MCFSP: Mining Closed frequent Similar Patterns

Mr. Ramesh N Koppar[1], Dr. Ramesh D[2]

[1]*Assistant Professor, Information science and Engineering, SKSVMACET, Lakshmeshwar, Karnataka, India*

[2]*Professor, Computer Science and Engineering, SSIT, Tumkur, Karnataka, India*

Abstract

The enormous amount of useful information in the form of frequent pattern poses the key challenge of how to reduce the number of frequent patterns without information loss. Despite the quality of solution given by frequent pattern mining algorithm, scalability and efficiency issues persist. Current closed frequent item set mining algorithms gives solutions only to frequent itemset by discovering reduced sets of frequent itemsets from which entire frequent itemsets can be recovered . But the solution in the case of frequent similar pattern mining wherein the number of pattern is even more than for frequent itemset mining is not present. In this paper as a solution we are extending closed frequent pattern mining technique to closed frequent similar pattern mining for discovering reduced set of frequent similar pattern without any information loss. We are proposing a novel closed frequent similar pattern mining algorithm, named MCFSP Mining closed frequent similar pattern algorithm. By traversing a tree which contains all closed frequent similar pattern the algorithm discovers frequent patterns. The MCFSP resolves scalability and efficiency issues by finding closed similar patterns, yielding a reduced size of the discovered frequent similar pattern set without information loss.

Keyword : *- Data mining, frequent closed itemset mining, frequent closed itemsets, Association rules*

---

1. Introduction

Frequent pattern mining is a technique that consists of finding patterns (i.e., feature sets with their corresponding values) that frequently occur (more than or equal to a minimum frequency threshold) in a dataset. It is considered a key task in data mining because of its application to discover useful information, such as risk factors , user's profiles, human behavior, malicious software among others. In addition, Frequent pattern mining can be used as a previous or internal step for other data mining tasks, like association rule, classification, and clustering.

Since 1990, most of the frequent pattern mining algorithms were based on the exact matching of boolean features to compare and count patterns. This subclass of frequent pattern mining algorithms was called frequent itemset mining (considered as the traditional approach for frequent pattern mining). However, real life objects, such as objects in sociology, geology, medicine or information retrieval are rarely equal or they can be described by non boolean features. Thus, similarity functions different from the exact matching were proposed to compare object descriptions giving rise to a new approach named frequent similar pattern mining which can handle datasets containing non boolean features by using similarity functions (Danger, Ruiz-Shulcloper, & Llavori, 2004; Rodríguez-González, Martínez-Trinidad, Carrasco-Ochoa, & Ruiz-Shulcloper, 2008; 2011; 2013). This approach produces patterns which can not be found by those algorithms based on exact matching. The frequent patterns found using a similarity function are named frequent similar patterns (Rodríguez-González, Martínez-Trinidad, Carrasco-Ochoa, & Ruiz-Shulcloper, 2013).Despite the quality of solutions given by a frequent itemset mining algorithm or a frequent similar pattern algorithm, a critical drawback to both these approaches is that, although a complete set of frequent itemsets or frequent similar patterns can be found, big. It is helpful, therefore, to obtain a reduced set of all the frequent patterns without

information loss (i.e., from which the entire frequent pattern set can be recovered). One way to do that, is through the use of closed frequent itemsets mining. Closed frequent itemsets mining algorithms define that a frequent itemset is closed if it has no super-patterns with the same frequency, and use this definition to find the closed frequent itemsets. From such closed itemsets, the complete set of frequent itemsets can be generated without information loss. The so-called closed frequent itemsets mining algorithms also have more efficient runtimes than frequent itemset mining algorithms (Pei et al., 2000; Uno, Asai, Uchida, & Arimura, 2003; Zaki & Hsiao, 2002).

However, the concept of a closed patterns has not been exploited for the frequent similar patterns to the best of our knowledge. In this paper, we introduce the concept of a closed frequent similar pattern and a novel Mining closed frequent similar pattern algorithm, named MCFSP, that finds a reduced closed set of frequent similar patterns without information loss (see Fig. 1 to see the scope of our work).

