International e-Conference on New Horizons And Multidisciplinary Applications In Science And Technology
In Association withInternational Journal of Scientific Research in Science and Technology
Volume 9 | Issue 6 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

Implementation of Sequential Pattern Classifications Using SVM

Rais Allauddin Mulla[1], Mahendra Eknath Pawar[2]
[12]Assistant Professor, Vasantdada Patil pratishthan's college of engineering Sion, Mumbai, India.
rais.mulla@pvppcoe.ac.in[1],mahendraepawar@gmail.com[2]

ABSTRACT

Sequence classification has a widely utilized in applications like as genomic analysis, information retrieval, health informatics, finance, and abnormal detection. Particular from the classification undertaking on feature vectors, sequences do not have undertaking features. Indeed, even with complex feature selection methods, the dimensionality of potential features in any case be high and the sequential nature of features is difficult to capture. This makes sequence classification a very extreme undertaking than classification on feature vectors. This paper resolve the issues of sequence classification by utilizing rules made out of interesting patterns or itemsets found in a dataset of named sequences and accompanying class labels. For pattern generation we will utilize FPGrowth, and will demonstrate that it is better than Apriori and Eclat algorithm. Interesting patterns from class of sequences are produced by consolidating the cohesion and the help of the pattern. After this found patterns are changed over into classification rules which will be further classified by utilizing SVM classifier. Proposed system is tested on NEWS dataset and experimental results demonstrate that the rule based classifier (SVM) is better than existing classifier in terms of efficiency and stability.

Index Terms : Sequence classification, interesting patterns, classification rules, FP growth algorithm, SVM.

I. INTRODUCTION

Datasets in real word is gathering of texts, videos, speech signals, biological structures and web usage logs, those are consist of sequential events or elements. Because of wide range of applications, the vital issue in statistical machine learning and data mining is sequence classification. The sequence classification task is described as assigning class labels to new sequences depend on the knowledge consumed in the training stage. There are some studies exist in integrating pattern mining techniques and classification, such as classification depends on association rules, sequential pattern based sequence classifier, and many others. These composed methods can give good outcomes as well as give users with information useful for understanding the characteristic of the datasets.

Sequential pattern mining is data mining technique and is locating statistically relevant patterns from data patterns where the values are delivered in a sequence. It is a part of data mining. It presumed that the values are discrete, and thus time series mining is closely related, but usually considered a numerous activity. A special case of structured data mining is sequential pattern mining. There are an amount of key traditional

computational problems inside this field. These consist of building efficient databases and indexes for sequence information. Extracting the frequently occurring patterns and comparing the sequences for similarity. Recovering missing sequence members. Sequence mining problems can be classified as string mining. String mining is based on string processing algorithms and itemset mining, it is based on association rule learning.

SVM is machine learning technique for classification, which gives highly correct results in the passive learning scenario and SVMs learn a linear decision boundary, by using kernel-induced feature space and estimating the distance of a sample to this boundary is straightforward and provides an measurement of its in formativeness, Efficient online learning algorithms make it possible to obtain a sufficiently accurate approximation of the optimal SVM solution without retraining on the whole dataset and the SVM can weight the influence of single samples in a simple manner.

The section I explains the Introduction of proposed system. Section II presents the literature review of existing systems and Section III present proposed system implementation details Section IV presents experimental analysis, results and discussion of proposed system. Section V concludes our proposed system. While at the end list of references paper are presented.

II. LITERATURE REVIEW

Sequence classification is important task in data mining Zhou et al. [1] address the problem of sequence classification. In a dataset of labeled sequences interesting patterns are found and accompanying class labels. Authors finds the interestingness of a pattern in a given class of sequences and combining the cohesion and the support of the pattern. Author utilizes the found patterns and create confident classification rules and presenting the two different classifier. The primary classifier is improved version of the existing method of classification and totally based on association rules. The secondary ranks the rules by first measuring their specific value to the new data object. Experimental results are our rule based classifiers outperform existing comparable classifiers in terms of accuracy and stability. They test a some of pattern feature based models that use different kinds of patterns as features to represent each sequence as a feature vector. Then by using a variety of machine learning algorithms to sequence classification. Experimentally presetting the patterns they discover the sequences and prove effective for the classification task.

