International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue X, October- 2016

Categorization of Sequential Data using Associative Classifiers

Mrs. R. Meenakshi , MCA., MPhil., Research Scholar,
Mrs. J.S. Subhashini, MCA., M.Phil., Assistant Professor,
Department of Computer Science,
SSM College of Arts & Science, Komarapaliyam, Tamilnadu, India

Abstract

The association rule mining techniques are applied to discover the frequent patterns. Frequent rules are discovered with minimum support levels. The classification techniques are employed to assign category information for the transactions. The learning phase identifies the class patterns using the labeled transactions. The testing phase assigns the class labels for the unlabeled transactions with learned patterns. The associative classification method integrates the association rule mining and classification techniques for classification process.

The sequence classification methods are applied to assign class labels for data sequences. The Sequence Classification based on Interesting Patterns (SCIP) scheme is adapted to perform classification on sequential data elements. Support and cohesion measures are estimated to discover the interesting rules. Interesting item set discovery, interesting subsequence identification, rule pruning and classifier building operations are carried out under the SCIP scheme. The class rules are updated in the classifier. The classifier is used in the class assignment process.

The Sequence Classification based on Refined Interesting Patterns (SCRIP) scheme is constructed to perform the data classification with refined rules. Redundant timestamp information are analyzed in the interesting subsequence discovery process. The rule refinement process is performed with automatic threshold estimation for support maximization process. The patterns are derived for each class level to build the classifier with optimal rules. The SCRIP scheme is applied to classify the Yeast gene expression data values. The dimensionality reduction process is called to minimize the gene expression data groups. Rule retrieval rate, time complexity and accuracy level parameters are used to verify the system performance levels.

1. Introduction

Mining frequent sequential patterns from a sequence database is an important data mining problem which has been attracting researchers for more than a decade. Dozens of algorithms find sequential patterns effectively have been proposed. However, relatively few researchers have addressed the problem of reducing redundancy, ranking patterns by interestingness, or using the patterns for solving further data mining problems.

Redundancy is a well-known problem in sequential pattern mining. Let us consider the *Journal of Machine Learning Research* (JMLR) dataset which contains a Mining compressing sequential patterns database of word sequences, each corresponding to an abstract of an article in the *Journals of Machine Learning Research*. The 20 most frequent closed sequential patterns ordered by decreasing frequency. This set of patterns is clearly very redundant, so many patterns with very similar meaning are shown to users. Besides redundancy issues, the set of frequent patterns usually contain trivial and meaningless patterns. In fact, the set of frequent closed patterns contains random combinations or repeats of frequent terms in the JMLR abstracts such as *algorithm, result, learn, data* and *problem*. These patterns are meaningless given our knowledge about the frequent terms.

To solve these issues, we have to find alternative interestingness measures rather than relying on frequency alone. For itemset data, an interesting approach has been proposed recently. The Krimp algorithm mines patterns that compress the data well using the minimum description length (MDL) principle. This approach has been reduce redundancy and generate patterns that are useful for classification, component identification and change detection. We extend these ideas to sequential data. The key issue in designing an MDL-based algorithm for sequence data is the encoding scheme that determines how a sequence is compressed given some patterns. In contrast to itemsets we need to consider the ordering of elements in a sequence and need to be able to deal with gaps, as well as overlapping and repeating patterns; all properties that are not present in itemset data.

2. Related Work

The existing sequence classification techniques deploy a number of different approaches, ranging from decision trees, Na¨ıve Bayes, Neural Networks, K-Nearest Neighbors (KNN), Hidden Markov Model (HMM) and, lately, Support Vector Machines (SVMs) [7].

In this section, we give an overview of pattern-based classification methods [1]. Most such work can be divided into the domains of classification based on association rules and classification based on sequential

patterns [2]. The main idea behind the first approach is to discover association rules that always have a class label as their consequent. The next step is to use these patterns to build a classifier, and new data records are then classified in the appropriate classes [3]. The idea of classification based on association rules (CBA) was first proposed by Liu et al. In another work, Li et al. proposed CMAR, where they tackled the problem of overfitting inherent in CBA. In CMAR, multiple rules are employed instead of just a single rule. Additionally, the ranking of the rule set in CMAR is based on the weighted Chi-square of each rule replacing the confidence and support of each rule in CBA. Yin and Han proposed CPAR which is much more time-efficient in both rule generation and prediction but its accuracy is as high as that of CBA and CMAR.

