Improvement in Performance of Classificstion Using Lift Method

Neha P. Sonar
Department of Computer Engineering
SSBT's COET Bambhori
Jalgaon, India

Ankit K. Dixit
Department of Computer Engineering
SSBT's COET Bambhori
Jalgaon, India

Kajal E. Shelke
Department of Computer Engineering
SSBT's COET Bambhori
Jalgaon, India

Viren A. Patil
Department of Computer Engineering
SSBT's COET Bambhori
Jalgaon, India

Rahul N. Patil
Department of Computer Engineering
SSBT's COET Bambhori
Jalgaon, India

***Abstract*—Data mining is the process of analyzing data from different perspective and summarizing it into useful information. Classification is one of the important task in data mining. Interestingness of pattern in a given class is measured by the two terms- support and cohesion. The set of classification rules are generated from these interesting pattern. In existing system these set of classification rules are used to classify the dataset. But the set of rules may contain independent rules or rules which are not correlate to each other. Such independent rules increases classification time. So, there is need to prune such independent rules, for which lift technique is used. The lift technique will reduce the amount of time require for classification of data.**

Keywords— Interesting pattern, Classification rules, lift Method, Classification

I. INTRODUCTION

Data mining is the process of identifying the valid information from huge dataset. There are various tasks in data mining such as classification, clustering, prediction, time series analysis, pattern mining, etc. Data mining is helpful in different domains such as market analysis, decision support, fraud detection, business management. Pattern Mining is a popular technique which consists of finding subsequences or item set appearing frequently in a set of sequence. Classification has been an important problem in statistical machine learning and data mining. It also finds out frequent item sets as patterns from a datasets and used in various domains such as medical treatments, natural disasters, customer shopping sequences, DNA sequences and gene structures. In real world, as massive amount of data are needed to be collected continuously and stored in the databases. Many industries are becoming interested in mining patterns from these databases. There are large number of possible patterns in a huge dataset. The patterns that occur in particular individual items can be found and also the patterns between different items can be found. The number pattern available in datasets, and also the users have different interests and requirements. In proposed, pattern mining and classification is used for market basket analysis. There exist a number of techniques for integrating pattern and classification, such as classification based on association rules (CBA), sequential pattern based sequence classifier, the Classify-By-Sequence (CBS) algorithm. In proposed system, Classification based on Interesting Patterns. First of all, present algorithms such as support and cohesion to mine interesting patterns — item sets. As a second step, discovered patterns convert into classification rules, and then classification is done. The rest of this paper is organized as follows. In the next section, related work is described with classification technique. Section III describes Implementation of the proposed system. Next Result and Discussion is given in section IV. In V section conclusion and Future work is described.

II. RELATED WORK

In the research it is important to understand which methods are more appropriate than others so as to implement a faster market basket classification system. Interesting pattern find using support and cohesion methods. Lift technique is used to prune the classification rules.

A. Interesting Pattern

The pattern interestingness is depends on two methods: support and cohesion. The support method count of a pattern is defined as the number of sequences in which the pattern occurs. Cohesion method find how close pattern appear to each other.

a) Support: Support is an indication of how frequently the item-set appears in the dataset. Suppose item set α, Denote the set of sequence that contain all item if α as T (α). Denote the set of sequence that contain all item of α labelled by class label $C_{K.}$ The support of a pattern P in a given class of sequences Q_K can now be defined as.

$$\pi_K(P) = \frac{|T_K(P)|}{|Q_K|}$$

Where P is item set

b) Cohesion: Cohesion refers to the degree to which the elements of an item set belong together. Cohesion measures how 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. An item set α in a sequence β € T (α) as W (α, β) = min {t_2 - t_1 +1, where $t_1 \leq t \leq t_2$}. In order to compute the cohesion of a pattern P within class k, we now compute the average length of such shortest intervals in.