T. C. Silva and L. Zhao [2] proposed a system, which joins both low and high-level information classification techniques. The low-level classification actualized by any classification technique, while the high-level classification fundamental systems features (graph) built from the info information, which estimates the consistence of the test examples with the pattern development of the training data.

Computation necessities confine the algorithm from managing substantial informational collections and may restrain its application in numerous domains. Chang et al. [3] authors have address this issue by updating the algorithm for execution on exceptionally parallel Graphics Process Units (GPUs). They have explored a few ideas of GPU programming and built up a dynamic programming algorithm, which is appropriate for implementation on GPUs.

Egho et al. [4] authors identify that there are two important problems related to pattern-based sequence classification, the curse of parameter tuning and the instability of common interestingness measures. To handle

these problems, system suggests a new approach and framework for mining sequential rule patterns for classification purpose. System presents a model space. This model space is defining a Bayesian criterion for calculating the interest of sequential patterns and also develops a parameter-free algorithm to efficiently mine sequential patterns from the model space. Extensive experiments show that the new criterion identifies interesting and robust patterns, the direct utilization of the mined rules as new features in a classification process describe better performance than the state-of-the-art sequential pattern based classifiers.

Mao et al. [5] authors proposed data imbalance issue become more prominent in the applications of pattern recognition and machine learning. For a new online sequential extreme learning machine method with sequential SMOTE strategy is proposed for getting fast and efficient categorization for this particular issue. This method is utilized to reduce the randomness while generating virtual minority samples by means of the distribution characteristic of online sequential data. A baseline algorithm is utilized for using online-sequential extreme learning machine method contains two stages are generated by synthetic minority oversampling technique (SMOTE) generates each class distribution based on which some virtual samples. In online stage, each class's membership is calculated according to the projection distance of sample to principal curve. The redundant majority samples and unreasonable virtual minority samples are all excluded to lighten the imbalance level in online stage. The proposed system is estimates four UCI datasets as well as the real-world air pollutant forecasting dataset. The experimental results show that, the proposed method outperforms the classical ELM, OSELM and SMOTE-based OS-ELM in terms of generalization performance and numerical stability.

A privacy preserving association rule mining algorithm given a privacy preserving scalar product protocol [6], and an efficient protocol for computing scalar product while preserving privacy of the individual values. Author shows that it is possible to achieve good individual privacy with communication cost comparable to that required to build a centralized data warehouse.

Zhou et al. [7], proposed a sequence classification method based on interesting item sets named SCII with two variations. Authors also found the issues of sequence classification by making usage of rules composed of interesting item sets found in a dataset of labeled sequences as well as accompanying class labels.

Themis et al. [8] proposed a technique for sequence classification, which employs sequential pattern mining and optimization, in a two-stage process. The method provides high classification results in the sequence classification issue, similar or better with previously reported works.

Holat et al. [9] authors analyzed a type of patterns in sequential data, the free sequential patterns. These patterns are the shortest sequences of equivalence classes on the support with respect to the threshold.

Dafe et al. [10] authors proposed sequential pattern mining. The extension of familiar methods from many other classical patterns to sequences is not a small task. While this notion has extensively been discussed for itemsets. System defines an efficient algorithm devoted to the extraction of $^{\delta}$ free sequential patterns. In proposed system shows the advantage of $^{\delta}$ free sequences and highlight their importance when building sequence classifiers.

III. SYSTEM ARCHITECTURE

System Architecture

1. Input Dataset

The dataset News was formed by selecting the five biggest groups of documents from the 20 Newsgroups dataset. The five groups are rec.sport.hockey (999 documents), rec.motorcycles (996 documents), soc.religion.christian (996 documents), rec.sport.baseball (994 documents) and sci. crypt (991 documents).