The concept of sequential pattern mining was first described by Agrawal and Srikant and further sequential pattern mining methods, such as Generalized Sequential Patterns (GSP), SPADE, PrefixSpan and SPAM, have been developed since. A number of sequence classifiers have been based on these methods.

Lesh et al. combined sequential pattern mining and a traditional Na¨ıve Bayes classification method to classify sequence datasets [4]. They introduced the FeatureMine algorithm which leveraged existing sequence mining techniques to efficiently select features from a sequence dataset. The experimental results showed that BayesFM is better than Na¨ıve Bayes only. Although pruning is used in their algorithm, there was still a great number of sequential patterns used as classification features. As a result, the algorithm could not effectively select discriminative features from a large feature space.

Tseng and Lee proposed the Classify-By-Sequence (CBS) algorithm for classifying large sequence datasets. The main methodology of the CBS method is mining classifiable sequential patterns (CSPs) from the sequences and then assigning a score to the new data object for each class by using a scoring function, which is based on the length of the matched CSPs. They presented two approaches, CBS ALL and CBS CLASS. In CBS ALL, a conventional sequential pattern mining algorithm is used on the whole dataset. In CBS CLASS, the database is divided into a number of sub-databases according to the class label of each instance. Sequential pattern mining was then implemented on each sub-database. Experimental results showed that CBS CLASS outperforms CBS ALL. Later, they improved the CBS CLASS algorithm by removing the CSPs found in all classes [8]. Furthermore, they proposed a number of alternative scoring functions and tested their performances. The results showed that the length of a CSP is the best attribute for classification scoring.

Exarchos et al. [6] proposed a two-stage methodology for sequence classification based on sequential pattern mining and optimization. In the first stage, sequential pattern mining is used, and a sequence classification model is built based on the extracted sequential patterns. Then, weights are applied to both sequential patterns and classes. In the second stage, the weights are tuned with an optimization technique to achieve optimal classification accuracy. However, the optimization is very time consuming, and the accuracy of the algorithm is similar to FeatureMine. Additionally, several sequence classification methods have been proposed for application in specific domains. Exarchos et al. [5] utilised sequential pattern mining for protein fold recognition, while Zhao et al. [9] used a sequence classification method for debt detection in the domain of social security.

The main bottleneck problem for sequential pattern based sequence classification being used in the real world is efficiency. Mining frequent sequential patterns in a dense dataset with a large average sequence length is time and memory consuming. None of the above sequence classification algorithms solve this problem well.

3. Pattern Based Sequence Classification

Sequential data is often encountered in a number of important settings, such as texts, videos, speech signals, biological structures and web usage logs, where a sequence is generally an ordered list of singletons. Because of a wide range of applications, sequence classification has been an important problem in statistical machine learning and data mining. The sequence classification task can be defined as assigning class labels to new sequences based on the knowledge gained in the training stage. There exist a number of studies integrating pattern mining techniques and classification, such as classification based on association rules (CBA), sequential pattern based sequence classifier, the Classify-By Sequence (CBS) algorithm and so on. These combined methods can produce good results as well as provide users with information useful for understanding the characteristics of the dataset. In practice, most datasets used in the sequence classification task can be divided into two main cases. In the first case, the class of a sequence is determined by certain items that co-occur within it, though not always in the same order.

In this case, a classifier based on sequential patterns will not work well, as the correct rules will not be discovered, and, with a low enough threshold, the rules that are discovered will be far too specific. In the other case, the class of a sequence is determined by items that occur in the sequence almost always in exactly the same order. At first glance, a sequence based classifier should outperform an itemset based classifier in this situation.

Itemset based classifiers will do better when the pattern sometimes occurs in an order different from the norm. This robustness means that itemset based classifiers can handle cases where small deviations in the subsequences that determine the class of the sequences occur. A simpler candidate generation process, itemset based methods are much faster than those based on sequential patterns.

