IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Sequence Classification by Using Auto Calculation of Support

G. R. Mane [1*], S. B. Bhagate[2]

[1]Dept. of Computer Science and Engineering, D.K.T.E‘S TEI (An Autonomous Institute), Ichalkaranji, India.
[2] Dept. of Computer Science and Engineering, D.K.T.E‘S TEI (An Autonomous Institute), Ichalkaranji, India.

**Corresponding Author: goutami.kamat@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 23/Jul/2018, Published: 31/July/2018

Abstract — Sequence classification is an efficient task in data mining. Sequence classification problem can be solved by rules that consist of interesting patterns. Another major problem in data mining is pattern mining. In pattern mining, patterns can be used as rules. These rules may be more accurate or simpler to understand while classifying the data object. The cohesion and support of the pattern are used to define interestingness of a pattern. The degree of interest in patterns in a given class of sequences can be measured by combining these two factors. The patterns found can be used to generate reliable classification rules. There are two different ways to build a classifier. The first classifier consists of advanced classification methods that rely on association rules. In the second classifier, the value belonging to the new data object is first measured then the rules are ranked. A well-known methods of association classification are CBA (Classification based on Association rules), CMAR (Classification based on Multiple class-Association Rules), and CPAR (Classification based on Predictive Association Rules) etc. mine the frequent and confident patterns for building a classifier. All these approaches do not consider the cohesion of a pattern and applicable to only one type of pattern. These limitations can be overcome by taking into account a cohesion factor to define interestingness of pattern and can consider another type of pattern.

Keywords— Sequence classification, interesting patterns, classification rules.

I. Introduction

Data mining is the process of mining information from data sets and translating them into understandable structures for further use. Data mining is also known as knowledge discovery (KDD) in data. It is the procedure to analyse concealed patterns of data according to different viewpoints. After analysing data they are classified into valuable information and assembled in data warehouses. In data mining association rules are created by analysing data for frequent patterns. Association rule mining is a process to find frequent patterns, associations and correlations from data sets. These datasets can be found in different types of databases such as transactional databases, relational databases etc. Data mining uses confidence and support factors to discover the most efficient relationships within the data.

A sequence is generally an ordered list of events. Ordinal data can often be found in many important settings, such as web usage logs, videos, biological structures, and text. A sequence may have a class label. In traditional sequence classification, only one category label can be assigned to each sequence. After classification, the entire sequence can be provided to the classifier prior to classification. For example, suppose a sequence of indications of a patient that lasts to a very long time then according to that situation the health state of the patient may vary. A streaming sequence is a sequence that is virtually unlimited. In a streaming sequence only one class label can be predicted. But it is very convenient to predict a sequence of labels. The strong sequence classification task aims to solve the problem. Data sets used in sequence classification tasks fall into two categories. In the first case, the category of a sequence is determined by some items that appear internally, but not always in the same order. In the second case, the category of the sequence is determined by the majority of the items that appear in the 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 function and structure 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 behaviour 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.

The sequence classification methods are divided into three major types. The classification that based on features is first type. A sequence is converted into a feature vector. After transforming into feature vector, conventional classification methods can be applied in the classification. The selection of feature is a significant task in this type of methods. The second category is sequence classification that based on distance. The distance function measures the resemblance between sequences and decides the quality of the classification efficiently. The third category is model based classification. Hidden Markov Model (HMM) and other statistical models can be used to classify sequences. Sequence classification is suitable to large number of applications. It can be defined as assigning class label to new sequences. There exist a number of studies which integrate pattern mining techniques and classifications such as CBA, sequential pattern based sequence classifier, the Class by Sequence (CBS) algorithm and so on. 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 determine a small ruleset in the database classification rule mining is used. The revealed set of rules are used to build a classifier. To discover the rules in the database association rule mining is used. 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 techniques can be integrated, it may produce better results. Relevance classification is used as an integrated framework. Integration of association rules mining and classification can be done through CARs. The integration framework is used to apply association rule mining to classification tasks. Some problems such as understandability problem, a detection of interesting or valuable rules can be solved by this integration. But it requires discretization of constant 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 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 unstable 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 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.

C. Zhou proposed an itemset based sequence classification approach [10]. 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 confident classification rules are generated by using revealed itemsets. To form a classifier two different methods are used. The CBA 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 [11]. 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. Methodology

Existing pattern mining techniques do not consider cohesion of pattern, which affects performance of classification. It can be overcome by Sequence Classification based on Interesting Patterns (SCIP) which uses cohesion and support of the pattern. It aims to mine interesting patterns with improved accuracy in classification and reducing number of user chosen parameters needed for classification.

