Elysium Journal of Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

A SURVEY ON ERGODIC RATING PREDICTION USING SEQUENCE CLASSIFICATION

Dr C Sundar[1] and L.Manjula[2]

[1]*Associate Professor, Department of Computer Science and Engineering, Christian College of Engineering and Technology, Dindigul, Tamil Nadu -624619 India.*

[2]*PG Scholar, Department of Computer Science and Engineering, Christian College of Engineering and Technology, Dindigul, Tamil Nadu -624619 India.*

Email: sundarc007@yahoo.com[1], lakme_1983@yahoo.co.in[2]

Abstract— Sequence classification has an extensive of applications range such as analysis of the genome, retrieval of information, information about fitness, economics, and abnormal detection. Thus, it creates sequence classification, a more interesting task while the current methods only emphasis of frequency on the concept but not over the time. There exist many more algorithms for upholding sequential patterns. The older datasets are removed and other datasets get updated. Hence, it is clear that timestamp was an important attribute of each dataset for the process of data mining and it provides more accurate and useful information. Sequential patterns use static databases, as time passes by new datasets are inserted and makes the current databases into the dynamic database enhancing the ergodicity of prediction. A brief review of the contemporary work on sequence classification is presented along with highlighting rating prediction.

Keywords— Sequence classification, interesting patterns, and classifier types

1. INTRODUCTION

Sequential important settings such as texts, videos, speech signals, data is often encountered in the number of biological structures and web usage logs. Extensive applications range for sequence classification leads to a significant problem in data mining and the statistical machine learning. Sequence classification is the process of allocating class labels to new sequences are based on the knowledge gain in the training stage. There are some integrating pattern mining techniques, such as association rule based classification, sequential pattern based classification and interesting pattern-based classification. The combined methods of these techniques produce good results as well as provide users with information useful for understanding the characteristics of the dataset. The quantity of determining all frequent sequences in large databases is relatively challenging since the research space is very large. Creating the sequential pattern mining is a very problematic and time-consuming one, such as the formation of a pattern has not limited to a single items but item sets are neither the number of item sets in the pattern nor the number of items in an item set is known as apriori, and patterns can be formed by any permutation or by any combination of possible items in the database.

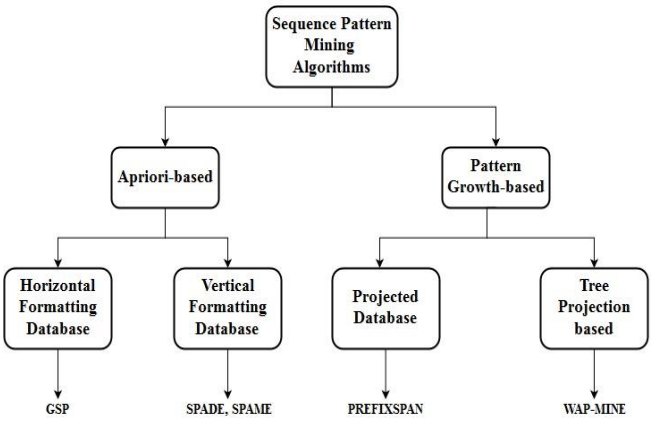
2. SEQUENTIAL PATTERN MINING ALGORITHMS

Many approaches towards the sequential pattern mining ensure two concerns: 1. Extension of pattern mining sequence into time associated patterns. Time-related patterns relate the potential applications for the sequential patterns and numerous extensions, such as discovering constraint and sequential pattern in time interval. Some extensions of those algorithms for special purposes such as multidimensional, closed, time interval, and proposed the constraint-based sequential pattern mining. 2. Competence of the sequential pattern mining has to be improved. To increase the efficiency (competence), the sequence differ in two ways,

i. To minimize the cost of I/O through reducing the number of candidate sequence generated.

ii. To support the counting purpose in all the time, the elimination of any database or data structure is the key strategy.

Based on these criteria's, mining could be divided broadly into two parts: Apriori-based and pattern growth based.



P-ISSN: 2347-4408
E-ISSN: 2347-4734

Fig.1 Classification of Sequential Pattern Mining Algorithms

2.1 Apriori algorithms

These set a basis for all breeds of algorithms and link procedure to produce candidate sequences. The Apriori statement is that each and every non-empty or filled subset of item sets should be frequent. Here, if a sequence is not able to pass through the minimum support test, then it fails the entire super sequences. Important terms of the Apriori -based algorithms are,

i. Breadth-first search: The sequence in the j^{th}iteration of the Apriori algorithm traverses the search space.

ii. Generate-and-Test: It creates huge number of candidate sequences and tests each one sequentially to satisfy user constraints that consume lots of memory.

iii. Multiple scan of the database: It is undesirable because it requires the lots of processing time and IO cost.

