AN ADAPTIVE PATTERN GENERATION IN SEQUENTIAL CLASSIFICATION USING FIREFLY ALGORITHM

Dr. P. Radha[1], M. Thilakavathi[2]

[1]*Head and Assistant Professor, Dept. of Computer Technology, Vellalar College for Women, Erode, Tamilnadu,india*
[2]*Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India*

---***---

Abstract -*Sequence classification is an important task in data mining. This work addresses the problem of sequence classification using rules composed of interesting patterns found in a dataset of labeled sequences. It accompanies a class label and also measures the occurrence of a pattern in a given class of sequence by using the cohesion and the support of the pattern. Discovered patterns are used to generate confident classification rules, and present two different ways of building a classifier. The first classifier is based on an improved version of the existing method of classification based on association rules, while the second ranks the rule by first measuring their value specific to the new data object. Experimental results shows that rule based classifier outperform existing comparable classifier in terms of accuracy and stability. Additionally a number of pattern feature based models that use different kinds of pattern as feature to represents each sequence as a feature vector. Then apply a variety of machine learning algorithms for sequence classification, experimentally demonstrating the pattern that represent the sequence well, and prove effective for the classification task. Hence to overcome all these issues the Firefly algorithm is utilized in the current work.*

Key Words: *Sequential Pattern Mining, Sequence classification, SCII(Sequence classification interesting itemset), (SCIP) Sequence Classification based on Interesting Patterns, firefly algorithms.*

1. INTRODUCTION

Sequence classifications have a broad range of applications such as, information retrieval, genomic analysis, health informatics, finance, and abnormal detection. Different from the classification task on the feature vectors, sequences don't have explicit features. Even with sophisticated feature selection techniques, the dimensionality of potential features may still be very high and the sequential nature of features is too difficult to capture. This makes sequence classification is a more challenging task than classification on feature vectors. There are three major challenges in sequence classification. First, most of the classifier, such as decision trees and neural networks, can only take input data as a vector of features. However, there are no explicit features in a sequence data. Second, even with various feature selection methods that can be transform a sequence into a set of features; the feature selection is far from trivial. The dimensionality of the features space for the sequence data could be very high and the computation can be costly. Third, besides accurate classification results, in some applications, it is need to get an interpretable classifier. Building an interpretable sequence classifier is crucial since there are no explicit features.

2. LITERATURE REVIEW

Zhou.C, Cule.B, Goethals.B, [9] proposed that sequence classification is an important task in data mining. The discovered item sets is used to generate confident classification rules, and present two different ways of building a classifier. The first classifier is based upon the CBA (Classification based on associations) method, but uses a new ranking strategy for the generated is used to rules, is used to achieve better results. The second classifier ranks the rules by first measuring their value specific to the new data object. A sequence classification method is introduced based on interesting item sets named SCII (sequence classification interesting itemset) with two variations. Through experimental evaluation, the SCII methods provide higher classification accuracy compared to existing methods. The experimental results show that SCII is not sensitive to the setting of a minimum support threshold or a minimum confidence threshold. In addition, the SCII method is scalable as the runtime is proportional to the dataset size and the number of items in the dataset. The output rules of SCII are easily readable and understandably represent the features of datasets.

Srikant and Agrawal, [1] proposed an algorithm named GSP (Generalizations and performance improvements) which uses a breadth-first search and bottom up method to obtain the frequent subsequences. It also considers the mining sequential patterns with timing constraints regarding the minimal time gap, maximal time gap and sliding window size. The advantage of the inclusion of timing constraints is that patterns are not satisfying the timing constraints are filtered out. Thus the number of candidate sequential patterns would be reduced.

T. Sutouet al, [7] proposed a parallel Modified PrefixSpan method for extracting recurrent patterns from sequence databases. This method entails the use of several computers which are interlinked in a local area network in parallel. The Modified PrefixSpan method reduces the wild cards sent to the PrefixSpan thereby reducing the computational time in pattern extraction and redressing the functional problem. Unlike PrefixSpan, the Modified PrefixSpan method is able to produce recurrent patterns which include most wild cards.

Bing Liu, Wynne Hsu and Yiming Ma, [10] propose that Classification rule mining aims to discover a small set of rules in the database that forms an accurate classifier. Association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence constraints. The two mining techniques are integrated and the integration is done by focusing on mining a special subset of association rules, called a class association rules (CARs). An efficient algorithm is a given for building a classifier based on the set of discovered CARs. Experimental results show that the classifier built this ways are , in general, more accurate than produced by the state-of-the-art classification system C4.5. An algorithm is presented to generate all the class association rules (CARs) and to build an accurate classifier. The new framework not only gives a new way to construct classifiers, but also helps to solve a number of problems that exist in current classification systems.