$$T_K(P) = W_K(P) = \frac{\sum_{\beta \epsilon T_K(P)} W(P, \beta)}{|T_K(P)|}$$

P is item set

Here, $\overline{W_K}$ (P) is greater than or equal to the number of item in P, denote as | P |. Furthermore, for a fully cohesive pattern, $\overline{W_K}$(P) =| P|. Therefore, define cohesion of P in T_K(P) as,

$$\bar{A}(P) = \frac{|P|}{\overline{W_K}(P)}$$

All Pattern containing just one item are fully cohesion, that is,

$$\bar{A}_K(P) = 1 \quad \text{If } |P| = 1$$

The cohesion of P in a single sequence γ is defined as,

$$\bar{A}(P, \gamma) = \frac{|P|}{W(P, \gamma)}$$

c) Interestingness: In a class of sequence $Q_{K,}$ now a define the interestingness of a pattern P as.

$$I_K(P) = \pi_K(P) * \bar{A}_K(P)$$

Where P is α

Minimum support threshold min_sup and a Minimum interestingness threshold min_int, a pattern P is considered interesting in a set of sequences labelled by class label C_K, if

$$\pi_K(P) \geq \min\_sup \quad \text{And} \quad I_K(P) \geq \min\_int$$

B. Lift Method

There are various technique used for finding classification rules but the problem is that there are redundant rules, which occupy a lot of space and take up CPU time for processing. To avoid such a scenario, pruning of rules is to be done. This pruning of rules is done is using "Lift" method. Lift interestingness measure defines the number of transaction that contain the item used to find interesting patterns. The Lift is denoted by Lift(X = Y) as follows.

Lift(X => Y) = sup (XUY) / sup(X) * sup(Y)

The lift of a rule is defined as, the ratio of the observed support to that expected if X and Y were independent. If some rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events. If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

Algorithm 1: Lift Method

Input => Interesting Pattern, set of values calculated by lift. For each rule R in IP.

Output => A set of prune rules PR.

Step 1: Choose co-relation of association rule based on Lift'

a. If (L>1), positive correlation α (1+α)

b. If (L<1), negative correlation β (1-β)

c. If (L=1), Antecedent and consequent are independent of each

Step 2: Prune the rules having value 1.

Step 3: Store the values of α and β for each rule.

Step 4: Scan the data set.

Step 5: Classify the rule on the basis of the values of α and β.

Step 6: Store the rules on the basis of priority

III. Implementation

The overview of proposed system is shown in the Figure 1. In this system, grocery dataset use for classification. Firstly for finding interesting patterns (IP), apply two methods namely support and cohesion. The product of Support and Cohesion will provide us with Interestingness measure over a class of sequential patterns. i.e.

IP= support * cohesion

Once discovered all interesting patterns, the next classification of data. Define, Γ = Þ => β as a classification rule where Þ is an interesting pattern and β is a class label. Þ is the antecedent of the rule and β is the consequent of the rule. The redundant rules occupy an extra space and take more time for classification. After identify all classification rules, pruning of rules is to be done for redundant rules. This pruning of rules is done is using "Lift" Method. The lift of a rule is defined as, the ratio of the observed support to that expected if the antecedent and consequent were independent. If some rule had a lift of less than 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events. If the lift is greater than 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in data sets. In the next step SCIP_MA (Sequence classification based on interesting pattern-Matching cohesion rules based classifier) algorithm use for build the classifier [1]. In system, as a result the grocery dataset is labeled into four categories: Dairy Product, Bakery Product, Fruits and Vegetables. All products are classified in these type based on classification rules.

Algorithm 2: SCIP_MA Classifier

Input: PR, default_r, a new unclassified data object d = (s, $L_?$)

```
Output: a class label
Step 1: N=0
Step 2: foreach rule r in PR do
    a)  r matches d then store r into NP
Step 3: if N >0 then
Step 4: foreach rule r: P= $C_K$ in N do
Step 5:  use SCIP_MA then
Step 6: r.value = r.lift * C (P, d.s)
Step 7: sort rules in N by descending r.value
Step 8: CR = {the top rules in sorted N}
Step 9: Score = a new array of size |L|
Step 10: foreach rule r: P= $C_K$ in CR do
Step 11: score[K] = score[K] + r.value
Step 12: return the class label $C_K$ with largest score [K]
Step 13: else return the class label of default r
```