2. Data Preprocessing:

Three common text preprocessing methods that can handle the problem of noise word and help improve the overall performances of the learning algorithms for numerous text mining tasks such as categorization are: stop words, stemming process, and pruning.

a) Stop words:
A stop word, also known as stop word, is a word that is filtered during the preprocessing of text. For example, words such as is, are, you and me can appear in any text document. The longer the document is the greater chance of encountering them.

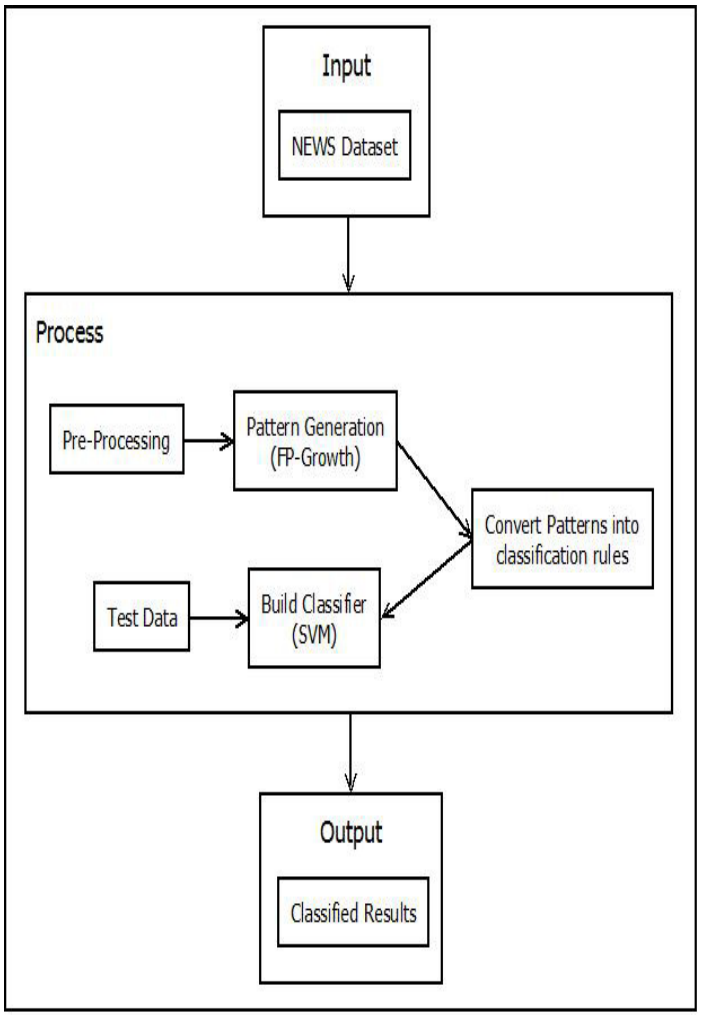

Fig 1. System Architecture

b) Stemming:

The terms process, processing and processed derive from the root word process. The question is, in a document, should we consider these words separately, or should we collapse them into a single root form. The stemming process addresses this exact issue. For example, if we apply the stemming algorithm on. A data processing program processes a document; we have A data process program process a document. The stemming algorithm refers to the above process that reduces the in affected and/or derived words to their stems. In practice, the words with the same stem root share a single summed frequency count. For stemming, the user does not have to provide the list of stems. The stemming process can be done automatically by using various methods.

c) Pruning:
Pruning, in Machine Learning, refers to an action of removing non relevant features from the feature space. In text mining, pruning is a useful preprocessing concept because most words in the text corpus are low-frequency words.

