ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Survey on Interesting Pattern Classification

Satyani Manthale, Prof. K. P. Moholkar

Student, Department of Computer Engineering, Rajashri Shahu College of Engineering, Savitribai Phule Pune University, Pune, India

Professor, Department of Computer Engineering, Rajashri Shahu College of Engineering, Savitribai Phule Pune University, Pune, India

ABSTRACT: Sequence classification used in several applications such as retrieval of data, genomic analysis, health informatics and so on. Unlike the process of classification on feature vectors, sequences don't have explicit features. Indeed, even with refined feature selection methods, the dimensionality of potential features may in any case be high and the sequential nature of features is difficult to catch. Because of this, sequence classification is very hard task as compared with the classification on feature vectors. This paper solves the issue of sequence classification by making use of rules generated by interesting patterns or item sets found in a labeled sequences dataset as well as having class labels. In this study, we will go through some of the present work by different researchers on sequence classification. This paper also presents the solution of pattern generation by using FP-Growth algorithm, and proves that it outperforms Apriori algorithm. Also convert these patterns into classification rules. Finally SVM classifier is used for rule classification with efficient and stable results.

KEYWORDS: Sequence Classification, Interesting Patterns, Classification Rules, Feature Vectors, Text Mining, Sequential Patterns, FP Growth, SVM.

I. Introduction

Real word datasets is collections of texts, videos, speech signals, biological structures and web usage logs; those are composed of sequential events or elements. Because of wide range of applications, the important problem in statistical machine learning and data mining is sequence classification. The sequence classification task is described as assigning class labels to new sequences based on the knowledge gained in the training stage. Classification based on association rules, sequential pattern based sequence classifier, and many others [1]. These combined methods can give good outcomes as well as provide users with information useful for understanding the characteristic of the datasets.

Sequential pattern mining is locating statistically relevant patterns among data examples where the values are delivered in a sequence [4] [10]. It is a part of data mining. It is also includes presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. A special case of structured data mining is sequential pattern mining. There are a number 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 item set mining; it is based on association rule learning [4].

SVM is a state-of-the-art method, which gives highly accurate results in the passive learning scenario and The properties of SVM :(i) SVMs learn a linear decision boundary, by using kernel-induced feature space and measuring the distance of a sample to this boundary is straightforward and provides an estimation of its in formativeness. (ii) Efficient online learning algorithms make it possible to obtain a sufficiently accurate approximation of the optimal SVM solution without retraining on the whole dataset. (iii) The SVM can weight the influence of single samples in a simple manner [6].

DOI: 10.15680/IJIRCCE.2017. 0503024

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

II. RELATED WORK

Zhou et al. [1] have developed a technique which combines both low and high-level data classification techniques. The low-level term can be implemented by any classification technique, while the high-level term is realized by means of the extraction of the underlying network's features (graph) constructed from the input data, which measures the compliance of the test instances with the pattern formation of the training data. They have made use of the dynamical features that are generated from a tourist walker in a networked environment.

K. W. Chang et al. [2] authors have addressed this issue by redesigning the algorithm for implementation on highly parallel Graphics Process Units (GPUs). Computation requirements restrict the algorithm from dealing with large data sets and may limit its application in many domains. They have investigated several concepts of GPU programming and developed a dynamic programming algorithm, which is suitable for implementation on GPUs.

M. A. Salama et al [3] have proposed model, which handles most of the requirements of data classification techniques in a single model. The developed model uses pattern based clustering concept to perform the classification of the data. At last, the developed model is compared with the six other models for performance check.

M. A. Salama et al. [4] have presented a model of a supervised machine learning approach for classification of a dataset. The model extracts a set of patterns common in a single class from the training dataset according to the rules of the pattern-based subspace clustering technique. These extracted patterns are used to classify the objects of that class in the testing dataset. The user-defined threshold dependence problem in this clustering technique has been addressed in the proposed model.

G. Chicco et al. [5] have developed two approaches for customer classification—a modified follow-the-leader algorithm and the self-organizing maps. They include an overview of basic theory for these methods and discuss the performance of the customer classification on the real case of a set of customers supplied by a distribution company.

C. Zhou et al. [6] have proposed a sequence classification method based on interesting item sets named SCII with two variations. In given paper authors also address the issues of sequence classification by making use of rules composed of interesting item sets found in a dataset of labelled sequences and accompanying class labels.

Exarchos et al. [7] have developed technique for sequence classification, which employs sequential pattern mining and optimization, in a two-stage process. The methodology provides high classification results in the sequence classification problem, comparable or better with previously reported works.