The above observations motivate the proposed research, Sequence Classification based on Interesting Patterns (SCIP). First of all, we present algorithms to mine both types of interesting patterns — itemsets and subsequences. As a second step, we convert the discovered patterns into classification rules, and propose two methods to build classifiers to determine the class to which a new instance belongs. In the first method, we select the rules to be applied based on their confidence, while the second uses a novel approach by taking into account how cohesive the occurrence of the pattern that defines the rule is in the new instance. Finally, we step away from pattern based classification and evaluate the quality of our pattern miner by using our patterns as features in a variety of feature based classifiers.

When looking for interesting patterns in sequences, a pattern is typically evaluated based on how often it occurs. The proximity of the items that make up the pattern to each other is important, too. If two classes are determined by exactly the same items, traditional pattern based classifiers may struggle. For example, if class A is determined by the occurrence of ha; bi with a shortest interval of 2 and class B by the occurrence of ha; bi with a shortest interval of 5, pattern ha; bi will not be enough to be able to tell the difference, and this will be solved by considering the cohesion information. Therefore, we use both cohesion and support to define interesting patterns in a sequence dataset. Finally, we utilise these interesting patterns to build classifiers. Experiments show the effectiveness of our classifiers and test which kind of pattern works better in given circumstances. An additional advantage of our method is that the classifiers we build consist of easily understandable rules.

In this paper, we build on our previous work both improving and extending it. Firstly, we now present algorithms to mine two different types of patterns. Secondly, we replace the CBA based classifier used previously with a new classifier, based on the HARMONY scoring function, which achieves better performance. In addition, we now deploy a new top-k strategy for all the presented classifiers, instead of using only the highest ranked rule, as we did in our previous work. Moreover, we now propose using the discovered patterns as features in order to transform each sequence into a feature vector. We present a new feature vector representation approach by setting each feature value as the cohesion of the feature in a sequence. After this, we apply machine learning algorithms for sequence classification and find that this new feature vector representation approach outperforms the traditional presence-weighted feature vector representation approach.

4. Issues on Pattern Based Sequence Classification

The association rule mining methods are applied to discover the frequent rules or patterns from the transactional data values. The support and confidence levels are used to filter the frequent rules. The classification methods are applied to assign class labels for the unlabeled transactions. The associative classification method integrates the association rule mining methods and classification methods. The Sequence Classification based on Interesting Patterns (SCIP) scheme is employed to discover the class labels with derived patterns. Support and cohesion measures are used to identify the interesting item sets. The interesting item set discovery, interesting subsequence identification, rule pruning and classifier construction tasks are carried out in the SCIP scheme. The classification process is performed with the constructed classifier. The following issues are identified from the existing system.

- Dimensionality reduction operations are supported in the system
- Redundant timestamp values are not handled in the sequence identification process
- Rule refinement process is not handled
- Classifier is build with noisy rule information

5. Associative Classifiers for Sequence Classfication

The sequence classification scheme is used to perform the yeast gene classification process with pattern information. The association rule mining method supports the pattern extraction on labeled data values. The discriminative patterns are adapted to carry out the gene classification process. The associative classification system is optimized to mine rules on continuous data environment. The system is improved to assist the user to estimate the threshold value with support boundaries. Support distribution based rule summarization model is also proposed for frequent item-set identification. The Sequence Classification based on Interesting Patterns (SCIP) and Sequence Classification based on Refined Interesting Patterns (SCRIP) schemes are employed for the gene data classification process.

The system is designed to analyze yeast gene expressions. High dimension and dense data analysis is provided in the system. Pattern mining operations are performed on labeled data values. The system is divided into six major modules. They are gene data analysis, relationship analysis, pattern mining process, rule

summary analysis, classifier building process and classification process.

The gene data analysis module is designed to analyze yeast gene expression data values. Candidate set and item sets are prepared under relationship analysis module. Pattern mining process module is designed to mine frequent rules. Rule distribution is identified under rule summary analysis module. The classifiers are constructed under the classifier building process. The classification process module is designed to perform the data classification process.

5.1. Gene Data Analysis

The yeast gene expression data values are collected from UCI (University of California Irwin) machine learning repository. The data values are selected from CSV files. The data values are transferred into Oracle database. Data cleaning is performed to correct noisy data values. The yeast gene data values are listed with its weight values. The optimized gene data values are prepared with the attribute group information. The timestamp information are also optimized in the optimization process. Redundant timestamp details are considered in the sequence preparation process. The attribute summary shows the attribute group names and attribute count details. The gene data values are extracted and analyzed with timestamp values.

