Evolving temporal fuzzy association rules from quantitative data with a multi-objective evolutionary algorithm

MATTHEWS, Stephen G., GONGORA, Mario A. and HOPGOOD, Adrian A.

Published version

MATTHEWS, Stephen G., GONGORA, Mario A. and HOPGOOD, Adrian A. (2011). Evolving temporal fuzzy association rules from quantitative data with a multi-objective evolutionary algorithm. In: Hybrid Artificial Intelligent Systems. London, Springer, 198-205.

Evolving Temporal Fuzzy Association Rules from Quantitative Data with a Multi-Objective Evolutionary Algorithm

Stephen G. Matthews, Mario A. Gongora, and Adrian A. Hopgood

Centre for Computational Intelligence,
De Montfort University, Leicester, UK
{sgm,mgongora,aah}@dmu.ac.uk
http://www.cci.dmu.ac.uk/

Abstract. A novel method for mining association rules that are both quantitative and temporal using a multi-objective evolutionary algorithm is presented. This method successfully identifies numerous temporal association rules that occur more frequently in areas of a dataset with specific quantitative values represented with fuzzy sets. The novelty of this research lies in exploring the composition of quantitative and temporal fuzzy association rules and the approach of using a hybridisation of a multi-objective evolutionary algorithm with fuzzy sets. Results show the ability of a multi-objective evolutionary algorithm (NSGA-II) to evolve multiple target itemsets that have been augmented into synthetic datasets.

Keywords: multi-objective evolutionary algorithm, fuzzy association rules, temporal association rules, NSGA-II, hyrbid

1 Introduction

Association rule mining is a well established method of data mining that identifies significant correlations between Boolean items in transactional data [1]. This paper extends the classical problem by exploring the composition of two variants of association rule mining with a hyrbid approach.

It is often assumed in classical association rule mining that the dataset is static, meaning that discovered rules hold across the entire period of the dataset. However, real-world datasets can have underlying temporal patterns. For example, an increase in association rule frequency may occur before a large sports event or when an unforeseen events occur, such as hurricanes (e.g., [2]). Quantitative association rule mining [3] discovers rules that express associations between intervals of item attributes (e.g. height, pressure), but common approaches of discretisation can lead to a loss of information. Evolutionary Computing (EC) has been used to remove the requirement for prior discretisation and the synergy of hybridising EC with fuzzy sets has become popular for data mining tasks [4, 5] such as classification and association rule mining.

The composition of temporal association rule mining and quantitative association rule mining is treated as a multi-objective optimisation problem. The aim is to extract temporal association rules from quantitative data using fuzzy sets. The temporal association rules sought are those that occur more frequently over an interval of the dataset, which are seen as an area of greater density. The advantages of fuzzy sets are they allow a linguistic interpretation, a smoother transition between boundaries, and better handle uncertainty. The itemset/association rule space, temporal space and quantitative space are simultaneously searched and optimised. This paper extends our previous work in [6] by including a quantitative element, mining multiple occurrences of association rules and by directly mining association rules.

This paper is organised as follows: Section 2 presents an overview of related works on association rule mining; Section 3 describes the multi-objective evolutionary algorithm for mining temporal fuzzy association rules from quantitative data; Section 4 presents results; and conclusions are drawn in Section 5.

2 Quantitative and Temporal Association Rule Mining

A disadvantage of classical quantitative association rule mining is the crisp boundaries of discretised values that potentially hide rules and lose information [8]. Soft computing techniques can overcome this issue, for example, in [8], a genetic algorithm evolves attribute intervals for a fixed number of attributes. Fuzzy association rules deal with inaccuracies in physical measurements and better handle unnatural boundaries found in crisp partitions. They provide a linguistic interpretation of numerical values for interfacing with experts. Evolving fuzzy association rules [9] enhances the interpretability of quantitative association rules.