Generalized Sequential Pattern (GSP): This makes multiple passes on the data and acts faster than the Apriori algorithm. It involves two steps: candidate generation method and candidate pruning method. They are not main memory algorithm hence, it generates candidates equal to fit in the memory and scanning of datasets help to find out the support of the candidate. If the support from the candidate is minimum, it is deleted while the frequent sequences from those candidates are written to the disk. The process gets repeated until all the candidates are counted. The algorithm finds the length of all candidates and orders them with their support value. At each level, (length of the sequences) the algorithm scans the datasets to acquire the count of each candidate sequences and they generate length (K+1) sequence from the frequent length-K sequences using Apriori. It is continued until there is no frequent sequence or candidates found. The algorithm has very good properties in terms of a number of transactions per data sequence and the number of items per transaction. But it is inefficient in large sequencing of databases that have numerous pattern or long pattern as they are not able to generate more candidates sequence and it also requires multiple scans for database since the length of each candidate grow by one at each database scan.

SPADE: The horizontal formulation methods are transformed into vertical datasets format that consists of item id list. The vertical datasets consists of list of sequential ids and timestamps that indicate the timestamp of the item in that sequence. The searching progress is done with the id-list interaction and it completes the mining in three passes of database scanning. Meanwhile, the computation time required to transform and additional storage spaces are several times the original sequence databases.

SPAM: It associated with the ideas of GSP, SPADE, and Free span. It uses the vertical bitmap data structure that is similar to the given id-list of SPADE. It fits the main memory with a whole algorithm and its data structure. To increase the performance of SPAM, depth-first traversal is used. Though SPAM is similar to SPADE, it uses the bitwise operations instead of regular and temporal join. The SPAM outperforms the SPADE while the SPADE is SPACE-efficient than SPAM.

2.2 Pattern-growth Sequential Pattern Mining Algorithms

It acts as a problem solver for generating and test. It avoids the candidate generation step and focuses on the restricted portion of the initial database. The feature termed as a search space partitioning plays a vital role in pattern growth. The method of pattern growth start with a representation of the database mining, then it makes way for search space partitioning to generate a few candidate sequence as possible from already mined frequent sequence and Apriori method act as the search space to look for frequent sequences while recursive traversing. The early algorithms started with using projected databases.

2.2.1 Prefix Span

It is the only projection based algorithms from all sequencing pattern in mining algorithms. It outperforms all other algorithms like Apriori, Freespan, and SPADE. They find frequent items by scanning the sequence database only once. According to the frequent item sets, the database is estimated into several small databases. The complete set of sequential patterns is formed with the recursive growth of sequence fragments in every projected database. The concept behind the Prefix span algorithm is to discover patterns, which is employed by a divide-and-conquer strategy. The algorithm requires high memory space compared to other algorithms because of creation and processing a huge number of projected sub-databases.

2.2.2 FreeSpan:

It reduces the cost for candidate generation and testing of Apriori and satisfies its basic features. They use frequent items to iteratively project the sequence database into projected database. They grow sequences frequently in each projected datasets. Every projection divides the database and confines testing to progress smaller and more manageable units. The key issue is to consider the amounts of sequences appear is more than the single projected database and the size of database decreases with each iteration.

2.2.3 WAP-MINE:

These are based on tree-structure mining technique, in which the sequence database is scanned twice to build up the WAP-tree from the frequent sequences with their support values. The header table is maintained to point where the first occurrence of each item in frequent item sets that helps to mine the tree for frequent sequences so that it built up on their suffix. It is more scalable than GSP and it performs bitterly with marginal points. It scans the databases twice and avoids

P-ISSN: 2347-4408
E-ISSN: 2347-4734

the problem of generating huge candidate in case of Apriori approach. Since it iteratively regenerates, n increases automatically.

