ISSN 1870-4069

Development of a Metaheuristic to Obtain Frequent Similar Patterns

Gretel Bernal-Baró, Ansel Y. Rodríguez González,
Rosa M. Valdovinos Rosas

Universidad Autónoma del Estado de México,
Facultad de Ingeniería,
Mexico

gretelbernalbaro@gmail.com

Abstract. Most frequent pattern mining algorithms assume that two sub-descriptions of instances are similar if and only if they are equals. However, other similarity functions are used in the soft sciences. In fact, some algorithms find patterns using similarity functions other than equality, but those algorithms are shown difficulties for mining datasets that involve a large variety of attribute values. Although these algorithms find a complete set of Frequents Similars Patterns (FSPs), the number of patterns mined is often too large, requiring high analysis and processing costs. The research proposed here is focused on developing a metaheuristic for obtaining a representative subset of FSP, which describes the whole dataset, from large mixed datasets in a shorter time than the traditional algorithms used for mining FSPs.

Keywords: Frequent patterns, similarity, mixed data, data mining.

1 Introduction

A frequent pattern is a combination of feature values of an instance in the dataset that appears with a frequency not less than a user-specified frequency threshold [1]. Most current algorithms for frequent patterns mining assume that two instances are similar when they are equal. However, in many real-world problems, two instances can be considered similar even though they are not identical. In these problems, the concept of similarity between instances descriptions is used to compare instances and count how many times an instance appears in a dataset. When similarity functions different than equality are used, new patterns emerge called Frequent Similar Patterns (FSP). Thus, a FSP is a combination of feature values of the instance in the dataset, such that the accumulation of frequency of its similar patterns is not less than a user-specified frequency threshold. [4].

In the literature several algorithms are proposed for finding FSPs (ObjectMiner [2], STreeDC-Miner [4] and STree NDC-Miner [4]), but these algorithms' behavior is affected negatively when large volume or high-dimensional datasets are used. When the similarity is the criteria considered to obtain the frequency calculation, more operations need to be done, and as a consequence, the runtime required by these algorithms is increased. Furthermore, although new patterns can be found using these algorithms, the number of mined patterns is often too large, making additional analysis and processing costs. The research here presented proposes to develop a metaheuristic to mine a subset of representative FSPs, which explores the search space more efficiently than the existing algorithms used in mining FSPs. Specifically, the research is focused on: To develop a mechanism for the efficient retrieval of instance sub-descriptions and their frequencies. To propose a quality measure of a FSP that allows obtaining a representative subset of FSPs that describes the dataset. If is required, to adapt the existing theory about the mining frequent patterns to mining FSPs. Finally, to develop a metaheuristic for mining FSPs in mixed datasets.

1.1 Motivation and Justification

Several studies demonstrated that, when the similarity concept is included in the frequency calculation, it is possible to discover hidden patterns in the traditional frequent patterns context [2]. Similarly, other studies validate that when using the FSPs in tasks such as classification, the classifier obtains higher precision when new objects are classified. In fact, it is unnecessary to extract the entire set of FSPs; for example, a reduced set of frequent similar patterns without information loss, named Closed Frequent Similar Patterns, can be extracted [3]. In addition, some important difficulties of the FSP mining algorithms are: high computational cost required for finding the solutions, prohibitive in datasets with high dimensionality (more than hundreds of attributes), and too many FSPs obtained from the mining process, resulting in analysis and processing costly.

2 Previous Works in the Area

In the state of art, there are several algoritms for obtain frequent pattenrs, however, related with FSP are too few. The referent research could be the CFSP-Miner algorithm [3] which discover a subset of FSPs, but the number of FSPs mined and the runtime can be high too. On the other hand, ObjectMiner [2] was the first algorithm that uses similarity functions for mining frequent patterns, is inspired in the Apriori algorithm, and includes a pruning method to diminish the search space.

STreeDC-Miner [4] works by following a depth-first search strategy, using a tree structure called STree. In the STree each path from the leaf to the root represents a sub-description, where the same sub-descriptions are grouped. Each leaf stores the repetitions of the pattern and also holds the similarity among this pattern and its similar patterns. STreeNDC-Miner [4], like STreeDC-Miner, assumes an order among the features that describe the instances. But, unlike ObjectMiner and STreeDC-Miner, it does not include a method for pruning the search space. However, the computational effort for searching all FSPs is reduced using a top-down strategy and the STree data structure.