The outline of this paper is as follows. In Section 2 related work is reviewed. Section 3 provides basic concepts. In Section 4 a novel algorithm for mining closed frequent similar patterns is proposed. Section 5 some conclusions and future work are discussed.

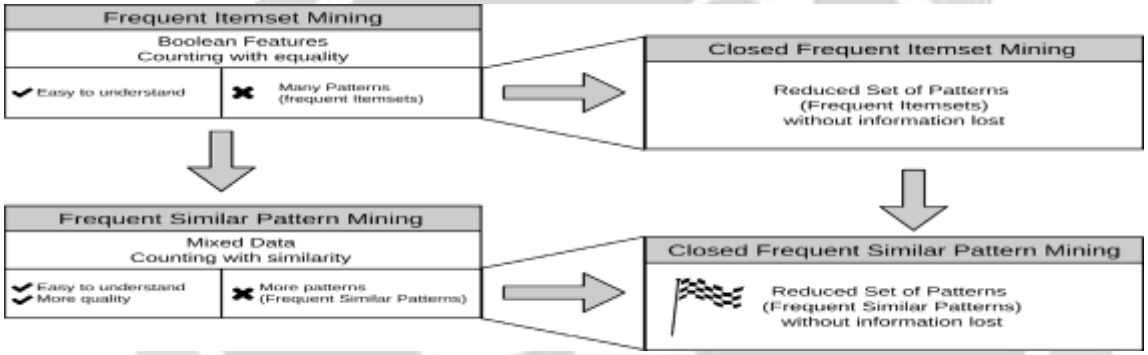

Fig 1. The journey from frequent itemset mining to closed frequent similar pattern mining.

2. Related work

Danger, Ruiz-Shulcloper, and Berland(2004) have proposed ObjectMiner which was the first algorithm that used similarity functions for mining frequent patterns. In order to allow pruning the search space of frequent similar patterns, this algorithm was designed for similarity functions that hold: if two objects are not similar with respect to a feature set S then they are not similar with respect to any superset of S. The main weakness of ObjectMiner is that although descriptions or subdescriptions of objects are usually repeated in the datasets, it does not use this fact in order to reduce the number of operations on subsequent steps.

Rodriguez-Gonzalez et al, (2008) have proposed STreeDC algorithm which fulfils Downward closure property. STreeDC algorithm builds a structure called STree. Each STree is a tree where each path from the root to leaf represents a sub description. The problem with this algorithm is the similarity function must be booleanized, which could lead to losing information.

Rodriguez-Gonzalez et al, (2008) have proposed STreeNDC algorithm which does not fulfils Downward closure property. STreeNDC is one efficient solution to the problem of frequent similar pattern mining for collections of objects described by a small set of features. The problem with this algorithm is the similarity function must be booleanized, which could lead to losing information.

Rodriguez-Gonzalez et al, (2008) have proposed DC-SP Miner algorithm, which uses various properties like Monotony of the frequency and fs-Downward Closure etc to prune the search space of frequent similar pattern. The problem with this algorithm is huge frequent pattern sets are generated which leads to increased computational time.

Closet (Pei et al.,2000) is based on: i) compressing frequent patterns in tree structure containing the frequent patterns for mining closed itemsets without candidate generation, ii) com- pressing a single path in the tree to do a fast identification of the frequent closed itemsets, iii) performing a partition-based projection mechanism for scalable mining in large databases. *Closet* uses a divide and conquer method for mining frequent closed patterns. First, frequent items are found and sorted in descending frequency order. Then, the search space is divided into non-overlapping subsets and each subset of frequent closed itemsets is mined recursively by constructing related conditional databases.

CHARM (Zaki & Hsiao, 2002), on the other hand, uses a bottom up approach for mining the closed frequent itemsets. It explores both itemset and transaction spaces, through a dual itemset-tidset search tree, using an efficient hybrid search that skips many levels in the tree during the search. CHARM also uses a technique called diffsets to reduce the memory footprint of intermediate computations. Finally, it uses a fast hash-based approach to remove any non-closed sets found during the search.