3. PATTERN BASED SEQUENCE CLASSIFICATION

The determination of protein folding helps to compute and connect the pattern of the disulfide bonds that significantly reduce research space in solving the protein-folding problem. Thus, it develops an effective means of predicting the disulfide pattern and estimates three-dimensional structure of a protein and its function[1]. The importance of feature selection or extraction in the classification of patterns is presented. Here predicted cytosine connectivity patterns based on four features. Experimental results indicate that the proposed method achieves an accuracy of 79.8% tested using fourfold cross-validation with SP39 (SWISS-PROT) dataset, which is better than the prediction performance reported in previous studies. The results indicate that the proposed method achieves prediction accuracies as high as 70.6% and 69.8% when the SP39 dataset is used as the training dataset, and the SP43 and SP56 datasets as the testing datasets, respectively. The use of Support Vector Machine (SVM) along with the Multiple Trajectory Search (MTS) to adjust the optimal parameters of SVM and the window size for various features are proposed to accurately predict a disulfide connectivity pattern efficiently.

A novel selection approach for mining a representative summary of a set of frequent 3D-motifs[2]are presented. Unlike current methods that are based on the relations between patterns in the transaction space, this approach considers the distance between patterns in the pattern space. The proposed approach exploits a specific domain knowledge, in the form of a substitution matrix, to select a subset of representative un-substituted patterns from a given set of frequent 3D-motifs. The results of this analysis incorporate the domain knowledge to allow the proposed approach to detect relations between the patterns while the existing pattern methods[5] fail. They also allow the initial set size of 3D-motifs to obtain an easy and efficient exploration, which is more interesting and act as a representative. This approach helps in other motif based analysis of protein sequence classifications. An encouraging future direction considers the insertions and deletions over the nodes and edges. Hence, it considers the substitution over patterns with different sizes. Although it increases the complexity and the difficulty of the selection exponentially, it is closer to the real world substitution phenomenon. Since the proposed approach is a filter approach, it would be also interesting to embed the selection within the extraction process in order to mine directly the representative patterns from data.

The review of the proposed work [3] is to provide a bridge between various application domains [11],[7], and promote an added value towards the structural properties of DNA to have functional genomics. Thus, the DNA structural properties play an important role in many bio-molecular processes. The characterization of different genomic elements is essential to generate a complete understanding. This work presents different genomic elements that require different representation methods to capture potential defining structural patterns. The methodologies used are not able to learn from one another as their functional or bio-molecular relationships between most of their elements. For instance, the inherent flexibility of the DNA molecule is often found to be an informative feature in a large number of genetic elements. Different genomic elements share a number of structural similarities that results a false positive prediction during classification and it leads to problems. Hence, a proper structural characterization helps to understand such intricate relationships between different elements like promoter regions and splice sites. The DNA structural properties play a role in more critical areas and hence, a further study is needed to contribute the underlying bio-molecular mechanisms.

The key perspective elements considered are the pattern type and data structure independence[4]. The work proposed is then categorized into two dimensions, such as Model dependent or model independent (whether the pattern model independently or the pattern is guided by a model), Iterative or post processing (whether the set of patterns are pre-computed by post processor iteratively perform pattern mining algorithms). For almost any quality measure and mining techniques, both the pattern language and the language in which data are expressed are not relevant to the pattern set selection phase as long as there is a well-defined matching operator between the two. Furthermore, almost all techniques for mining class-sensitive patterns themselves are independent of these aspects as well, with the exception of data structures used. Such data structures, however, typically do not influence the applicability of mining techniques but only their implementation. This means that it is possible to transfer approaches freely between different representations and settings, albeit possibly at a certain cost of efficiency. Iterative approaches[8], [12] and [13]have the advantage of taking the effects of already selected patterns into account by adjusting the scoring function in some way. This allows focusing on interesting areas of the pattern space, pruning subspaces that would have been explored in non-iterative mining. The downside to this is that the space of potential solutions is far larger than in the non-iterative case, requiring the adoption of heuristic techniques and less control over, and looser guarantees for the quality of resulting sets.

Resulting pattern sets are used as an input to a different modeling step but might perform worse than pattern sets produced by model-independent approaches[14], [15]. A major issue in the current state-of-the-art [7]is that so far it is not very clear to what degree the merits and drawbacks that are derived analytically for different approaches materialize empirically. In most of the papers that proposed pattern-based classification algorithms, experiments were performed to

P-ISSN: 2347-4408
E-ISSN: 2347-4734

show the benefits of the approaches. However, these comparisons were (understandably) often limited; they did not exhaustively consider all relevant comparable approaches that derive if one would take pattern-type independence into account and recognize that graph-based approaches may also be used in simpler pattern domains; also, the number of data sets in most publications is limited and usually restricted to one data type.

4. RATING PREDICTION

Prediction models continuous-valued functions, i.e., predicts unknown or missing values .The typical applications include credit approval, target marketing, medical diagnosis and fraud detection. Prediction is a two-step process involving model construction and model usage.