Masseglia et al, [5] considered handling time constraints in the previous stage of the data mining process in order to provide better performance. GSP (Generalized sequential pattern) met more real-world requirements than AprioriAll. After all when the database or the number of possible items grows, the voluminous candidate sequential patterns and the increased number of database scans have a huge impact on the performance. Also the methodology cannot find a pattern whose interval between two consecutive items is not in the range and the sequential patterns include only the temporal order of the items.

2.1 Problem Statement

The existing scenario contain the pattern which comprises of both itemset or subsequences (sequential patterns).The support count of a pattern is defined as the number of different sequences in which the pattern occurs; regardless of how many times the pattern occurs in any single sequence the support count of a pattern alone, stop looking at a sequence as soon as it detected for the first time of the pattern. To determine the interestingness of a pattern, however, it is not enough to know how many times the items making up the pattern occur. So it is necessary to know how close pattern appear to each other. Hence to overcome all these issues the Firefly algorithm is used.

2.2 Existing Scenario

In the existing system sequential data are classified by mining interesting patterns in itemset and subsequence. In this algorithm mine both types of interesting patterns - itemsets and subsequences. As a second step, 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, 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. Final step is away from pattern based classification and evaluate the quality of pattern miner by using patterns as features in a variety of feature based classifiers. Initially Apriori-like algorithm is used to find frequent itemsets in the sequential data then interesting itemset can be determined using user defined parameter are min-sup,min-int. Then SPADE algorithm is used to determine frequent subsequence in the sequential data. Then CMAR (Classification based multiple association rule) is used to find a subset of rules of high quality to build an efficient classifier and prune the unnecessary rules.

Finally SCIP-CB algorithm (sequence classification based on interesting patterns classifier building)is presented for building a classifier using the interesting patterns discovered by SCIP-RG (Rule generation), SCIP HAR (Harmony) and SCIP MA are used to score the class labels. Hence, to overcome all these issues the Firefly algorithm is used in the research work.

2.3 Proposed Scenario

In this research work, two important tasks of sequence mining are considered, that is sequence generation and sequence searching. In sequence generation process, important and frequent sequences are generated from the database based on the user-defined minimum support. Many applications have the need to check whether a given search sequence is found in the sequence database or not and some applications have the need to count the occurrence of a given search sequence in the sequence database. This proposed technique intends to explore more events to predict where several events may sometimes occur at the same time stamp. It will attempt to reduce the maximum number of frequencies by reducing the reoccurred events. Better accuracy is detected from comparing the sequence classification with firefly. Parameter value (0, 4) is considered for parameter reduction. This algorithm deals with highly non-linear multi-model optimization problems naturally and efficiently. Firefly can deal with highly non-linear multi-model optimization problems naturally and efficiently .The speed of convergence of firefly is very high in probability of finding the global optimized answer.

3. SYSTEM METHODOLOGY

3.1 Sequence Classification

Sequence classification has a broad range of applications such as genomic analysis, information retrieval, health informatics, finance, and abnormal detection. Even with sophisticated feature selection techniques, the dimensionality of potential features may still be very high and the sequential nature of features is difficult to capture. This makes sequence classification an extra challenging task than classification on feature vectors.

3.2 Dataset

The dataset contains multiple transactions. Each transaction contains the set of items. The dataset used in this research work is taken from Frequent Item Set Mining Repository. mlg.ucd.ie/datasets/bbc.html. Retail dataset is used in this research work. It is a real time dataset collected from a sports news dataset. The dataset consists of the documents from the BBC Sports corresponding to sports area news articles.

3.3 Preprocessing

The preprocessing step is used to reduce the size of the specified dataset and improve the classification results. The preprocessing has been performed on the specified dataset.

The preprocessing step which includes stemming, stop word removal and word count then corresponding results are classification task on simple symbolic sequences and simple time series data. Although there are a few works on multiple variate time stored. Initially the sports news dataset is collected.

Then perform the process of tokenize, stemming, stop word and filtering process on the dataset. Stop words are such as ‘a’, ‘this’, ‘is’ and ‘so’, ‘on ‘are removed as well as dependency words like ‘not’,’ no’ are considered. Stemming is the process where the words suffixes are removed. Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. Data must be preprocessed in order to perform any data mining functionality.