3. Pattern Generation:

After a text document is converted to a feature vector format, the pattern must be identified and analyzed to help extracting hidden information. There are various pattern mining algorithms available depending on the nature of text mining task. FP-Growth Algorithm for Generating Interesting Item-sets: The rule generator for interesting itemsets (SCII-RG) generates all interesting itemsets in two steps. Due to the fact that the cohesion and interestingness measures are not anti-monotonic, we prune the search space based on support alone. We use an FP-Growth like algorithm to find the frequent item sets. The FP-Growth Algorithm, proposed by Han, is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree). The example of FP-tree are as follow:
FP-Tree is constructed using 2 passes over the dataset:

Pass 1:
(a) Scan data and find support for each item.
(b) Discard infrequent items.
(c) Sort frequent items in decreasing order based on their support.
(d) Minimum support count = 2
(e) Scan database to find frequent 1-itemsets s(A) = 8, s(B) = 7, s(C) = 5, s(D) = 5, s(E) = 3
(f) Item order (decreasing support): A, B, C, D, E Use this order when building the FP-Tree, so common prefixes can be shared.

Pass 2:
Nodes correspond to items and have a counter
(a) FP-Growth reads 1 transaction at a time and maps it to a path

(b) Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix). In this case, counters are incremented

(c) Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines). The more paths that overlap, the higher the compression. FP-tree may fit in memory.

(d) Frequent itemsets extracted from the FP-Tree.

4. Classification Rules:

Once we have extracted all interesting patterns in each class of sequences, the next step is to identify the classification rules we will utilize to build a classifier.
We define r : P) Lk as a classification rule where P is an interesting pattern in Sk and Lk is a class label. P is the antecedent of the rule and Lk is the consequent of the rule. We further define the interestingness, support, cohesion and size of r to be equal to the interestingness, support, cohesion and size of P, respectively.

5. Build Classifier:

SVM has been proved to be an effective method for sequence classification. The basic idea of applying SVM on sequence data is to map a sequence into a feature space and find the maximum-margin hyperplane to separate two classes. Sometimes, we do not need to explicitly conduct feature selection. A kernel function corresponds to a high dimension feature space. Given two sequences x, y, some kernel functions, K(x, y), can be viewed as the similarity between two sequences. The test data is the data from which we want to classify the results.

6. Classification Result:

Finally the classification results shows the classification results, which predict the results that the news is from politics, sports or etc.

IV. RESULT AND DISCUSSIONS

V. Experimental Setup

All the experimental cases are implemented in Java in congestion with Netbeans tools and MySql as backend, algorithms and strategies, and the competing classification approach along with various feature extraction technique, and run in environment with System having configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM.

VI. Comparison Results

This section presents the performance of the SVM, Apriori, FP-Growth, Eclat algorithms in terms of time required and memory. Fig 2 Shows memory Comparison of algorithms for various Threshold. X-axis shows Algorithm & Y-axis shows memory required in bytes. Eclat require less memory as comared with other algorithms.

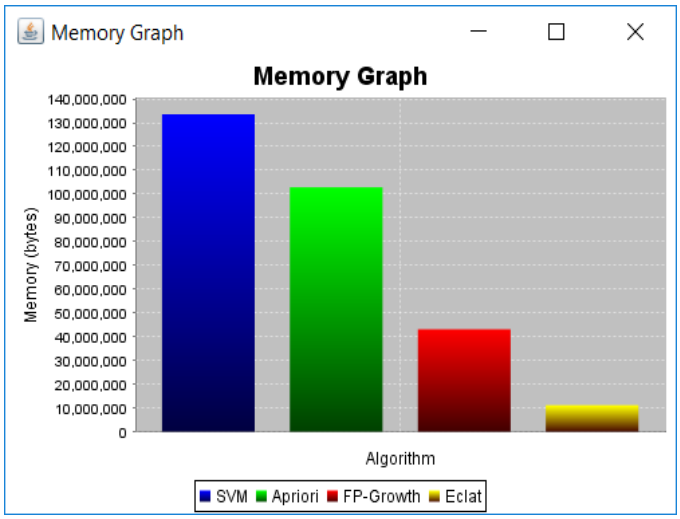

Fig. 2: Memory Comparison Graph

Fig 3 shows the Time comparison of SVM, Apriori, FP-Growth and Eclat algorithms for various size. The X-axis shows algorithms and Y- axis shows Time in ms. The Eclat takes very less time than other algorithms for classifying dataset.