LCM (Uno et al., 2003) is another closed frequent itemsets mining algorithm. That defines a parent-child relationship between closed patterns. It was proven that each parent-child relationship forms a tree from which all closed patterns can be found by traversing it. LCM , also introduced an efficient way to traverse each tree in polynomial time with respect to the amount of closed frequent itemsets in the datasets.

3 Basic concepts and notations

In this Section, some concepts related to frequent similar pattern mining and closed frequent pattern mining are introduced. First, common concepts are described. Secondly, frequent similar pattern mining concepts are enumerated. Thirdly, closed frequent pattern mining concepts are also enumerated.
Consider a dataset as a tuple $D = (O, A, V, P)$ where O is a non-empty and finite set of objects, A is a non-empty and finite set of features, V is a non-empty and finite set of values and P is an application such that $P : (O \times A) \rightarrow V$. For simplicity, $O[A]$ is denoted as $P(O, A)$, $\forall O \in O$, $\forall A \in A$.

Definition 1 (Domain of a feature)**.** The Domain of a feature $A \in A$ is the application $Domain : A \rightarrow 2^V$ defined as $Domain(A) = \{V \in V \mid \exists O \in O : V = O[A]\}$.

Definition 2 (Pattern)**.** A pattern in D is a pair $T = (O, A_T) \in (O \times \{2^A \backslash \{\}\})$. $T.O$ denotes O and $T.A$ denotes A_T . Also, T denotes the set of all patterns in D . T is a non-empty and finite set. Given two patterns $T1 \in T$ and $T2 \in T$, $T1 = T2$ iff $T1.A = T2.A$ and $\forall A \in T1.A$; $T1.O[A] = T2.O[A]$.

Definition 3 (Boolean similarity function)**.** A Boolean similarity function in D is an application $F : (T \times O) \rightarrow \{$ True, False$\}$, such that $\forall\ T \in T$, $O \in O$; $T.O = O \Rightarrow F(T, O) = True$. B denotes the set of all Boolean similarity functions that can be defined in D.

Definition 4 (Occurrences)**.** The occurrences of a pattern in D is the application $OccurrencesF : T \rightarrow 2^O$, such that $OccurrencesF(T) = \{O \in O \mid F(T, O) = True\}$.

Definition 5 (Frequency)**.** The frequency of a pattern in D is the application $FrequencyF : T \rightarrow \{1, 2, \ldots, \|O\|\}$, such that $FrequencyF(T) = \|OccurrencesF(T)\|$.

Definition 6 (Frequent Similar Pattern)**.** A Pattern $T \in T$ is a frequent similar pattern in D if $FrequencyF(T) \geq M$ where M is a mini- mum threshold. M denotes the domain of M, $M = \{1, 2, \ldots, \|O\|\}$. Also, S denotes the set of all frequent similar patterns in D.

With the above definitions , a *Frequent Similar Pattern Mining* problem can be stated as follows: Given a dataset $D = (O, A, V, P)$, $F \in B$ and $M \in M$, the frequent similar pattern mining problem consists in finiding the set of all frequent similar patterns S.

However, pruning the search space of frequent similar patterns is needed for frequent similar pattern mining. Some definitions useful to this end are:

Definition 7 (Non-increasing monotonic boolean similarity func- tion)**.** F is a non-increasing monotonic boolean similarity function iff $\forall O, T, T_{sup}; O \in O ; T \in T ; T_{sup} \in SupPatterns(T)$ $[F (T, O) = False] \Rightarrow [F (T_{sup}, O) = False]$. N denotes the set of all non-increasing monotonic boolean similarity functions that can be defined in D.

Definition 8 (Closed Pattern)**.** Given a dataset $D = (O, A, V, P)$and $F = F_{eq}$, a pattern $T_e \in T$ is a closed pattern in D if $\forall TeSup \in SupPatterns(T_e)$ $FrequencyF (TeSup) < FrequencyF (T_e)$. E_{eq} denotes the set of all closed patterns in D using $F = F_{eq}$

Definition 9 (Closure)**.** Given a dataset $D = (O, A, V, P)$ and $F = F_{eq}$, the closure is the application $Closure : T \rightarrow E_{eq}$, such that $Closure(T) = T_{Cl} \in E_{eq} | T_{Cl} \in SupPatterns(T)$ y $Frequency F (T) = FrequencyF (T_{Cl})$.

