International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
458 – 461

Various Sequence Classification Mechanisms for Knowledge Discovery

Goutami R. Mane
(PG Scholar): Computer Science and Engineering.,
DKTE Society's Textile & Engineering Institute (An Autonomous Institute)
Ichalkaranji, India.
goutami.kamat@gmail.com

Suhas B. Bhagate
(Asst. Prof.): Computer Science and Engineering.,
DKTE Society's Textile & Engineering Institute (An Autonomous Institute)
Ichalkaranji, India.
suhas.bhagate@gmail.com

***Abstract*—** Sequence classification is an efficient task in data mining. The knowledge obtained from training stage can be used for sequence classification that assigns class labels to the new sequences. Relevant patterns can be found by using sequential pattern mining in which the values are represented in sequential manner. Classification process has explicit features but these features are not found in sequences. Feature selection techniques are sophisticated, but the potential features dimensionality may be very high. It is hard to find the sequential nature of feature. Sequence classification is a more challenging task than feature vector classification. Sequence classification problem can be solved by rules that consist of interesting patterns. These patterns are found in datasets that have labeled sequences along with class labels. The cohesion and support of the pattern are used to define interestingness of a pattern. In a given class of sequences, interestingness of a pattern can be measured by combining these two factors. Confident classification rules can be generated by using the discovered patterns. Two different approaches to build a classifier are used. The first classifier consists of an advanced form of classification method that depends on association rule. In the second classifier, the value belonging to the new data object is first measured then the rules are ranked.

Keywords- *Sequence classification, interesting patterns, classification rules.*

*****

I. INTRODUCTION

Data mining is the process to analyze hidden patterns of data into useful information. It determines patterns and setup relationships to solve problems by using data analysis. By analyzing data for frequent patterns association rules are generated. Factors such as support and confidence finds the relationships within the data.

A sequence can be defined as an ordered list of events. Sequential data can be often found in some important settings such as videos, web usage logs, texts and biological structures. In sequence classification datasets can be divided into two cases. In first case the class of sequence is determined by certain items that co-occur within it, but not always in same order. In the second case the class of sequence is determined by items that mostly occur in same order. Sequential pattern mining is used to find relevant patterns among data where the values are delivered in sequence. One of the important tasks in data mining is sequence classification. The information is arranged into sequences in sequence classification. In biology, for example, to understand structure and function of DNA or protein sequences, it is necessary to classify it into different categories. In medicine, to identify pathological cases, it is required to classify time series of heart rates. Many sequence classification models are application-dependent. The sequence classification models such as speech recognition, text classification, bioinformatics, and customer behavior predictions apply certain domain knowledge to build the classifier. The complexity of these models is very high; most of these algorithms are not capable of working on large datasets.

Sequence classification is suitable to a large number of applications. It can be defined as assigning class label to new sequences. There are previous approaches which integrate classifications techniques such as Class by Sequence (CBS) algorithm, sequential pattern based sequence classifier, classification based on association rule (CBA), and pattern mining techniques. These systems have their own advantages. Such techniques can be combined together to achieve better results. These methods provide information which is useful for users to understand the characteristics of data. All these methods mine the frequent and confident patterns to build a classifier, but they don't consider cohesion of a pattern that affects the performance of classification. To overcome this, a method called as Sequence Classification based on Interesting Patterns (SCIP) can be used.

II. RELATED WORK

B.Liu proposed a system that integrates association rule mining and classification technique [1]. To find a small set of rules in the database classification rule mining is used. To form a classifier discovered set of rules can be used. Association rule mining is used to find the rules in the database. These rules can satisfy minimum confidence and minimum support criteria. The aim of discovery is not previously determined in association rule mining. Class is one and only one pre-determined target in. If the two mining

 ISSN: 2454-4248
458 – 461

techniques can be integrated, it gives better results. The associative classification is used as an integrated framework. The integration of association rule mining and classification can be done by the class association rules (CARs). This integrated framework is used to apply association rule mining techniques to classification tasks. Some problems such as understandability problem, a discovery of interesting or useful rules can be solved by this integration. But it requires discretization of continuous attributes. All the class association rules (CARs) has to be generated and by using these CARs it builds a classifier.

B. Liu proposed a system that uses classification with association rule [2]. In data mining to build an effective classification systems is an important tasks. Many techniques have produced in past research (e.g. Naive-Bayes classification, rule learning, decision trees). These techniques are mainly based on greedy search. To form a classifier, they aim to find only a subset of the regularities existing in data. The user specified minimum support and minimum confidence are satisfied by the set of rules. To discover such rules is the objective of association rule mining. To build an effective classifier, these rules can be used. An extensive search based classification system is Classification Based on Associations (CBA). The most accurate rules are used to build a classifier. It is the main advantage of CBA system. Association rule mining uses only a single minimum support in rule generation. A single minimum support is not sufficient in an unbalanced class distribution. The number of rules is very large in classification. To generate rules having many conditions is a difficult task for rule generator even if such rules are necessary for correct classification.