There are two common approaches to mining quantitative association rules. One approach is to tune membership functions and use a deterministic method to induce rules afterwards (e.g., [10]). Membership functions are tuned to produce maximum support for 1-itemsets before exhaustively mining rules. Another approach is to extract association rules whilst defining attribute intervals [8] or membership functions [9]. The latter approach is adopted in this paper.

A key issue of classical methods, based on the support-confidence framework, is that temporal patterns with low support can escape below the minimum support threshold. For example, supermarket items may be sold only during particular seasonal periods, resulting in annual support values dropping below a minimum threshold, despite having sufficient support values in a seasonal period. The *lifespan* property [11] is an extension on the Apriori algorithm [19] that incorporates temporal information. This measure of support is relative to the lifespan of the itemset defined by a time interval, known as temporal support, corresponding to the first and last occurrences of the itemset. But this does not consider datasets where the frequency of rules may be skewed towards particular areas whilst still occurring throughout the entire dataset.

A step towards analysing areas of a dataset where rules occur more frequently is cyclic association rule mining [12]. Cyclic rules are induced from user-defined partitions of regular periods and pattern matching is performed on binary sequences. Other temporal patterns that can potentially be extracted with our method are partially periodic rules [13] and calendar-based schemas [14].

Our previous work [6] has demonstrated mining association rules that occur more frequently over single areas of a dataset with a single objective genetic algorithm. A multi-objective evolutionary algorithm (MOEA) is used in [7] and extended here to include association rules and multiple targets.

3 Multi-Objective Evolutionary Search and Optimisation

Extracting a set of fuzzy association rules from areas of the dataset where the occurrence is greater is treated as a multi-objective problem. This is the optimisation of two or more functions, whilst satisfying optional constraints [15]. Optimal solutions found with a MOEA are compromises between objectives and such trade-offs are often managed with the concept of Pareto optimality. A solution is said to be Pareto optimal when no change in the solution will improve one objective without degrading another objective.

A Pareto based MOEA is capable of producing multiple association rules from a single run through utilising a maintained set of maximally-spread Pareto-optimal solutions. This is desirable when the cardinality of the optimal set may be more than one, for instance in the case of multiple temporal patterns. This improves our previous work [7] which mines single temporal patterns. From the plethora of MOEAs, we selected NSGA-II [16] for its popularity and ability to maintain a diverse set of solutions suitable for extracting multiple patterns. Previous works have used NSGA-II for Subgroup Discovery [17], a closely related area, and motif sequence discovery [18], a different form of temporal mining.

3.1 Representation

A Michigan approach and mixed coding scheme is used to represent the temporal interval and fuzzy association rules as

$$C = (t_0, t_1, i_0, a_0, b_0, c_0, A_0, \ldots, i_k, a_k, b_k, c_k, A_k) \tag{1}$$

where the temporal interval is defined with t_0 and t_1 as integers. The items are integers denoted with i and the basic parameters of the triangular membership functions are real numbers indicated with a, b and c for association rules with k distinct items. A binary value in A_k determines if this item belongs to the antecedent or consequent.

3.2 Objectives

Temporal Support: The temporal support objective, ts, guides the MOEA to find itemsets that occur more frequently in areas of the dataset. Modified from [11], this is redefined as a minimisation function

$$ts(X, l_X) = 1 - \frac{\sigma(X)}{l_X} \tag{2}$$

with l denoting a time interval i.e. $l_X = t_1 - t_0$ where t_0 is the lower endpoint, t_1 is the upper endpoint and $\sigma(X)$ is the itemset support. A minimum temporal support is used to prevent solutions evolving towards the smallest time interval of length 1, which would produce a 100% temporal support.