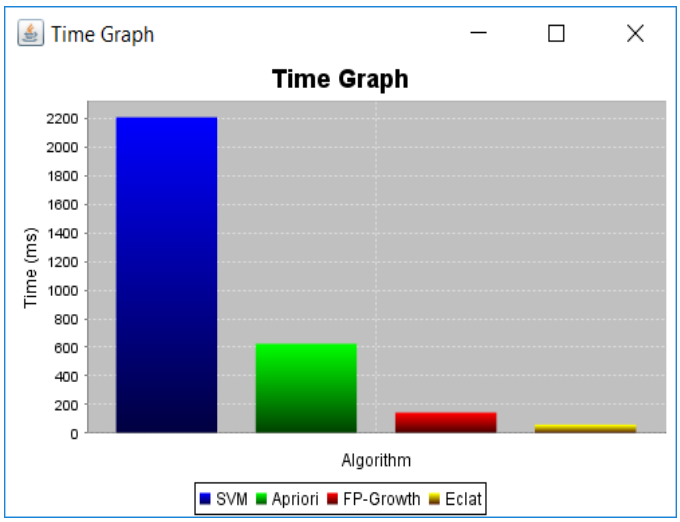

Fig. 3: Time Comparison Graph

VII. CONCLUSION

System solves the problem of sequence classification by using rules composed of interesting patterns or item sets found in a dataset of labeled sequences and accompanying class labels. This system works in two important steps: pattern generation and sequential classification. This system uses FP-Growth algorithm for pattern generation and proves that this is more time efficient than Apriori and Eclat algorithm. For sequential classification SVM classifier is used which generates more accurate classification results. The performance of system is evaluated on news Dataset and prove the time effectiveness of proposed system.

VIII. REFERENCES

[1] Cheng Zhou, Boris Cule, and Bart Goethals, "Pattern Based Sequence Classification", IEEE Transaction on knowledge and data engineering, Vol 28, No. 5, May 2016.

[2] T. C. Silva and L. Zhao, "Pattern-Based Classification via a High Level Approach Using Tourist Walks in Networks," 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, Ipojuca, 2013, pp. 284-289.

[3] K. W. Chang, B. Deka, W. M. W. Hwu and D. Roth, "Efficient Pattern-Based Time Series Classification on GPU," 2012 IEEE 12th International Conference on Data Mining, Brussels, 2012, pp. 131-140.

[4] E. Egho, D. Gay, M. Boull, N. Voisine and F. Clrot, "A Parameter-Free Approach for Mining Robust Sequential Classification Rules," 2015 IEEE International Conference
on Data Mining, Atlantic City, NJ, 2015, pp. 745-750.

[5] Wentao Mao, J.Wang and L.Wang, "Online sequential classification of imbalanced data by combining extreme learning machine and improved SMOTE algorithm," 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, 2015, pp. 1-8.

[6] G. Zhang and M. Piccardi, "Sequential labeling with structural SVM under the F1 loss," 2014 IEEE International Conference on Image Processing (ICIP), Paris, 2014, pp. 5272-5276.

[7] 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.

[8] Exarchos, Themis P., et al. "A two-stage methodology for sequence classification based on sequential pattern mining and optimization." Data and Knowledge Engineering 66.3 (2008): 467-487.

[9] P. Holat, M. Plantevit, C. Rassi, N. Tomeh, T. Charnois and B. Crmilleux, "Sequence Classification Based on Delta-Free Sequential Patterns," 2014 IEEE International Conference on Data Mining, Shenzhen, 2014, pp. 170- 179.

[10] G. Dafe, A. Veloso, M. Zaki, W. Meira, "Learning sequential classifiers from long and noisy discrete-event sequences efficiently", Data Mining Knowl. Discovery, vol. 29, no. 6, pp. 1685-1708, 2014.