W. Li proposed a method called as efficient and accurate classification based on multiple class-association rules [3]. In Classification Based on Association (CBA), the accuracy of classification is high. The flexibility to handle unstructured data is strong, but a number of rules are very large. The classification uses only single high-confidence rule. It creates the problem of overfitting in classification. This problem can be overcome by the CMAR (Classification based on Multiple class-Association Rules) approach. The FP-Growth algorithm is used in CMAR approach. This algorithm is used to generate frequent itemsets. To classify object by using just one rule, the matching rules subset can be used. The accuracy of classification is improved in CMAR. The FP-tree structure is more efficient. This FP-tree structure has used in CMAR. The multiple rules are used in CMAR approach. These rules can be used to predict associated weights. As a result, higher accuracy can be obtained in CMAR.

X. Yin and J. Han proposed a new classification approach CPAR (Classification based on Predictive Association Rules) [4].Both associative classification and rule-based classification have some advantages. The CPAR approach combines these advantages. Associative classification approach generates more association rules. High processing is required for these rules. But it results in high processing overhead. This limitation is overcome by CPAR method. In CPAR approach such large numbers of association rules are not generated. The greedy algorithm is used by CPAR approach. From training dataset, rules are generated. Traditional rule-based classifier can be used to generate and test rules. But some important rules are missed. This problem can be solved by CPAR. CPAR approach generates more rules to include important rules. The associative classification approach also has overfitting problem. This problem is solved by CPAR approach. The accuracy can be used by CPAR to evaluate each rule and k rules are used for prediction. The CPAR approach is used to create predictive rules of good quality. These rules are generated directly from dataset but in a smaller quantity. CPAR generates each rule by taking into account previously generated rules. In such a way that CPAR is used to solve the problem of generating repetition of rules. Dynamic programming is used by CPAR approach. As a result of rule generation, redundant calculation can be avoided.

J. Pei proposed the prefixspan approach for mining sequential patterns by pattern growth [5]. The sequential pattern mining is an important task in data mining. Frequent subsequences can be discovered as patterns in a database of sequence. To effectively mine sequential patterns, sequential pattern growth approach can be used. This approach is projection based. A sequence database is recursively calculated into smaller projected databases. By expanding locally frequent patterns sequential patterns are increased in a database. A pattern-growth approach is also called as divide and conquer approach. It is an extension of FP growth. Pattern growth algorithm is very effective. It is used to mine frequent patterns. There is no need to do the candidate generation. There are various pattern growth-based sequential pattern mining methods. prefixspan is one of the pattern based method. It is an efficient pattern based algorithm. The physical projected dataset is generated in a prefixspan method. A pseudoprojection technique is used in prefixspan to reduce these datasets.

T. P. Exarchos proposed a two-stage methodology for sequence classification [6]. It is based on sequential pattern mining and optimization. The sequential pattern mining is used in the first stage. The sequential patterns are calculated which is used to create a sequence classification model. Then, classes and sequential patterns use weights. The optimization technique is used in the second stage. In this technique, weights are tuned. As a result, optimal classification accuracy is achieved. In this technique the optimization is very time-consuming.

 ISSN: 2454-4248
458 – 461

X. Zhang proposed a highly compact and accurate associative classifier [7]. To build an effective associative classifier, a GARC (Gain based Associative Rule Classification) type approach can be used. This approach is called as GEAR (Gain based Effective Association Rule Classification). A good classification accuracy can be achieved in GEAR. The smaller set of rules can be generated in GEAR. The redundancy/conflicts resolution, the compact set and rule intuition are all advantages of GEAR approach.

L. T. Nguyen proposed a lattice-based approach [8]. In decision support systems classification is an important task. The CBA (Classification based on Association Rules) can be used to mine CARs (Class Association Rules). High accuracy can be obtained by CBA method than heuristic and greedy methods. But more time can be taken to mine CARs. The execution time required to mine rules can be reduced by the lattice based approach. In this approach lattice structure of class rules can be used. The lattice structure includes nodes. Each node consists of values of attributes and their information. The main objective of lattice structure is to compare rule generated from lattice node with all its parent nodes. By comparing whether a rule is redundant or not can be determined. If a parent node's confidence rule is higher than the current node, then the current node generated rule can be decided as a redundant. In this approach, there is no need to check a rule that is generated with a lot of other rules generated. As a result, the time required to mine rules can be surprisingly reduced.