5.2. Relationship Analysis

The candidate set and item sets values are prepared with labels. Attribute name, value and transaction labels are used to prepare candidate sets. Item sets are prepared using the combination of candidate sets. Frequency values are updated for each candidate set and item set. Candidate sets and item sets are listed in separate forms. The system also lists the labeled candidate sets and labeled item sets in different forms.

5.3. Pattern Mining Process

The pattern mining process is applied to extract frequent patterns. Minimum support and minimum confidence values are used to fetch frequent rules. Rules are filtered on labeled patterns. Rules are filtered for each label levels. The interesting pattern extraction also uses the support and cohesion values. The attribute frequency shows the attribute name, value and frequency details. The support / confidence details for all the item sets are listed in the separate form. The interesting patterns are listed with user threshold values.

5.4. Rule Summary Analysis

Rule summary analysis is carried out on item sets with support and confidence values. Support distribution is prepared with the support ratio values. Rule summary is prepared for each support level. Rule ranges and rule count values are produced in rule summary. The rule refinement is carried out with rule summary information.

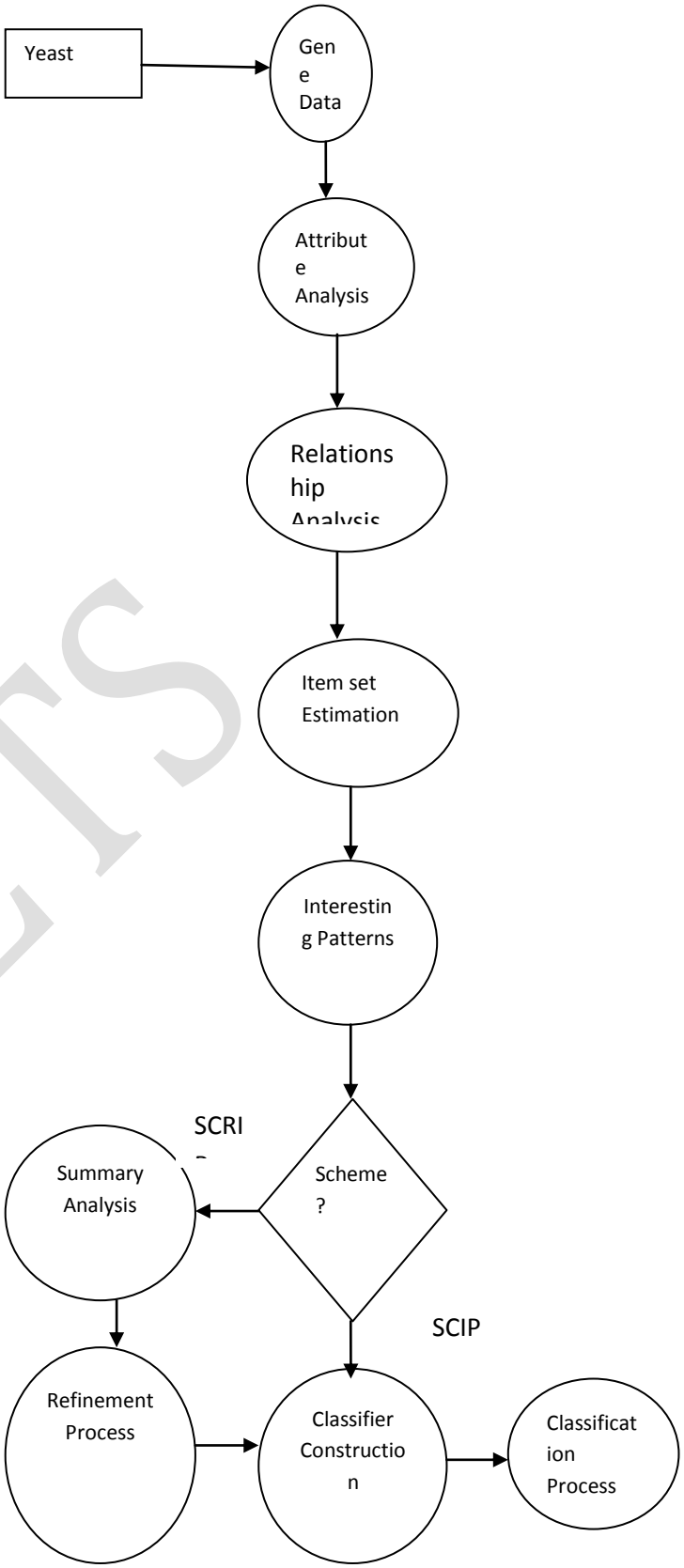

Figure. No: 5.1: Yeast gene classification framework

5.5. Classifier Building Process

The classifier construction process is used to build the classifier with discriminative rules. The pattern extraction process is applied on discriminative pattern collections. The sequence construction process and rule identification operation are called on the interesting item set values. The classifier construction is carried out using two ways. They are Sequence Classification based on Interesting Patterns (SCIP) scheme and Sequence Classification based on Refined Interesting Patterns (SCRIP) scheme. The Interesting Pattern (IP) method is applied to fetch the patterns that are used for the classifier rules. The Refined Interesting Patterns (RIP) method is used to discover the patterns for classifier rules. The RIP method eliminates the infrequent and irrelevant rules from the discovered patterns. Support maximization threshold is used for the refinement process. The classifier is build to extract the rules for all the classes in the Yeast data values.

5.6. Classification Process

The classification process is designed to perform the gene expression categorization process. The associative classification model based classification framework is used in the system. Pattern based sequence classification is carried out on the gene expression data values. The Sequence Classification based on Interesting Patterns (SCIP) and Sequence Classification based on Refined Interesting Patterns (SCRIP) schemes are used in the classification process. Redundant timestamp information are analyzed to correct the sequence constructions process. The pattern refinement technique reduces the classifier rules. Time and accuracy parameters are estimated and analyzed in the classification process.

