Fusing classifiers is not generalized to other distributions and the investigation is not handled to different prior distributions. The wide range of experiments on existing techniques calculates the comparative results of the many chemical structured index bonding for sequential pattern mining techniques and its limitations. Finally from the result, the research work can be carried out with chemical structured index bonding for sequential pattern mining to evaluate the structural patterns of chemical bonding in chemical information data sets.

References

[1] HyunJi Kim, Byong Su Choi and Moon Yul Huh, “Booster in high dimensional data classification”, IEEE Transactions on Knowledge and Data Engineering, 2016, Volume 28, Issue 1, Pages 29-40.

[2] Chuanren Liu, Kai Zhang, Hui Xiong, Geoff Jiang and Qiang Yang, “Temporal Skeletonization on Sequential Data: Patterns, Categorization, and Visualization”, IEEE Transactions on Knowledge and Data Engineering, Year 2016, Volume 28, Issue 1, Pages 211-223.

[3] Massimiliano Albanese, Cristian Molinaro, Fabio Persia, Antonio Picariello, and V.S. Subrahmanian, “Discovering the Top-k Unexplained Sequences in Time-Stamped Observation Data”, IEEE Transactions on Knowledge and Data Engineering, March 2014, Volume 26, Issue 3, Pages 577-594. [4] Minyoung Kim, “Probabilistic Sequence Translation-Alignment Model for Time-Series Classification”, IEEE Transactions on Knowledge and Data Engineering, February 2014, Volume 26, Issue 2, Pages426-437. [5] Shuo Wang, Leandro L. Minku, and Xin Yao, “Resampling-Based Ensemble Methods for Online Class Imbalance Learning”, IEEE Transactions on Knowledge and Data Engineering, May 2015, Volume 27, Issue 5, Pages 1356-1368.

[4] Buyue Qian, Xiang Wang, Jieping Ye, and Ian Davidson, “A Reconstruction Error Based Framework for Multi-Label and Multi-View Learning”, IEEE Transactions on Knowledge and Data Engineering, March 2015, Volume 27, Issue 3, Pages 594-607.

[5] Dominik Fisch, Edgar Kalkowski, and Bernhard Sick, “Knowledge Fusion for Probabilistic Generative Classifiers with Data Mining Applications”, IEEE Transactions on Knowledge and Data Engineering, March 2014, Volume 26, Issue.3, Pages 652-666.