The system architecture of improving pattern based sequence classification is as follows:

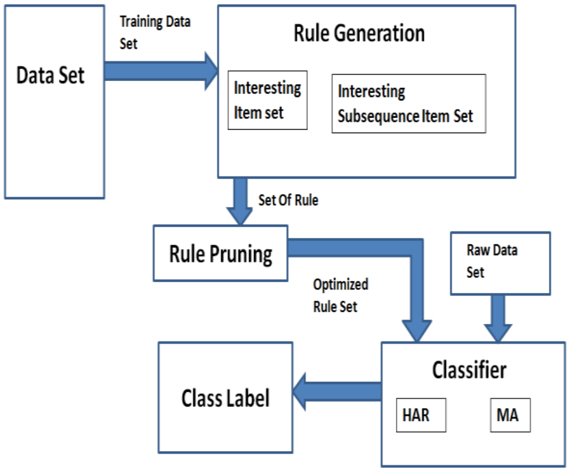

Fig. 1. System architecture.

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 create 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 form an efficient and effective classifier. An optimized ruleset obtained from rule pruning can be used to build the classifier.

Improving Pattern Based Sequence Classification consists of three modules.

1. Rule generation
   - Using interesting item set
   - Using subsequence item set
2 .Rule pruning

3. Building classifier

Rule Generation [11]

The rule generation module takes training dataset as input and uses two variants. First variant is rule generation using interesting itemsets. Generating interesting itemset uses A_n as the set of frequent n-item sets, C_n as the set of candidate n-item sets and T_n as the set of interesting n-item sets. S_k is the set of sequences with class label L_k. The minimum interestingness threshold min_int is a user chosen parameter. The min_support is calculated by using an equation of min_sup(x) = (e raise to (-ax-b)) + c, where x is number of transactions and a,b,c are positive constants. So the equation min_sup(x) = (-0.4x-0.2) is used for calculated auto support. The max_size parameter can be used for the interesting itemsets with a size smaller than or equal to max_size and these itemsets can be used as output. In this algorithm, first it can find the frequent items from S_k and stores in A1. Then from that frequent items interesting itemsets can be found and stores in T_1. All interesting itemsets of size n (max_size $\geq$ n $\geq$2) can be discovered. First frequent itemsets of size n-1 can be used to generate the candidate item sets C_n. The frequent itemsets can be stored from C_n into A_n and the interesting itemsets can be stored from A_n to T_n. Finally the set of all interesting itemsets in S_k can be stored in X_k as output.

The Second variant is rule generation using interesting subsequences. This algorithm uses S_k,min_int and max_size as input parameters. First frequent items or 1-sequences can be found and stores in $A_{1.}$ Then from that frequent subsequences interesting subsequence can be found and stores in T_1 by using a formula of F_k (s').

$$F_k(s') = \frac{|N_k(s')|}{|S_k|}$$

where, s' is subsequence, $F_k(s')$ is the support of a subsequences', S_k is the set of sequences carrying class label $L_{k,}$ N_k (s') is the set of sequences that contain s' labelled by class label L_k. To get interesting n- sequences ($2 \leq n \leq$ max_size) Enumerate-Sequence technique can be used. In this technique any two l-sequences α_i and α_j that share the same (l-1) length can be used. The candidate sequence s' of length (l+1) can be generated by adding the last item in α_j to α_i. The cohesion of subsequence s' can be determined by computing $I_k(s')$. Finally the complete set of interesting subsequences can be stored in Y_k as output.

Rule pruning [11]

Discovered interesting patterns can be used to generate a set of rules. However, the no of patterns is very large which leads to a large number of rules. So there is a need to prune rules and find subset of rules of high quality to build an efficient and effective classifier. Rule pruning algorithm consists of two steps. The training dataset D, a set of confident rules R and coverage threshold δ can be used as input parameters. In first step the set of rules are sorted according to definition of rule pruning. This definition can be described as below.

Suppose there are two rules r_i and r_j in R. The rule r_i has higher precedence than r_j if:

1. the confidence of $r_i > r_j$, or
2. the confidence of r_i and r_j is the same, but the interestingness of $r_i > r_j$, or
3. both the confidence and interestingness of r_i and r_j are the same, but the size of $r_i > r_j$
4. all of the three parameters are same, but r_i is generated earlier than r_j.

The second module is rule pruning. Rule pruning is required to find a subset of rules of high quality. By applying these steps good rules can be obtained for classifying. The second step of this algorithm consists of pruning the rules using database coverage method. For each rule r in sorted R, the dataset D can be used to find all the data objects accurately classified by r. The count of those data objects can be increased. If it correctly classifies a data object, the rule r can be stored in to PR. The data object can be removed if the count of data object passes the coverage threshold. Finally the new set of rules PR can be considered as output.

Building classifiers [11]

Optimized set of rules obtained from rule pruning are used to build the classifier. There are two methods for building the classifier. First method is called as SCIP_HAR (**HAR**MONY based classifier). HARMONY calculates the score of a class label L_k when classifying a new data object. The second method is called as SCIP_MA (**Ma**tching cohesive rules based classifier). In this method not only confidence of rule but also cohesion of the rule is considered. The cohesion of the rule in new data object is included into measure of the correctness of a rule to classify the object. Building classifier algorithm consists of getting the default rule first. The default rule can be calculated by using default rule technique. The training dataset D and pruned rules PR can be used as input parameters. In this technique the data object that matches the rules in PR can be deleted. The counter can be used to indicate how many times each class label appears in the remainder of the dataset. The label with largest counter can be set as the default class label L_d. Finally the default rule can be generated and stored into default_r as output.