Data preprocessing requires reduplication and other word errors correction, normalization of dates, places and acronyms. Hence the dataset does not include redundancy and irrelevant information. It is in the form of consistent review comments and reduction dataset is maintained for further process.

The transactions consist of unique pattern that are given for each pattern is converted into binary format whereas similar as 1 and otherwise non-similar 0.It will attempt to reduce the maximum number of frequencies by reducing the reoccurred events.

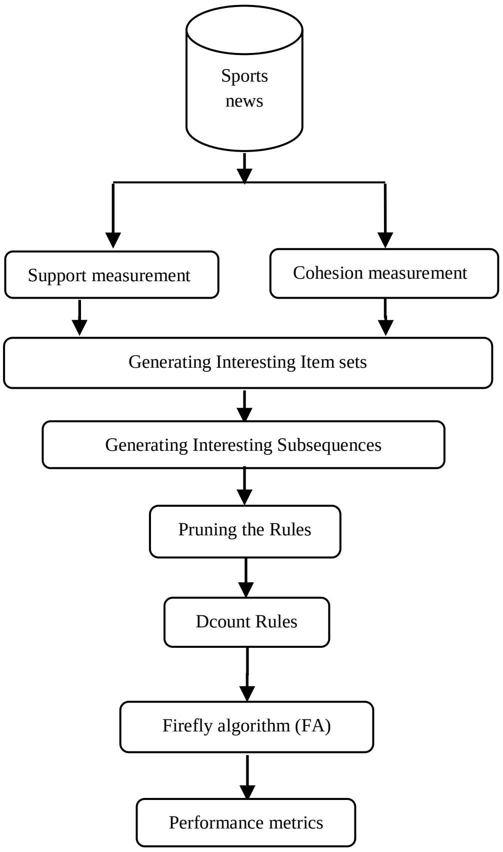

Fig- 1: System Architecture

3.4 Module Description

The modules in this work are listed below

- Definition of an Interesting Pattern
- Generating Interesting Itemsets
- Rule Based Classifiers
- Pruning the Rules
- Building the Classifiers
- Firefly algorithm using classification

Definition of an Interesting Pattern

In this research work the interesting patterns are used. The interestingness of a pattern depends on two factors: its support and its cohesion. Support measures is how many sequences the pattern appears; while cohesion measure show close the items making up the pattern are to each other on average, using the lengths of the shortest intervals containing the pattern in different sequences.

Consider the pattern that could be item sets or subsequences, which has the definition of an interesting pattern based on item sets and subsequences respectively

Support

The support of a pattern P in a given class of sequences S_k can now be defined as
$F_k(P) = \frac{|N_k(p)|}{|s_k|}$
Where P could be X or s′.

Cohesion

All patterns containing just one item are fully Cohesive,

i.e., $C_k(P) = 1$ if $|P| = 1$.

The cohesion of P in a single sequence s is defined as
$C\ (P, s) = \frac{|P|}{W(P,s)}$

Interestingness

In a given class of sequences S_k, define the interestingness of a pattern P as

$$I\kappa(P) = F\kappa(P) \times C\kappa(P)$$

where P could be X or s'.

Given a minimum support threshold min_sup and a minimum interestingness threshold min_int, a pattern P is considered interesting in a set of sequences labeled by class label L_k, if $F\kappa(P) \geq \text{min\_ sup}$ and $I\kappa(P) \geq \text{min\_ int}$.

Rule Based Classifiers

SCIP (sequence classification based on interesting (patterns), consists of two stages, rule generation(called SCIP-RG, with one variant using interesting itemsets(SCII RG) and another using interesting subsequences(SCIS-RG), and classifier building (called SCIP-CB).

Generating Interesting Item sets

The rule generator for interesting itemsets (SCII-RG) generates all interesting item sets in two steps. Due to the fact that the cohesion and interestingness measures introduced in problem statement are not anti-monotonic, prune the search space based on support alone. In the first step, use an Apriori-like algorithm to find the frequent item sets. In the second step, determine which of the frequent itemsets are actually interesting. An optional parameter, max size, can be used to limit the output only to interesting itemsets with a size smaller than or equal to max size. This algorithm is used for generating the complete set of interesting Itemsets in a given class of sequences. Analyze the time needed to evaluate each candidate. The first need is +to find a shortest interval W(X; s) of an itemset X in each sequence. GetInterval function, computes the size of the shortest occurrence of X around time stamp.