Temporal Rule Confidence: Temporal confidence, tc, helps extract association rules from itemsets. This aims to identify specific association rules that have a temporal occurrence based on temporal support.

$$tc(X \Rightarrow Y, l_{X \cup Y}) = \frac{ts(X \cup Y, l_{X \cup Y})}{ts(X, l_X)} \tag{3}$$

Fuzzy Rule Support: This objective optimises the membership function parameters of matching association rules. The quantitative values are modelled with triangular fuzzy sets and the objective's optimal solution is one where the fuzzy sets support the quantitative values associated with the association rule to the highest degree of membership. Fuzzy rule support, fs, is the sum of the degrees of memberships, $sum(\mu(x^{(i)}))$, for a chromosome itemset, $x^{(i)}$, in the ith transaction.

$$fs = (k \cdot (t_1 - t_0)) - \sum_{i=t_0}^{t_1} sum(\mu(x^{(i)})) \tag{4}$$

$$sum(\mu(x^{(i)})) = \sum_{j=0}^{k} \begin{cases} \mu(x_j^{(i)}), & \text{dataset item matches gene item} \\ 0, & \text{otherwise} \end{cases} . \tag{5}$$

$$\mu(x_j^{(i)}) = \begin{cases} \frac{x_j^{(i)} - a}{b - a}, & \text{if } a \leq x_j^{(i)} < b \\ \frac{c - x_j^{(i)}}{c - b}, & \text{if } b \leq x_j^{(i)} \leq c \\ 0, & \text{otherwise} \end{cases} . \tag{6}$$

Equation 4 subtracts the sum of the actual degrees of memberships from the maximum possible sum if all items in a transaction match those in the chromosome. Equation 5 performs the summation of actual degrees of memberships for chromosome items matching dataset transaction items. Equation 6 is the triangular memerbship function.

Membership Function Widths: The aim of this objective is to prevent the membership function parameters evolving to cover the entire range of values i.e. the feet of the membership function (a and c) nearing the limits of the attribute values. Without this objective solutions evolve to cover the entire range of attribute values because this yields higher support values as it includes more items.

$$mf\_widths = \begin{cases} \sum_{j=0}^{k} c_j - a_j, & \text{if } c_j - a_j > 0 \\ nitems, & \text{otherwise} \end{cases} . \tag{7}$$

3.3 Initialisation and Genetic Operators

The initial population is randomly generated with lower and upper endpoints being within proximity to the first and last transactions. An endpoint range is defined for two purposes: limit the range for creating endpoints and also for mutating endpoints. Time endpoints initialised near dataset boundaries provide starting solutions with large temporal coverages of the dataset. Without the endpoint range, random sampling of time intervals occurs. This may lead to some potentially strong itemsets being lost, so an initial large temporal coverage, combined with the mutation operator, provides more opportunity for solutions with great potential that initially may be weak.

Crossover is adapted to handle quantitative data from the method proposed in [6]. For mutating genes that form the time interval endpoints, the values are generated within the endpoint range (epr) where the midpoint is the value of the current gene (g), such that the mutated value is a member of the set $\{-epr/2, \ldots, g, \ldots, epr/2\}$. This reduces the effect of randomly sampling the dataset. The endpoint range is decremented every generation until reaching 10, to allow further mutations.

4 Experimental Study

4.1 Methodology

The IBM Quest Synthetic Data Generator [19] has been extended to include quantitative attributes. The dataset has these features: 1000 transactions, 50 items, an average transaction size of 10 and a maximum size for quantitative values of 20. The Apriori algorithm is used to identify relatively low (0.2%), medium (1.7%) and high support (3.5%) association rules that are augmented into areas of the dataset to produce temporal patterns. This is based on the process defined in [6] that creates target temporal patterns with varyling levels of difficulty and is extended to include multiple temporal patterns. The minimum temporal support was set to 30. Figure 1 depicts the frequency of a quantitative itemset augmented into the first half and second half of a dataset to demonstrate the increased occurrence of the same pattern in two areas. Table 1 shows the itemsets used for augmentation. Augmentation is based on itemset support because this is used to extract fuzzy association rules.