3.1 Combining frequent similar pattern and closed frequent similar pattern concepts

Definition 10 (Closed Similar Pattern) **.** Given a dataset D = (O, A , V, P) and $F \in$ N , a closed similar pattern in D is a pattern $T e \in$ T | $\forall$ $T eSup \in SupPatterns$ $(T e)$ $Frequency F (T eSup) < Frequency F (T e)$, where E denotes the set of all closed similar pattern in D using F .

Definition 11 (Closed Frequent Similar Pattern) **.** The closed frequent similar pattern mining problem consists in finding the set of all frequent similar patterns in S ∩ E .

4. Mining Closed frequent similar pattern algorithm (MCFSP)

In this section we introduce the MCFSP algorithm for mining closed frequent similar patterns when similarity functions hold the downward closure property from The algorithm works by traversing a tree, defined by a father-child relation that contains all the closed frequent similar patterns.

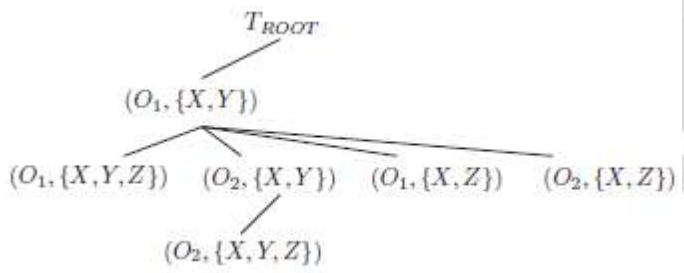

Fig. 2. Sample of father-child relation tree.

Definition 12 (Father-child relation among closed similar patterns) **.** Father of closed similar pattern is the application $Father : E \rightarrow E \cup T ROOT$,

Definition 13 (Father-child relation graph) **.** G is a the undirected graph, such that $G = (V, A)$ where $V = E \cup \{ T ROOT \}$ and $A = \{ (T 1 , T 2) \in (V \times V) | Father(T 2) = T 1 \}$.

Algorithm : MCFSP(*D, F, M, T*).

Input: Dataset *D = (O, A, V, P)*,

Similarity Function $F \in N$,

Minimum Frequency Threshold $M \in M$

Frequent Closed Similar Pattern $T \in S \cap E_{eq}$

Output: Frequent Closed Similar Patterns Set SE_{eq}

Consider a dataset as a tuple D = (O, A , V, P) where O is a non-empty and finite set of objects, A is a non-empty and finite set of features, V is a non-empty and finite set of values and P is an application such that P : (O ×A) → V . For simplicity, *O* [*A*] is denoted as P(*O, A*) , ∀ *O* ∈ O, ∀ *A* ∈ A .

Table 1. Description of datasets.

Datasets	Objects	Non-numeric features	Numeric features
Dermatology	**366**	**34**	**1**
Flags	**194**	**20**	**10**
Mushroom	**8124**	**22**	**0**
Waveform	**5000**	**40**	**1**
Vehicles	**946**	**1**	**18**
Wine	**178**	**1**	**13**

