Data Classification Using Sequential Pattern Mining

1

Under the Guidance of:-
Mrs. Dhanshree Tayade

[1] SSBT's College of Engineering And Technology, Bambhori Jalgaon - 425001, Maharashtra, India

Outline of Topics

Outline

Introduction

Sequence classification system classifies dataset into sequential pattern. User inputs dataset which is to be classified. The classification involves the measure of interesting pattern in a class of sequence. This involves use of techniques such as support and cohesion. The interesting patterns are then used to build a sequence classifier.

Introduction

The pattern can both,itemset and subsequences. The support measures in how many sequence the pattern appears. Cohesion measures how close the items are to each on an average. The classifier is built using the classification rules. There may be additional rules which occupy a lot of space. These rules also take time for processing. There is a need to prune such rules.

This pruning is done by using a technique known as Lift. Lift is the ratio of observed support to the expected , if x and y were independent . A rule that has independent events are pointless to process. Such rules are eliminated using Lift. The pruned subset classifies the interesting patterns of the dataset into a sequential pattern.

Outline

Literature Survey

Existing system classifies data using interesting patterns. The patterns are connects into rules to further refine the rules. The system uses confidence Method. The confidence value of a rule X = Y with respect to a set of transactions T, is the proportion of the transaction that contains X as well as Y.

Literature Survey

The propose system classifies data using interesting patterns. Later, rules are drawn from these interesting patterns. By using confidence, we may find a rule to be true for a particular instances, but the same rule wont be true for an instances. Hence, Lift method is used in the proposes system. 'Lift' prunes only those rules in which posses dependency are stored in the databases. These rules can be uses for classification the antecedent and consequent are independent of each other.The rules which possess dependency are stored in the database.There rules can be uses for classification depending on the different instances.

System Architecture

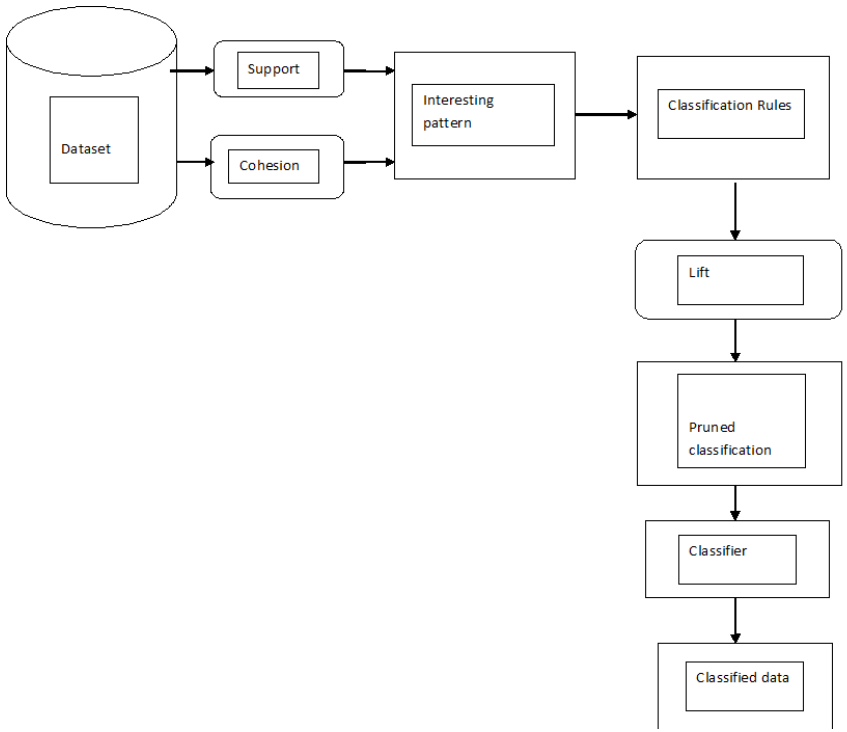


Figure: System Architecture

Outline

Interesting Pattern

- Interesting Patterns in sequences,a pattern is typically evaluated based on how often it occurs(support) along with the proximity of the items(cohesion).
- Utilise interesting patterns to build classifiers.
- Use cohesion and support methods to define interesting patterns in a sequence dataset.

$$I(P) = f(P) * c(P) \tag{1}$$

Outline

Support

- Support is an indication of how frequently the item-set appears in the database.
- 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.

$$F(P) = N(P)/S \tag{2}$$

where N=set of sequences containing all items of X.

