Survey on Frequent Pattern Mining

Vijay Sitaram Jadhav **Ritesh Kumar Yadav** **Dr. Varsha Namdeo**
M. Tech [CSE]-Scholar **Asst. Prof CSE Deptt** **Head of CSE Deptt**
Department of Computer Science & Engineering
RKDF Institute of Science and Technology, Bhopal

ABSRACT
Mining Huge data is a problem of today's great practical importance. However, there are some challenges for mining these huge data which includes
(1) The curse of dimensionality.
(2) Mining meaningful information based on the similarity measures.
In this paper, we consider the technique for analyzing data, e.g., frequent pattern mining. Frequent pattern discovery finds frequently occurring events in large databases. Such data mining technique can be useful in various domains. For instance, in recommendation and e-commerce systems frequently occurring product purchase combinations are essential in user preference modeling. In the ecological domain, patterns of frequently occurring groups of species can be used to reveal insight into species interaction dynamics. Most frequent pattern mining research has concentrated on efficiency (speed) of mining algorithms. However, it has been argued within the community that while efficiency of the mining task is no longer a bottleneck, there is still an urgent need for methods that derive compact, yet high quality results with good application properties. Many organizational areas has been dedicated to research ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications.
Keywords: Frequent pattern, Association rule mining, Classification.

I. INTRODUCTION

1. DATA MINING

Data mining is a very active and rapidly growing research area in the field of computer science. The task of data mining is to extract useful knowledge for human users from a database. Whereas the application of evolutionary computation to data mining is not always easy due to its heavy computation load especially in the case of a large database. Since Agrawal, the frequent pattern mining problem has been studied extensively with alternative problem formulations, as well as new variants of existing algorithms together with new application settings such as telecommunications, bioinformatics, web mining, text mining, and many more. In retrospect, however, the efficiency (run time) of enumerating the complete set of frequent patterns has attracted most research in the subject.

2. FREQUENT PATTERN

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread, which appear frequently together in a transaction data set, is a *frequent itemset*. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (*frequent*) *sequential pattern*. A *substructure* can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently in a graph database, it is called a (*frequent*) *structural pattern*. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data indexing, classification, clustering, and other data mining tasks as well. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research.

Frequent pattern mining was first proposed by Agrawal et al. (1993) for market basket analysis in the form of association rule mining. It analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets". For instance, if customers are buying milk, how likely are they going to also buy cereal (and what kind of cereal) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and arrange their shelf space.

Since the proposal of this new data mining task and its associated efficient mining algorithms, there have been hundreds of follow-up research publications, on various kinds of extensions and applications, ranging from scalable data mining methodologies, to handling a wide diversity of data types, various extended mining tasks, and a variety of new applications.

II. Literature Review

A frequent itemset (pattern) [11]is a set of items that occur in a dataset no less than a user-specified minimum support (*min sup*). Frequent patterns have been explored widely in classification tasks. The frequent pattern-based classification is related to associative classification. In associative classification, a classifier is built based on high-confidence, high support association rules [9]. The association between frequent patterns and class labels is used for prediction.

A recent work on top-*k* rule mining [6] discovers top-*k* covering rule groups for each row of gene expression profiles. Prediction is then performed based on a classification score which combines the support and confidence measures of the rules. Earlier studies on associative classification [6], mainly focus on mining high-support, high-confidence rules and build a rule-based classifier. Prediction is made based on the top-ranked rule or multiple rules. In related work we are going to focus on following methods:

1. HARMONY
2. PatClass
3. DDPMINE

1. HARMONY:

HARMONY is another rule-based classifier which directly mines classification rules. It uses an instance-centric rule-generation approach and assures for each training instance, that one of the highest confidence rules covering the instance is included in the rule set. HARMONY is shown to be more efficient and scalable than previous rule-based classifiers. On several datasets that were tested by both our method and HARMONY, the classification accuracy is significantly higher, e.g., the improvement is up to 11.94% on Waveform and 3.40% on Letter Recognition.

2. PATCLASS:

PatClass is the least efficient method. The performance of PatClass is very sensitive to *min sup*: as *min sup* lowers down, the running time increases dramatically, due to an explosive set of frequent itemsets produced. Besides the mining process, feature selection also slows down since the set of frequent itemsets as input is bulky.

3. DDPMINE:

DDPMine which directly mines the discriminative patterns and integrates feature selection into the mining framework. A branch-and-bound search is imposed on the FP-growth mining process, which prunes the search space in the database. DDPMine works in an iterative fashion and reduces the problem size incrementally by eliminating training instances which are covered by the selected features. Experimental results show that DDPMine achieves orders of magnitude speedup over the two-step method without any downgrade of classification accuracy. DDPMine also outperforms the state-of-the-art associative classification methods in terms of both accuracy and efficiency.