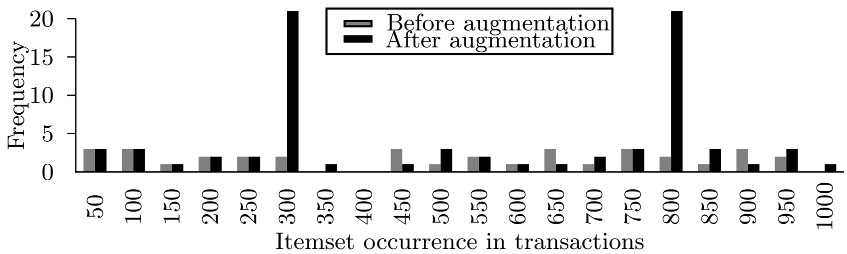

Fig. 1. Histogram of itemset {8, 12, 21, 45} with high support (3.5%) (Bins 250 and 750 have one extra itemset that does not contain the quantitative values).

4.2 Results

The augmented itemsets were identified with 50 runs of NSGA-II on each dataset, although with varying success for different levels of difficulty. The results are summarised in Table 1. The itemsets were deemed to be successfully identified if the entire itemset matched that of the augmented itemset and it was in proximity of the endpoints, t_0 and t_1. The number of temporal patterns identified increases with the support level of the augmented itemset. For each level of difficulty there is one area of the dataset that is more likely to be identified as a temporal pattern. For example, the high support (3.5%) dataset identified the 1st temporal pattern (transactions 250–289) in 12 runs while identifying the 2nd (transactions 750–788) in 38 runs. Also, with a higher support value of augmented itemsets there is an increase in identifying both temporal patterns. The correct identification of the quantitative attributes with fuzzy sets varies greatly and not all attributes were correctly identified in a solution.

Table 1. Results of augmenting same quantitative itemset, or temporal patterns (TP), in two locations

Endpoint		Itemset	Aug.	TP	Qty. of TP identified	
t_0	t_1		Sup.	identified	1	2
250	289	24 (3) 31 (7) 32 (12) 38 (16)	0.2%	8	7	1
750	788	24 (3) 31 (7) 32 (12) 38 (16)		1		
250	289	12 (3) 31 (7) 41 (12) 48 (16)	1.7%	19	15	8
750	788	12 (3) 31 (7) 41 (12) 48 (16)		12		
250	289	8 (3) 12 (7) 21 (12) 45 (16)	3.5%	47	12	38
750	788	8 (3) 12 (7) 21 (12) 45 (16)		41		

Three of the objectives are plotted in Figure 2, both augmented itemsets in the final solution are distinguished here. This graph can be used to view the trade-offs between fuzzy association rules, which is of particular use for knowledge discovery in real-world applications. This figure demonstrates how the objectives conflict, particularly for membership function widths and fuzzy rule support.

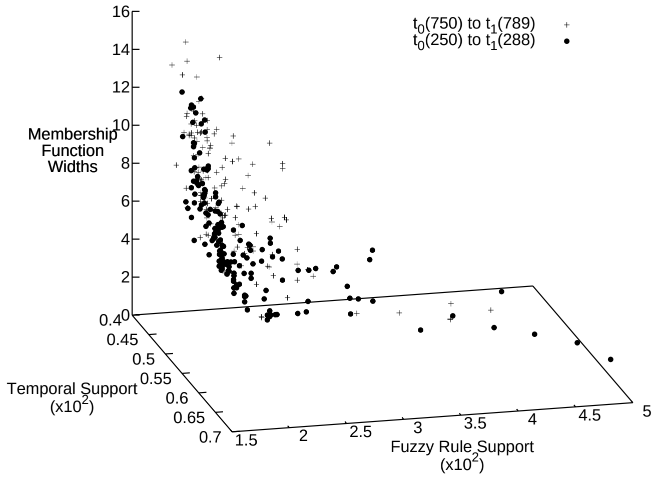

Fig. 2. Three objectives for best solutions in a portion of a final population augmented with a high support (3.5%) itemset

5 Conclusions

We have used a hyrbid approach of a MOEA (NSGA-II) and fuzzy sets to evolve multiple temporal association rules from quantitative transaction data. This demonstrates the ability to find association rules that occur more frequently in numerous areas of a dataset. A MOEA maintains diversity and so allows for numerous temporal patterns to evolve. The advantages of the proposed approach is that it does not exhaustively search the various spaces, it requires no discretisation and yields numerous diverse association rules.