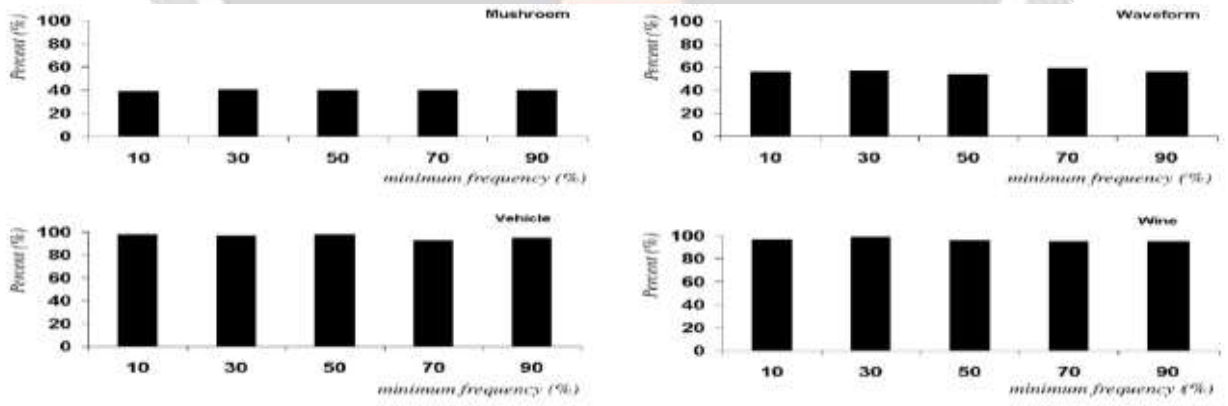

Fig. 3. Percent of closed frequent similar patterns respect to frequent similar patterns.

It is important to highlight that the main characteristic of MCFSP is its ability to find the "closed" similar patterns, yielding a reduction in the number of frequent similar patterns without information loss. To analyze in more detail the behavior of MCFSP , different datasets with a defined percent of frequent similar patterns that are closed were automatically generated.

First, three natural numbers X (for the amount of objects), Y (for the amount of features), and Z (for the cardinality of the feature domains) are fixed. Then, D is built such that O = X, A = Y and P(O, A) = V random ∀ O ∈ O and

∀ A ∈ A , where V random is a value for the random feature V with uniform distribution over { 1 , 2 , . . . , Z} . O (X) A (Y) D (Z) denotes the set of all datasets D defined by X, Y and Z .

5. Conclusion

In this paper we proposed the concept of closed frequent similar pattern mining for discovering a reduced set of frequent similar patterns without information loss. We also proposed a novel mining closed frequent similar pattern algorithm, named MCFSP, that uses boolean monotonic similarity functions to find all the closed frequent similar patterns. For future work, we visualize improving the efficiency of MCFSP, exploring the ideas proposed in the most recently works to improve the efficiency of traditional closed frequent itemset mining algorithms and studying the feasibility of extending these results to closed frequent similar pattern mining. Extending closed frequent similar patterns mining and association rules mining for non-boolean similarity functions and non-monotonic similarity functions is another interesting future work.

6. References

[1] Han, J., Cheng, H., Xin, D., Yan, X.: Frequent Pattern Mining: Current Statusand Future Directions. Data Mining and Knowledge Discovery 15(1), 55–86 (2007)

[2] Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: 1993 ACM SIGMOD International Conference on Management of Data, Washington, USA, pp. 207–216 (1993)

[3] Danger,R.,Ruiz-Shulcloper, J., Berlanga, R.: Objectminer: A New Approach for Mining Complex Objects. In: Sixth International Conference on Enterprise Information Systems, Oporto, Portugal, pp. 42–47 (2004)

[4] Rodr´ıguez-Gonz´alez, A.Y., Mart´ınez-Trinidad, J.F., Carrasco-Ochoa, J.A., Ruiz-Shulcloper, J.: Mining Frequent Similar Patterns on Mixed Data. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 136–144.Springer, Heidelberg (2008)

[5] Rodr´ıguez-Gonz´alez, A.Y., Mart´ınez-Trinidad, J.F., Carrasco-Ochoa, J.A., Ruiz- Shulcloper, J.: Mining frequent patterns and association rules using similarities. Expert Systems with Applications, 40(17), 6823–6836.

[6] Pei, J. , et al. (2000). Closet: An efficient algorithm for mining frequent closed item- sets.. In Acm sigmod workshop on research issues in data mining and knowledge discovery: 4 (pp. 21–30) .

[7] Zaki, M. J. , & Hsiao, C.-J. (2002). Charm: An efficient algorithm for closed itemset mining.. In Sdm: 2 (pp. 457–473). SIAM .

[8] Uno, T. , Asai, T. , Uchida, Y. , & Arimura, H. (2003). Lcm: An efficient algorithm for enumerating frequent closed item sets.. Fimi : 90. Citeseer .

[9] Prabha, S., Shanmugapriya, S. , & Duraiswamy, K. (2013). A survey on closed frequent pattern mining. International Journal of Computer Applications, 63 (14) .