Generating Interesting Subsequences

This method is used for generating interesting sub sequences on the well-known SPADE algorithm, which is capable of efficiently finding all frequent sub sequences. To determine the support of any l-sequence, SPADE looks at intersections of the id-lists of any two of its subsequences of length $(l-1)$ since such an id-list keeps a list of the sequences in which a pattern occurs, as well as where in those sequences the items making up the pattern occur.

Pruning the Rules

- ➢ Find all interesting patterns in a given class, all confident classification rules can be found in a trivial step — for each pattern P that is interesting in class L_k, the generate rule P =>L_k.
- ➢ However, the number of patterns is typically very large, which leads to a large number of rules. Reducing the number of rules is crucial to eliminate noise, which could affect the accuracy of the classifier, and to improve the runtime of the algorithm.
- ➢ Therefore try to find a subset of rules of high quality to build an efficient and effective classifier. So the idea introduced in CMAR and prune unnecessary rules using the database coverage method.
- ➢ Before using the database coverage method, first define a total order on the set of all generated rules R. This is used in selecting the rules classifier.

Building the Classifiers

This subsection presents the SCIP-CB algorithm for building a classifier using the interesting patterns discovered by SCIP-RG. CBA has successfully shown the competitive accuracy performance of an associative classifier. However, HARMONY uses a different coring function to improve the overall accuracy of the classifier. When classifying a new data object, HARMONY computes the score of a class label L_K as the sum of the top _ highest confidences of the rules carrying class label L_K and matching the data object. Since HARMONY outperforms CBA. SCIP-CB algorithm is presented for building a classifier using the interesting patterns discovered by SCIP-RG. SCIP HAR and SCIP MA are used to score the class labels.

Finally, get the finding default rule and classifying a sequence. the investigate building classifiers by both itemset rules and sequence rules, the complete set of SCIP classifiers contains four classifiers — two itemset based classifiers(SCII HAR and SCII MA) and two sequence based classifiers(SCIS HAR and SCIS MA).

3.5 Firefly Algorithms

The firefly algorithm chooses the optimum weight in neural network based on the concept of firefly characteristics by using objective function. A firefly is attracted to another firefly despite the consequences of its sex, attractiveness is proportional to their brightness and brightness of each firefly is decided by the landscape of the objective function. Initially the population of fireflies is assigned. Then considered two significant points are formulation of the attractiveness and variation in light intensity. The attractiveness is determined by the brightness of the fireflies in which the objective function is associated. Also, the objective function is used to determine the brightness (I) of the firefly in a specific position. The objective function is to decrease the false error rate (FER) to determine the phishing websites. It can be explained by firefly algorithm.

Find $L(x)$ using equation

L (x) =minimum (FER)

$$If\,(L_m > L_n)$$

Calculate attractive fireflies

$$\beta(r) = \beta_0 . ew^{-\gamma.r^2}$$

Compute the distance between the fireflies

$$r_{mn} = \left||Y_m - Y_n|\right| = \sqrt{\sum_{i=1}^{d} (Y_{m,i} - Y_{n,i})^2}$$

`Move all firefly to the best solution

$$y_m = y_m + \beta_0 * \exp(-\gamma r_{mn}^2) * (y_n - y_m) + \alpha * (rand - \frac{1}{2})$$

4. RESULT AND DISCUSSION

To execute the proposed technique and generate various results using Matlab tool in this environment. The scenario is sports news dataset to discover the sequence pattern and relevant topics for particular pattern. In this section, the analysis has been done for existing and proposed research work by using algorithms. The performance metrics are such as accuracy values and time factors. From the experimental result, the conclusion decides that the proposed method provides higher performance results in terms of greater accuracy and reduction in time.

4. 1 Performance metrics

To measure the accuracy in classification, the Precision, Recall and F-measures values are calculated.

- TP (True Positive) represents the number of pattern correctly classified,
- FN (False Negative) refers to the number of pattern misclassified as non-patterns,
- FP (False Positive) expresses the number of non-pattern misclassified as patterns,
- TN (True Negative) is the number of non-pattern correctly classified.

Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative+False positive+ False negative}}$$

Precision

Precision (P) is the ratio of number of instances correctly classified to the total number of instances and is expressed by formula

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}}$$

Recall

Recall (R) is the ratio of the number of instances correctly classified to the total number of predicted instances and is expressed by formula

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False negative}}$$

F-measure

F-measure is the harmonic mean between precision and recall, and is defined as

$$\text{F-Measure} = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

4.2 Comparison Result

The metrics such as accuracy are used to compare the existing and proposed scenario using the performance metrics. In existing scenario, the accuracy values are lower and time complexity is high. In proposed scenario, the accuracy value is higher and time complexity is reduced significantly. The above table shows that comparison between the existing and proposed system. Accuracy, Precision, Recall, F-measure are to be increased.

Table-1: Comparison Table

Metrics	Sequence Classification	Firefly optimized Sequence Classification
Accuracy	82.5000	92
Precision	0.8250	0.9202
Recall	0.8250	0.9200
F-Measure	0.8250	0.9201

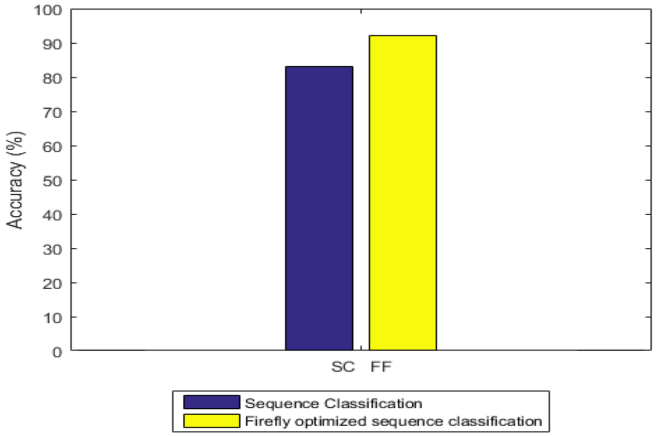

Chart-1: Accuracy

5. CONCLUSION AND FUTURE WORK

The problem of sequence classification using rules composed of interesting patterns establish in a dataset of labeled sequences and accompanying class labels. In existing work four concrete classifiers are used for various pattern types and classification strategies. Proposed work use firefly Optimization for the automatic tuning process of the parameters. Demonstrate that the pattern mining method is effective in finding informative patterns to represent the sequences, leading to classification accuracy that is in most cases more advanced than the existing work. The experimental results will prove that the proposed technique works more effectively and efficiently than the existing technique. The basic firefly algorithm is very efficient, but the solutions are still changing the optimal approach. It is possible to improve the solution quality by reducing the randomness gradually. A further improvement on the convergence of the algorithm is to vary the randomization parameter so that it reduction gradually as the optima are approaching. These form important topics for further research.

Furthermore, as a relatively straightforward expansion the Firefly Algorithm can be modified to solve multi objective optimization problems. In addition, the application of firefly algorithms in consolidation with other algorithms may form an exciting area for further research.

REFERENCES

1. Agrawal.R,Srikant.R, ”Pattern based sequence classification”,11th Int. Conf. on Data Engineering, IEEE Computer Society Press, Taiwan, pp. 3-14, 1995.
2. Deng.H, Runger.G, Tuv.E, and Bannister.W, “Cbc: An associative classifier with a small number of rules,” Decision Support Systems, vol. 59, pp. 163–170, 2014
3. Han.J, Pei.J, and Yin .Y, “Mining frequent patterns without candidate generation,” in ACM SIGMOD Record, vol. 29, no. 2. ACM, 2000.
4. Liu.B, Hsu W, and Ma.Y, “Integrating classification and association rule mining,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
5. Masseglia.F,Poncelet.P,Teisseire. M, “Efficient mining of sequential patterns with time constraints: reducing the combinations”, Expert Systems with Applications 36 (2), 2677–2690, 2009.
6. Niti Desai, AmitGanatra, “Sequential Pattern Mining Methods: A Snap Shot”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 4 (Mar. - Apr. 2013),
7. Sutou.T,Tamura.K, Mori.Y, and Kitakami.H, “Design and Implementation of Parallel Modified PrefixSpan Method”, ISHPC 2003, LNCS 2858, 2003, pp. 412–422.
8. Unil Yun, “A new framework for detecting weighted sequential patterns in large sequence databases” Knowledge-Based Systems 21 (2008) 110–1
9. Zhou.C,Cule.B, and Goethals.B, “pattern based sequence classification,” in Machine Learning and Knowledge Discovery in Databases. IEEE, 2015.