6. Conclusion and Future Enhancement

The sequence classification methods are applied to perform the classification operations on any domain. The gene expression classification process is carried out using the Sequence Classification based on Interesting Patterns (SCIP) scheme. The interesting item set discovery, interesting subsequence identification, rule pruning and classifier building operations are performed in the SCIP scheme. The gene data values are classified with the support of constructed classifier. Refined Interesting Pattern (RIP) discovery mechanism is integrated with the SCIP scheme. The Sequence Classification based on Refined Interesting Patterns (SCRIP) scheme supports the gene expression classification with minimum time complexity and better accuracy levels. The system can be enhanced with the following features. The associative classification scheme can be improved with weighted rule mining methods. The classification process can be upgraded with fuzzy logic, Genetic Algorithm (GA) and optimization based techniques.

References

[1] Cheng Zhou, Boris Cule and Bart Goethals, "Itemset Based Sequence Classification", 2013.

[2] Cheng Zhou, Boris Cule and Bart Goethals, "Pattern Based Sequence Classification", IEEE Transactions On Knowledge And Data Engineering, May 2015.

[3] Chuanren Liu, Kai Zhang and Qiang Yang, "Temporal Skeletonization on Sequential Data: Patterns, Categorization and Visualization", IEEE Transactions on Knowledge and Data Engineering, January 2016.

[4] Chung-Hsien Yu, Wei Ding, Melissa Morabito and Ping Chen, "Hierarchical Spatio-Temporal Pattern Discovery and Predictive Modeling", IEEE Transaction on Data and Knowledge Engineering, April 2016.

[5] Exarchos, T.P., Papaloukas, C., Lampros, C., Fotiadis, D.I.: Mining sequential patterns for protein fold recognition. Journal of Biomedical Informatics 41(1), 165–179 (2008)

[6] Exarchos, T.P., Tsipouras, M.G., Papaloukas, C., Fotiadis, D.I.: A two-stage methodology for sequence classification based on sequential pattern mining and optimization. Data & Knowledge Engineering 66(3), 467–487 (2008)

[7] Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques, third edition. Morgan kaufmann (2011)

[8] Tseng, V.S., Lee, C.H.: Effective temporal data classification by integrating sequential pattern mining and probabilistic induction. Expert Systems with Applications 36(5), 9524–9532 (2009)

[9] Zhao, Y., Zhang, H., Wu, S., Pei, J., Cao, L., Zhang, C., Bohlscheid, H.: Debt detection in social security by sequence classification using both positive and negative patterns. In: Machine Learning and Knowledge Discovery in Databases, pp. 648–663. Springer (2009).