C. Zhou proposed an itemset based sequence classification approach [9]. This approach includes a method called as sequence classification based on interesting itemsets. The cohesion and the support of the itemset are required to determine interestingness of an itemset. The discovered item sets are used to generate confident classification rules and present two different ways to build a classifier. The CBA (Classification based on associations) method is used to build the first classifier. To get better results a new ranking strategy for the generated rules is used. The second classifier is based on the approach in which rules ranked by first measuring their value specific to the new data object. This method improves the accuracy of the classifier but doesn't consider cohesion of itemset. It is considered only one type of pattern (itemset), but it is not applicable to another type of pattern.

Cheng Zhou proposed a sequence classification based on interesting patterns [10]. All these approaches mine the frequent and confident patterns to build a classifier. But cohesion of the pattern is not considered in above approaches. To overcome this problem, the cohesion of the pattern is considered to identify interesting patterns.

III. SEQUENCE CLASSIFICATION BASED ON INTERESTING PATTERN

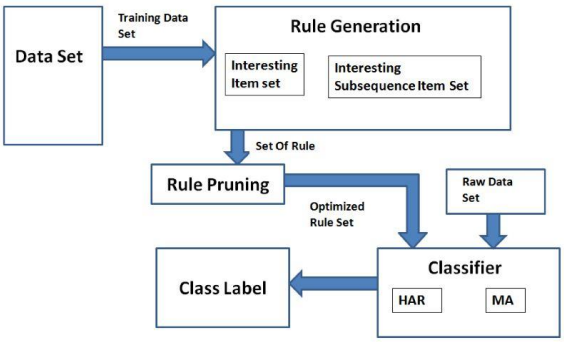

Fig.1. System architecture of improving pattern based sequence classification

Fig.1 shows system architecture of improving pattern based sequence classification. It consists of rule generation, rule pruning and building classifiers. Training dataset can be used as input in rule generation. Two variants can be used for rule generation. The first variant is rule generation using interesting itemsets. The Second variant is rule generation using interesting subsequences. Rule pruning aims to generate a set of rules by using discovered interesting patterns. However, the large number of patterns leads to a large number of rules. So there is a need to prune rules and find the subset of rules of high quality to build an efficient and effective classifier. An optimized set of rules obtained from rule pruning can be used to build the classifier.

IV. CONCLUSION

In this paper, a sequence classification method based on interesting patterns is proposed. Two concrete classifiers SCIP_HAR and SCIP_MA are presented. The cohesion of the itemset and another type of pattern (subsequences) is not considered in this technique. The Pattern Based Sequence Classification technique aims to overcome this limitation. This technique includes a method called Sequence Classification based on Interesting Patterns (SCIP). Interesting patterns are of two types. Itemsets and subsequences can be mined firstly. By using these mined patterns the classification rule can be constructed. The SCIP_HAR and SCIP_MA can be used to build classifiers. Therefore SCIP can be proved effective by considering cohesion of the pattern.

REFERENCES

[1] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 1998, pp. 80–86.

[2] B. Liu, Y. Ma, and C.-K. Wong, "Classification using association rules: Weaknesses and enhancements," in Proc. Data Mining Sci. Eng. Appl., 2001, pp. 591–605.

ISSN: 2454-4248
458 – 461

[3] W. Li, J. Han, and J. Pei, "Cmar: Accurate and efficient classification based on multiple class-association rules," in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 369–376

[4] X. Yin and J. Han, "Cpar: Classification based on predictive association rules," in Proc. SIAMInt. Conf. Data Mining, 2003, pp. 331–335.

[5] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, Mining sequential patterns by patterngrowth: The prefixspan approach," IEEE Trans. Knowl. Data Eng., vol. 16, no. 11, pp. 1424–1440, Nov. 2004

[6] T. P. Exarchos, M. G. Tsipouras, C. Papaloukas, and D. I. Fotiadis, "A two-stage methodology for sequence classification based on sequential pattern mining and optimization," Data Knowl. Eng., vol. 66, no. 3, pp. 467–487, Sep. 2008.

[7] X. Zhang, G. Chen, and Q. Wei, "Building a highly-compact and accurate associative classifier," Appl. Intell., vol. 34, no. 1, pp. 74–86, 2011.

[8] L. T. Nguyen, B. Vo, T.-P. Hong, and H. C. Thanh, "Classification based on association rules: A lattice-based approach," Expert Syst. Appl., vol. 39, no. 13, pp. 11 357–11 366, 2012.

[9] C. Zhou, B. Cule, and B. Goethals, "Itemset based sequence classification, "in Machine Learning and Knowledge Discovery in Databases. New York, NY, USA: Springer, 2013, pp. 353–368.

[10] Cheng Zhou, Boris Cule, and Bart Goethals, "Pattern Based Sequence Classification" vol.28, NO. 5, May 2016