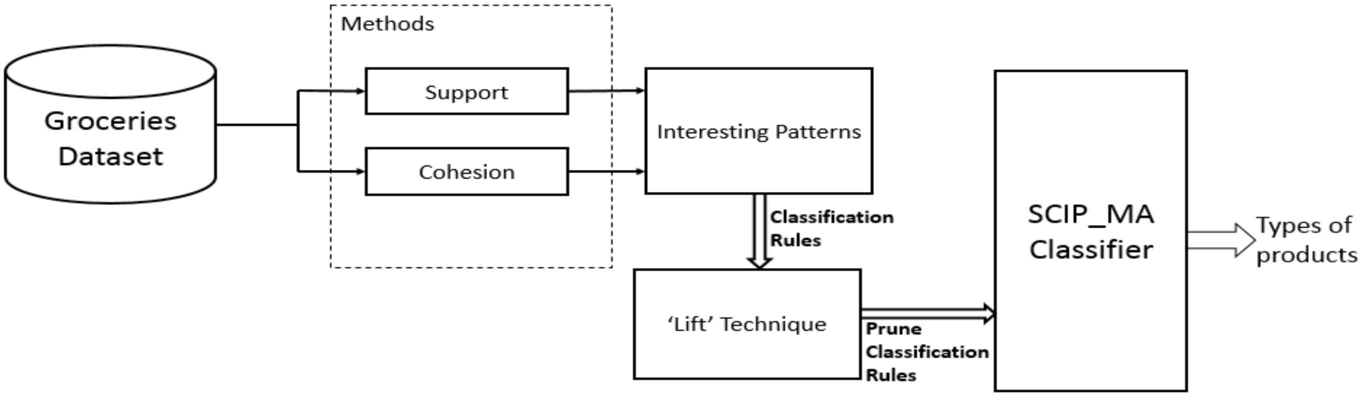

Fig. 1. Flow of Proposed System.

IV. Result and Discussion

From result analysis it is observed that the proposed system is more versatile than existing technique. Consider the example of grocery (Market Basket Analysis) dataset: {fruits, banana, orange, apple, grape, pineapple, and mango, peach} here, fruit is a class label and banana, orange etc. are all items in sequence. First finding interestingness of itemset using support and cohesion terms. Consider the itemset X= {banana, orange} then the support value of X is 0.120690 and 2 is cohesion value. Product of support and cohesion provide an interestingness value of given itemset i.e. interestingness value of above itemset is 0.241379. In the next step find the classification rule from interestingness. To prune unrelated classification rules applied the lift method. The lift value for the given itemset is 2.197802. Lift value of X is greater than 1 hence make rule potentially useful for predicting the consequent in data sets. Depend on lift value the classification rules are prune and only related classification rules are used for data classification. Due to independent rules system take more time for classification while using lift, classification time is reduced. Hence classification using lift method improve the proposed system performance.

Conclusion

In the proposed work, a classification is done by using Sequence Classification based on Interesting Pattern–Matching cohesion rule based classifier (SCIP_MA). For classification pruned rules are used. The lift method is used to prune the unrelated classification rule. It is observed that lift decreases the time required for classification of data as compared to existing system. The proposed work is useful in various data mining application such as Market Basket Analysis.

References

[1] Boris Cule, Cheng Zhou and Bart Goethals. "Pattern based sequence classification". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2016.

[2] L. T. Nguyen, B. Vo, T.-P. Hong, and H. C. Thanh, "Classification based on association rules: A lattice-based approach," Expert Systems with Applications, vol. 39, no. 13, 2012.

[3] R. Agrawal and R. Srikant, "Mining sequential patterns," in Proceedings of the 11th International Conference on Data Engineering,1995.

[4] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification" ACM SIGKDD Explorations Newsletter, vol. 12, no. 1, 2010.

[5] C. Zhou, B. Cule, and B. Goethals, "Itemset based sequence classification" in Machine Learning and Knowledge Discovery in Databases. Springer, 2013, pp. 353–368.

[6] D. Fradkin and F. Morchen, "Mining sequential patterns for classification" Knowledge and Information Systems, pp. 1–19, 2015.