P. Holat et al. [8] have analyzed a type of patterns in sequential data, the δ-free sequential patterns. These patterns are the shortest sequences of equivalence classes on the support with respect to the δ threshold. After that, they have proposed pruning properties depending on the projected databases of the sequence.

R. Nopsuwanchai et al. [9] has developed a new approach to assess arousal states based on the analysis of eye-blink patterns. The concentration is to introduce a non-intrusive and driver-independent system that can identify the struggling state, or the state where drivers spend a large effort to overcome drowsiness. The contributions of this paper are three folds.

L. T. Nguyen et al. [10] have implemented a lattice-based approach for mining class-association rules, and two algorithms for efficient mining CARs and PCARs were presented, respectively. The developed method has more advantages than the heuristic and greedy methods in that the former could easily remove noise, and the accuracy is thus higher. It can additionally generate a rule set that is more complete than C4.5 and ILA.

.

Table 1: Survey Table

Sr. No	Title	Method Used	Advantages	Disadvantages
1.	Pattern-based Classification via a High Level Approach using Tourist Walks in Networks	Combination of low and high level data classification techniques	successfully capture topological features of the underlying network in a local to global basis	high level classifier cannot extended by considering new network measures
2.	Efficient Pattern-Based Time Series Classification on GPU	shapelet discovery algorithm	performance improvements	Slightly high computing power
3.	Pattern-based subspace classification model	pattern based clustering	efficient	missing data imputation is not handled
4.	Uni-class pattern-based classification model	Thrombosis disease Classification	approach is more efficient and effective than the existing techniques	rule extraction is not considered
5.	Load pattern-based classification of electricity customers	—a modified follow-the-leader algorithm and the self-organizing maps	gives automatic clustering without modifying the search space	Not assign dedicated tariff rates to each customer class in such a way to maximize the profits of the electricity providers
6.	Itemset based sequence classification	discovered itemsets	provide higher classification accuracy compared to existing methods	Incomplete data is not handled
7.	A two-stage methodology for sequence classification based on sequential pattern mining and optimization	sequence classification model and an optimization technique	s high classification results in the sequence classification problem	Methodology is not extended in order to handle time series, through the use of discretization techniques
8.	Sequence Classification Based on Delta-Free Sequential Patterns	extraction of δ-free sequential patterns	Address the feature selection problem in statistical classifiers, as well as to build symbolic classifiers which optimizes both accuracy and earliness of predictions.	δ-free sequential patterns in natural language processing problems
9.	Driver-Independent Assessment of Arousal States from Video Sequences Based on the Classification of Eyeblink Patterns	Hidden Markov Models (HMMs) to classify eyeblink patterns	classify eyeblink patterns from the video of the drivers, and the arousal states are estimated from the histogram variations of these typical blink patterns	The results of arousal state assessment are not used.
10.	Classification based on association rules: A lattice-based approach	Classification Based on Associations (CBA)	more advantages than the heuristic and greedy methods in that the former could easily remove noise, and the accuracy is thus higher	Did not apply these measures in CARs/PCARs and discuss the impact of these interestingness measures with regard to the accuracy of the classifiers built

III. PROPOSED SYSTEM

The proposed system solve the problem of sequence classificationusing rules composed of interesting patterns or itemsets found in a dataset of labeled sequences and accompanyingclass labels. For pattern generation we will use FPGrowthAlgorithm, and will prove that it is better than Apriorialgorithm. Interesting patterns from class of sequences aregenerated by combining the cohesion and the support of thepattern. Discovered patterns are converted into classificationrules which will be further classified by using SVM classifier.Proposed system is tested on NEWS dataset and experimentalresults prove that the rule based classifier (SVM) is better thanexisting classifier in terms of accuracy and stability.