3 Hypothesis

With the development of a metaheuristic for mining a subset of FSPs, in mixed datasets, a subset of FSPs will be obtained with quality greater than or equal to 75%, in less runtime than the current algorithms that mine FSPs.

4 Methodology

To meet the stated objective, it is proposed to follow the following methodology:

1. To obtain from the repository textit UCI machine learning repository [1] datasets with dimensionality and size required for validate of the proposed metaheuristic.
2. To develop a metaheuristic for mining FSPs.
   - To develop a mechanism for the efficient retrieval of instance sub-descriptions.
   - To define a new fitness function to measure the quality of a FSP.
   - To implement the different metaheuristics used in traditional frequent pattern mining and develop a new metaheuristic to mine a subset of FSPs.
3. To perform tests and compare the results.
   - To evaluate the performance of the proposed metaheuristic according the run-time required and quality of the FSPs obtained.
   - To use the FSPs mined for classification purposes and to evaluate the subsets quality, based on the classifier accuracy.

5 State of the Research

The research is at the beginning of the second year. During the first year, a data structure was developed to represent the dataset efficiently. For that, the dataset is divided into several sub-trees, called FP-Similar-Tree. Each FP-Similar-Tree is associated with a FSP and represents a compact structure that stores quantitative information about the FSPs presents in the dataset. Several experiments were carried out using ten datasets from the UCI machine learning repository to evaluate the FP-Similar-Tree behavior. The number of instances in the selected datasets varies between 4,000 and 1,000,000 to assess their performance in different scenarios. To analyze the FP-Similar-Tree improvements, the number of mined patterns and the required run-time was measured concerning the STreeDC-Miner algorithm. This algorithm was selected for being one that shows the best performance FPS mining nowadays.

The results obtained show that both algorithms found the same number of patterns in the datasets tested (ten datasets), except in two. In these two datasets, the proposed algorithm, FP-Similar-Tree, mined all FSPs while

[1] https://archive.ics.uci.edu/ml/index.php

the STreeDCMiner algorithm did not work due to memory requirements. FP-Similar-Tree algorithm got the best results in terms of runtime in the other eight of the ten datasets tested, achieving an improvement of up to 57 %. Therefore, it can be assured that the proposed data structure is an excellent alternative to extract a subset of frequent similar patterns.

6 Conclusions

Nowadays, the mining of FSP is strongly attracting attention as an alternative solution in the development of descriptive strategies. The main problem identified in the existing methods is difficult for dealing with high dimensionality data and the large number of mined patterns. About it, the first development made in the research proposed was to build a data structure capable of representing the dataset more efficiently, the FP-Similar-Tree. The structure proposed was compared with the STreeDC-Miner algorithm to analyze its effectiveness. The preliminary results allow us to show that the FP-Similar-Tree mine a whole set of FSP in mixed data collections in lower run-time than the time reached by one of the most competitive algorithm, STreeDC-Miner. The main improvements of the FP-Similar-Tree are: Reduction the number of comparisons made between the sub-descriptions of the dataset and fewer accesses to the dataset because only is required single access. As future work, we plan to develop a quality measure that allows the generation of a subset of representative FSPs. In addition to developing a metaheuristic that uses the data structure and the quality measure proposed.

References

1. Baró, G. B., Martínez-Trinidad, J. F., Rosas, R. M. V., Ochoa, J. A. C., González, A. Y. R., Cortés, M. S. L.: A pso-based algorithm for mining association rules using a guided exploration strategy. Pattern Recognition Letters, vol. 138, pp. 8–15 (2020)
2. Danger, R., Ruíz-Shulcloper, J., Llavori, R. B.: Objectminer: A new approach for mining complex objects. In: ICEIS (2). pp. 42–47. Citeseer (2004)
3. Rodríguez-González, A. Y., Lezama, F., Iglesias-Alvarez, C. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., de Cote, E. M.: Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss. Expert Systems with Applications, vol. 96, pp. 271 – 283 (2018) doi: https://doi.org/10.1016/j.eswa.2017.12.018
4. Rodríguez-González, A. Y., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., Ruiz-Shulcloper, J.: Mining frequent patterns and association rules using similarities. Expert Systems with Applications, vol. 40, no. 17, pp. 6823 – 6836 (2013) doi: https://doi.org/10.1016/j.eswa.2013.06.041