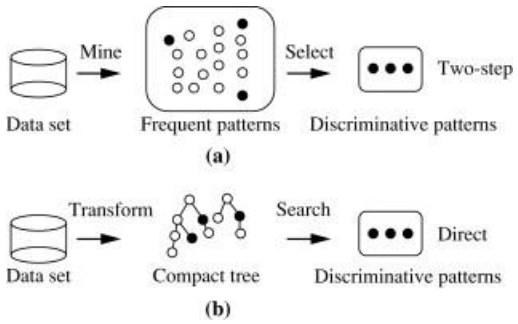

Fig. Two step Vs DDPMINE approach

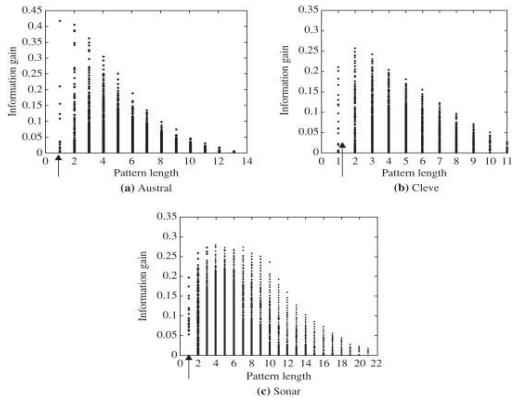

Fig. Single feature versus frequent pattern

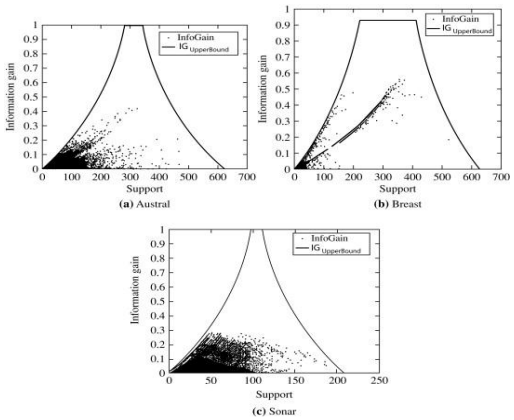

Fig. Information gain versus pattern

Datasets	Harmony	PatClass	DDPMine
adult	81.90	84.24	84.82
chess	43.00	91.68	91.85
crx	82.46	85.06	84.93
hypo	95.24	99.24	99.24
mushroom	99.94	99.97	100.00
sick	93.88	97.49	98.36
sonar	77.44	90.86	88.74
waveform	87.28	91.22	91.83
Average	82.643	92.470	92.471

RUNTIME COMPARISION

A classifier RCBT is constructed from the top-*k* covering rule groups and achieves very high accuracy. HARMONY [4] is another rule-based classifier which directly mines classification rules. It uses an instance-centric rule-generation approach and assures for each training instance that one of the highest-confidence rules covering the instance is included in the rule set. HARMONY is shown to be more efficient and scalable than previous rule-based classifiers is a innovative association rule-based classification method. Different from all the above studies, it is based on a *lazy* (non-eager) classification philosophy, in which the computation is performed on a demand-driven basis. This lazy classification method effectively reduces the number of rules produced by focusing on the test instance only. The concept of least frequent item sets in association [9] rule discovery is discussed. Apart from the traditional mining procedure for frequent pattern mining, the numerical frequent pattern approach is discussed. A combined approach which integrates association rule mining and classification rule mining called associative classification using medical data set is discussed.

Experimental studies show that the lazy approach outperforms both the eager associative classification approach and decision tree-based classifiers in terms of accuracy. Based on the experimental results on common datasets, is shown to achieve high accuracy than HARMONNY. Discriminative[12] is a frequent pattern-based[5] classification method. Highly discriminative frequent itemsets are selected to represent the data in a feature space, based on which any learning algorithm can be used for model learning. This method first mines a set of frequent itemsets[7], then performs feature selection on the mining results to single out a compact set of highly discriminative itemsets. This method is shown to achieve very high accuracy.

IV. CONCLUSIONS

Frequent pattern-based classification methods have shown to be very effective at classifying categorical or high dimensional sparse datasets. However, many existing methods which mine a set of frequent itemsets encounter nontrivial computational bottleneck in the mining step, due to the explosive combination between the items. In addition, the explosive number of features poses great computational challenges for feature selection.

In this study, we proposed an Associative classification based on Support Vector Machine (SVM)[8] mining approach for frequent pattern mining which directly mines the patterns and integrates feature selection into the mining framework. A branch-and-bound search is imposed on the FP-growth mining process, which prunes the search space significantly. It works in an iterative fashion and reduces the problem size incrementally by eliminating training instances which are covered by the selected features. This achieves orders of magnitude speedup over the two-step method without any downgrade of classification accuracy. SVM[8]-based discretization can reduce the number of generated rules greatly while improving classification accuracies. SVM[8] based Pattern mining also outperforms the state-of-the-art associative classification methods in terms of both accuracy and efficiency.

V. REFERENCES

[1] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proc. of KDD*, 1998, pp. 80–86.

[2] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules," in *Proc. of ICDM*, 2001, pp. 369–376.

[3] X. Yin and J. Han, "CPAR: Classification based on predictive association rules," in *Proc. of DM*, 2003, pp. 331–335.

[4] J. Wang and G. Karypis, "HARMONY: Efficiently mining the best rules for classification,"in *Proc. of SDM*, 2005, pp. 205–216.

[5] H. Cheng, X. Yan, J. Han, and C. Hsu, "Discriminative frequent pattern analysis for effective classification," in *Proc. of ICDE*, 2007, pp. 716– 725

[6] A. Veloso, W. M. Jr., and M. Zaki, "Lazy associative classification," in *Proc. of ICDM*, 2006, pp. 645–654.

[7] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate Generation," in *Proc. of SIGMOD*, 2000, pp. 1–12.

[8] Cheong Hee Park , Moonhwi Lee, "Associative Classification Using SVM-based Discretization," 2007 International Conference on Computational Intelligence and Security.

[9] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, In Proceedings of the

20th International Conference on Very Large Data Bases, 1994, pp. 487–499.
[10] B. Goethals. Survey on frequent pattern mining. Technical report, Helsinki Institute for Information Technology,03.
[11] Sunil Joshi et al: accepted research paper in The IEEE 2010 International Conference on Communication software and Netweorks (ICCSN 2010) *on "A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database" from* 26 - 28 February 2010.
[12] Hong Cheng , Xifeng Yan , Jiawei Han , Philip S. Yu, "Direct Discriminative Pattern Mining for Effective Classification"