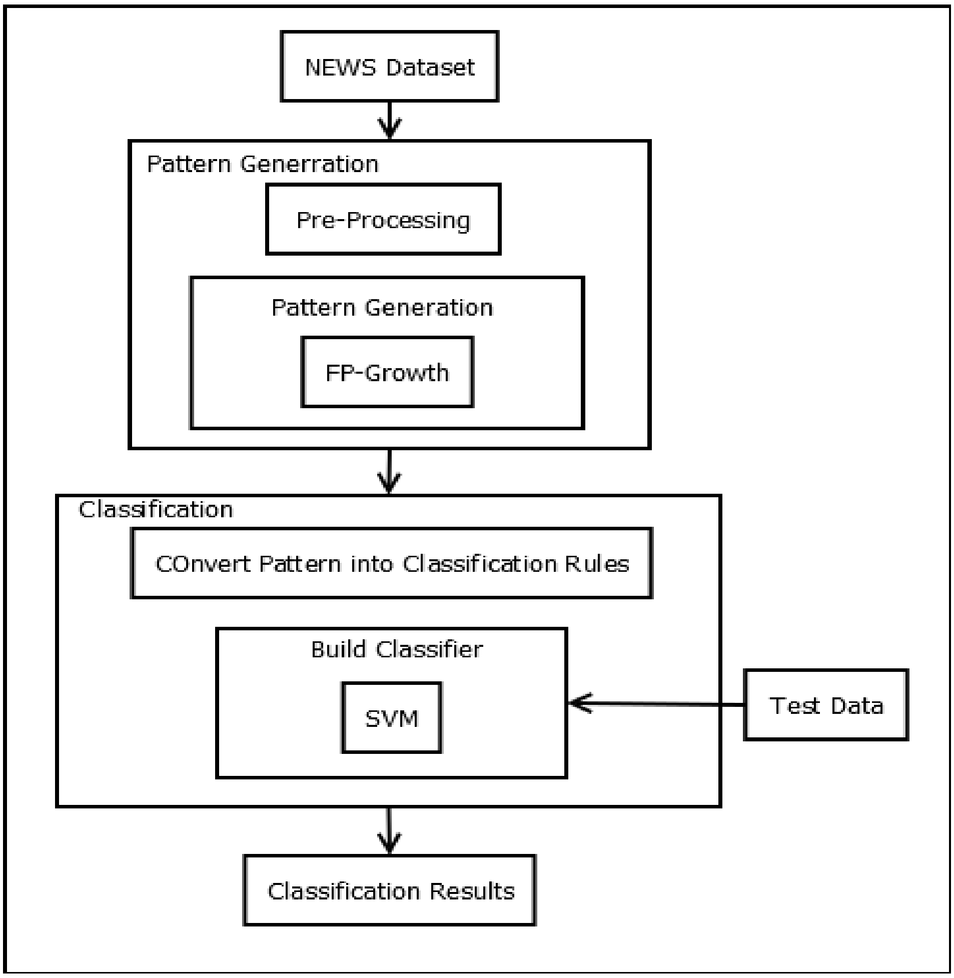

Fig 1. Propose System

System works as per following modules:

1. Input Dataset

The News dataset is formed with 20 Newsgroups dataset. The five groups are rec.sport.hockey, rec.motorcycles, soc.religion.christian, rec.sport.baseball and sci. crypt.

2. Data Preprocessing

To remove the noisy and unwanted data three data preprocessing methods are used named as: stop words, stemming process, and pruning.

3. Pattern Generation

A text document is converted into a feature vector format. In this feature vector format, the pattern must be identified and analyzed to extract hidden information. For pattern generation FP-Growth algorithm is used

IV. CONCLUSION

This paper provides a brief survey on sequence classification. Sequential classification achieves more accuracy and stability. Paper describes details description of some recent technologies along with comparative analysis. We compare recent sequence classification methods on the basis of methodology used, and respective advantages and disadvantages. By providing solution to the previous limitations, we also prove that FP growth outperforms Appriori for pattern generation. And for classification task, SVM provides more accurate and stable solution among all classification algorithms.

REFERENCES

[1] 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.

[2] 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.

[3] M. A. Salama, A. E. Hassanien and A. A. Fahmy, "Pattern-based subspace classification model," Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on, Fukuoka, 2010, pp. 357-362.

[4] M. A. Salama, A. E. Hassanien and A. A. Fahmy, "Uni-class pattern-based classification model," 2010 10th International Conference on Intelligent Systems Design and Applications, Cairo, 2010, pp. 1293-1297.

[5] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu and C. Toader, "Load pattern-based classification of electricity customers," in IEEE Transactions on Power Systems, vol. 19, no. 2, pp. 1232-1239, May 2004.

[6] 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.

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

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

[9] R. Nopsuwanchai, Y. Noguchi, M. Ohsuga, Y. Kamakura and Y. Inoue, "Driver-Independent Assessment of Arousal States from Video Sequences Based on the Classification of Eyeblink Patterns," 2008 11th International IEEE Conference on Intelligent Transportation Systems, Beijing, 2008, pp. 917-924.

[10] 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.