Cohesion

- 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.
- The cohesion of P in a single sequence s is defined as

$$C(P, s) = |P|/W(P, s) \tag{3}$$

'Lift' Method

- Lift interestingness measure defines the number of transaction that contain the item used to find interesting patterns.
- The Lift measure is denote by Lift(X = Y)as shown in

$$Lift(X -> Y) = supp(X + Y)/supp(X) * supp(Y) \quad (4)$$

Example

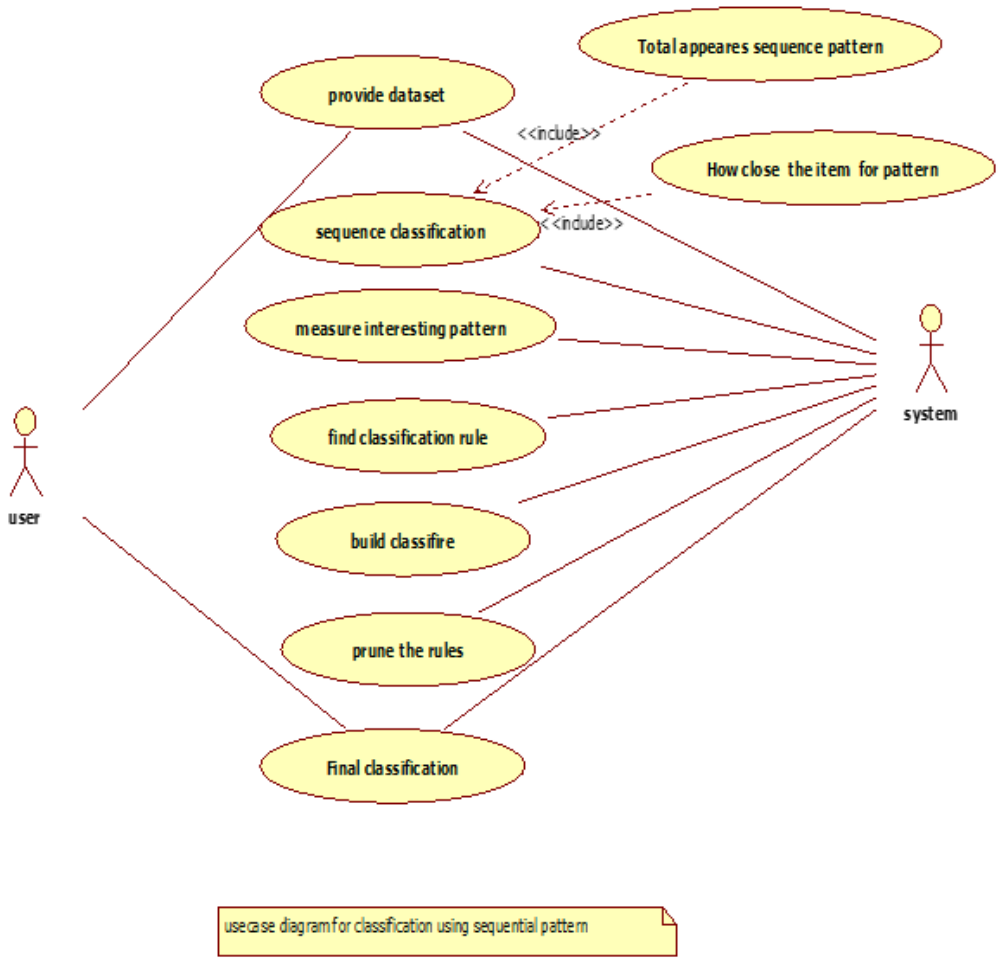
Example database with 5 transactions and 5 items

ID	Milk	Bread	Butter	Books	Pens
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

- Support=F(P)=N(P)/S
  F(P)=1/5
  F(P)=0.2
- Lift(X$\Longrightarrow Y) = supp(X \bigcup Y)/supp(X) * supp(Y)$
  $Lift(Milk, Bread \Longrightarrow Butter) = 0.2/0.4 * 0.4 = 1.25$

Outline

UseCase Diagram



Outline

Conclusion

The system classifies data using rules which are drawn from Interesting Pattern. These rules are pruned using Lift instead of Confidence. The proposed algorithm depends mainly on support compared to the algorithm proposed in the paper which uses both support and confidence. This results in simplification of the process as well as the rules that possesses dependency are not pruned. Thus, making the system more versatile.