1. Model construction

Describes a set of predetermined classes, where each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction is training set. The model is signified as classification rules, decision trees or mathematical formulae.

2. Model usage

Model usage is to classify future or unknown objects for estimation of accuracy of the model. The known label of the test sample is compared with the classified result from the model. The percentage of test set samples which are classified by the model is known as accuracy rate. The test set is independent of the training set, otherwise over-fitting will occur. If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.

5. SCOPE OF SURVEY

The scope of survey is to present a work that brings about higher prediction accuracy while existing systems use various learning models, involving multiple classifiers leading to small deviations. The Support Vector Machine (SVM) classifier analyses the datasets for classification and regression analysis. Identification of datasets based on training a set of data and compares it with threshold and generates a random value for further performance appraisal. The datasets are further clustered and decision trees are build using C4.5 algorithm to increase the rating accuracy.

6. RELATED WORKS

Rao,Lei and M. Chen in the year 2014 proposed a work on building emotional dictionary for sentiment analysis of online news[1] . A. Alashqur, in the year 2015 presented a novel methodology for constructing rule-based naïve bayesian classifiers [2].Wang and Yan in the year 2014 proposed a work on fast prediction of protein–protein interaction sites based on extreme learning machines [3].

R. H. Rajan and Dhas in the year 2012 created a method for classification based on association rules using ontology in Web data [4].Zhou,Cule and Goethals in the year 2013 proposed a work on itemset based sequence classification [5].

7. CONCLUSION

Using various types of their algorithms the sequential pattern mining is surveyed. The concept of sequential pattern mining has been introduced early, has gone through remarkable advancement in few years. Initial work of this concept has concentrated on the improvement of the algorithm performance by using different data structure or different representation. On account of problems present in sequential pattern mining, it categorized into various forms.The project has to present a methodology where rating prediction can be generated more accurately by using the classifier events and also the prediction algorithm. Then the different dataset can be given as an input for the related attributes. Finally the related dynamic result can be generated with efficient accuracy related to dynamic datasets. Experimental results show that embedded datasets are classified and analyzed to find a function which models the data with the least error. SVM through training, maps their input datasets into high-dimensional feature spaces thus enhancing the ergodicity of rating prediction. C4.5 algorithm chooses the attribute to make the decision for accurate prediction with the highest normalized information gain. The work can be extended to LDA (Latent Dirichlet allocation) process for classification and to generate the predicted result to increase more accuracy. Then the suggestion can be implemented to improve the accuracy with other events with different attributes.

REFERENCES

[1] H.-H. Lin and L.-Y. Tseng, "Prediction of disulfide bonding pattern based on a support vector machine and multiple trajectory search"

[2] W. Dhifli, R. Saidi, and E. M. Nguifo, "Smoothing 3D protein structure motifs through graph mining and amino acid similarities"

[3] P. Meysman, K. Marchal, and K. Engelen, "DNA structural properties in the classification of genomic transcription regulation elements"

[4] B. Bringmann, S. Nijssen, and A. Zimmermann, "Pattern-based classification: a unifying perspective"

[5] T. Kliegr and J. Kuchař, "Benchmark of Rule-Based Classifiers in the News Recommendation Task"

[6] C. Sun and R. Nevatia, "Semantic aware video transcription using random forest classifiers"

P-ISSN: 2347-4408
E-ISSN: 2347-4734

[7] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news"
[8] A. Alashqur, "A Novel Methodology for Constructing Rule-Based Naïve Bayesian Classifiers"
[9] I. Boudellioua, R. Saidi, M. Martin, R. Hoehndorf, and V. Solovyev, "Prediction of Metabolic Pathways Involvement in Prokaryotic UniProtKB Data by Association Rule Mining"
[10] D. D. Wang, R. Wang, and H. Yan, "Fast prediction of protein–protein interaction sites based on extreme learning machines"
[11] H.-C. Kuo and M.-Y. Tai, "Classifying Protein-Protein Interaction Type based on Association Pattern with Adjusted Support"
[12] R. H. Rajan and J. P. M. Dhas, "A method for classification based on association rules using ontology in Web data"
[13] C. Zhou, B. Cule, and B. Goethals, "Itemset based sequence classification"
[14] C. Premebida, D. R. Faria, and U. Nunes, "Dynamic Bayesian network for semantic place classification in mobile robotics"
[15] Z. Liu, C. Zhang, and Y. Tian, "3D-based Deep Convolutional Neural Network for action recognition with depth sequences"