Future work will explore enhancing the robustness of identifying quantitative attributes and evolving low support itemsets. Real-world datasets will be used as these are crucial to demonstrating the impact of this research. We will compare statistical methods, such as temporal based Apriori methods, and other MOEAs with this approach to explore its suitability.

Acknowledgements Supported by an Engineering and Physical Sciences Research Council Doctoral Training Account.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD ICDM, Washington, DC, USA, pp. 207–216 (1993)
2. Leonard, D.: After Katrina: Crisis Management, the Only Lifeline Was the Wal-Mart. FORTUNE Magazine (2005)

3. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: ACM SIGMOD ICDM, Montreal, Quebec, Canada, pp. 1–12 (1996)
4. Ishibuchi, H.: Multiobjective Genetic Fuzzy Systems: Review and Future Research Directions Fuzzy Systems Conference. In: FUZZ-IEEE, London, UK, pp. 1–6 (2007)
5. Corchado, E., Abraham, A., de Carvalho, A.: Hybrid intelligent algorithms and applications. Information Sciences, 180(14), 2633–2634 (2010)
6. Matthews, S. G., Gongora, M. A., Hopgood, A. A.: Evolving Temporal Association Rules with Genetic Algorithms. In Bramer, M. and Petridis, M. and Hopgood, A. (eds.) Research and Development in Intelligent Systems XXVII, pp. 107–120. Springer, London (2010)
7. Matthews, S. G., Gongora, M. A., Hopgood, A. A.: Evolving Temporal Fuzzy Itemsets from Quantitative Data with a Multi-Objective Evolutionary Algorithm. In: IEEE SSCI, Paris, France, Accepted for publication (2011)
8. Mata, J., Alvarez, J. L., Riquelme, J. C.: An evolutionary algorithm to discover numeric association rules. In: ACM SAC, New York, NY, USA, pp. 590–594 (2002)
9. Kaya, M.: Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules. Soft Computing - A Fusion of Foundations, Methodologies and Applications, 10(7), 578–58 (2006)
10. Hong, T.-P., Chen, C.-H., Lee, Y.-C., Wu, Y.-L.: Genetic-Fuzzy Data Mining With Divide-and-Conquer Strategy. IEEE Transactions on Evolutionary Computation, 12(2), 252–265 (2008)
11. Ale, J. M., Rossi, G. H.: An approach to discovering temporal association rules. In: ACM SAC, Como, Italy, pp. 294–300 (2000)
12. Özden, B., Ramaswamy, S., Silberschatz, A.: Cyclic Association Rules. In: ICDE, Washington, DC, USA, pp. 412–421 (1998)
13. Han, J., Gong, W., Yin, Y.: Mining segment-wise periodic patterns in time-related databases. In: KDD, New York, NY, USA, pp. 214–218 (1998)
14. Li, Y., Ning, P., Wang, X. S., Jajodia, S.: Discovering calendar-based temporal association rules. Data & Knowledge Engineering, 44(2), 193–218 (2003)
15. Coello, C. A. C., Lamont, G. B., van Veldhuizen, D. A.: Evolutionary Algorithms for Solving Multi-Objective Problems. Springer (2007)
16. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2), 182–19 (2002)
17. Carmona, C., Gonzalez, P., del Jesus, M., Herrera, F.: NMEEF-SD: Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery. IEEE Transactions on Fuzzy Systems, 18(5), 958–970 (2010)
18. Kaya, M.: MOGAMOD: Multi-objective genetic algorithm for motif discovery. Expert Systems with Applications, 36 (2, Part 1), 1039–1047 (2009)
19. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: VLDB, Santiago, Chile, pp. 487–499 (1994)