After finding the default rule the building classifier algorithm can be performed the second stage of it. The second stage consists of a technique called classifying a new sequence. The PR, default_r, the top λ rules and a new unclassified data object can be considered as input parameters. This technique aims to find the rules that can be matched the given data object and can be stored into MR. Then two different cases can be handled. In the first case if the size of MR > 0, the r.value of every rule in MR can be calculated. The rules can be sorted according to r.value. Then the score of a rule for SCIP_HAR and SCIP_MA can be calculated. The given data object can be classified by using top λ rules and can be stored into CR. The sum of r.values of each rule in CR can be calculated according to their class labels. Finally the sum which is largest can be obtained as the class label. The second case can be considered when MR is empty. The class label of the default rule can be obtained if MR is empty.

IV. Results and Discussion

The system is evaluated by using the auto calculation of min_support threshold of SCIP_HAR and SCIP_MA classifier and calculated the accuracy by using precision method. The different support thresholds and its corresponding accuracy is shown in table below.

Table 1. The different support thresholds and corresponding accuracy

Support	**Accuracy using SCIP_HAR**	**Accuracy using SCIP_MA**
0.1	86.38	86.52
0.118	86.07	86.19
0.125	85.69	85.90
0.133	85.21	86.14
0.15	84.79	86.59

Following fig.2. shows the line graph of auto support classification. The graph describes the impact of the support threshold on accuracy. Observation shows that, the accuracy decreases when the support increases in case of SCIP_HAR classifier. But there is a slight improvement in accuracy of SCIP_MA classifier as compared to SCIP_HAR classifier.

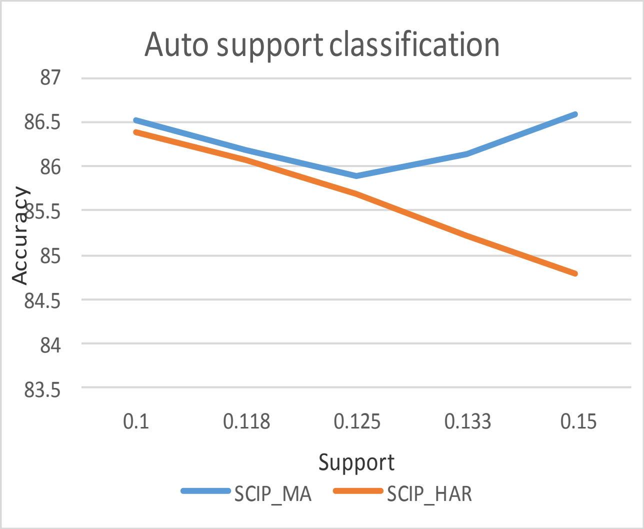

Fig.2. Line graph of auto support classification

V. Conclusion and Future Scope

Sequence classification method based on interesting patterns consists of two stages. First stage of this rule generation makes use of two variants. First variant is of using interesting item sets and second variant is of using interesting subsequences. To mine interesting patterns itemsets and subsequences is first task in SCIP. The discovered interesting patterns are converted into classification rules. The large number of patterns leads to large number of rules. To improve the accuracy of the classifier there is a need to reduce large number of rules. Rule pruning technique is used to prune unnecessary rules and find a set of rules of high quality. The optimized set of rules obtained from rule pruning are used to build the classifiers. These classifiers are used to determine the class to which a new instance belongs. The system assume that two events can never occur at the same time. In future, the system can be explored to a more general setting where several events may sometimes occur at the same time stamp.

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.

[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] M. J. Zaki, C. D. Carothers, and B. K. Szymanski, “Vogue: A variable order hidden Markov model with duration based on frequent sequence mining,” ACM Trans. Knowl. Discovery Data, vol. 4, no. 1,p. 5, 2010.

[8] 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.

[9] 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.

[10] 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.

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

Authors Profile

Miss. Goutami Ramakant Mane pursed Bachelor of Engineering from KIT’s college of Engineering, Shivaji University, Kolhapur, India in 2009. She is pursuing Master of Technology in Computer Science & Engineering from DKTE Society's Textile & Engineering Institute, (An Autonomous Institute), Ichalkaranji, *416115,* India. security

Mr. Suhas B. Bhagate pursed Bachelor of Engineering from Shivaji University, Kolhapur in 2003 and Master of Engineering from Walchand College of Engineering, Sangli in Shivaji University, and Kolhapur in year 2011. He is pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering, D.K.T.E. Society’s Textile and Engineering Institute, Ichalkaranji since 2004. He is IEEE Graduate Student Member. He has published more than 10 research papers in reputed international journals. His main research work focuses on Visual Cryptography Algorithms, Data Structures, Big Data Analytics and Data Mining.