Knowledge-Based Systems 54 (2013) 66–72

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Web usage mining with evolutionary extraction of temporal fuzzy association rules

Stephen G. Matthews [a,*], Mario A. Gongora [b], Adrian A. Hopgood [c], Samad Ahmadi [b]

[a] Intelligent Systems Laboratory, Department of Engineering Mathematics, University of Bristol, Bristol BS8 1UB, UK
[b] Centre for Computational Intelligence, Department of Informatics, De Montfort University, Leicester LE1 9BH, UK
[c] Sheffield Business School, Sheffield Hallam University, Sheffield S1 1WB, UK

ARTICLE INFO

Article history:
Available online 27 September 2013

Keywords:
Fuzzy association rules
Temporal association rules
Evolutionary fuzzy system
Genetic algorithm
Data mining
Analytics
Rule discovery
2-tuple linguistic representation

ABSTRACT

In Web usage mining, fuzzy association rules that have a temporal property can provide useful knowledge about when associations occur. However, there is a problem with traditional temporal fuzzy association rule mining algorithms. Some rules occur at the intersection of fuzzy sets' boundaries where there is less support (lower membership), so the rules are lost. A genetic algorithm (GA)-based solution is described that uses the flexible nature of the 2-tuple linguistic representation to discover rules that occur at the intersection of fuzzy set boundaries. The GA-based approach is enhanced from previous work by including a graph representation and an improved fitness function. A comparison of the GA-based approach with a traditional approach on real-world Web log data discovered rules that were lost with the traditional approach. The GA-based approach is recommended as complementary to existing algorithms, because it discovers extra rules.

1. Introduction

Web usage mining is one type of Web mining [23] that attempts to discover patterns of user behaviours that are recorded in the logs of Web servers as users browse Web sites [10]. In this paper, temporal fuzzy association rules are used for Web usage mining. For example, "On a Friday evening, visitors who viewed history.html for a large amount of time also viewed contact-us.html for a medium amount of time". Such rules extend traditional Boolean association rules [1] by incorporating temporal and fuzzy quantitative features. The temporal feature of the rule is *on a Friday evening*, and the fuzzy features are the *large* and *medium* descriptions.

Matthews et al. [25] discovered a problem of losing some rules when using traditional methods on synthetic market basket data. Traditional methods follow a two-step process of defining the linguistic labels and membership functions of those labels first, and using them in the mining process. However, the contextual meaning of the linguistic labels can change with events such as seasonal weather, sports games [28], or unforeseen events, e.g., hurricanes [22]. The problem is that although the meaning can change in a temporal period the membership functions remain the same. For example, a low quantity of ice cream sales in summer has a different meaning to a low quantity in winter. The membership function does not accurately define the linguistic label for some temporal periods.

Matthews et al. [25] created a solution that combined the flexibility of the 2-tuple linguistic representation [16] with the search power of a GA. The 2-tuple linguistic representation displaces membership functions laterally along the universe of discourse whilst the linguistic label remains the same. Previous work is improved in this paper by incorporating a graph data structure with an enhanced fitness function. The enhancements enable the approach to work on datasets with real-world complexity in a different domain.

This article is structured as follows: Section 2 provides an overview of related work, Section 3 describes the traditional approach and the original GA-based algorithm, Section 4 introduces enhancements to the GA-based algorithm that was applied to Web log data, Section 5 presents the evaluation of our approach compared with a traditional method, and conclusions are made in Section 6.

2. Related work

The application of Web usage mining has been categorised as either personalised for learning user profiles, or unpersonalised for user navigation patterns [30]. In this paper, we focus on user navigation patterns represented with fuzzy association rules. Web usage mining can be used for the personalisation of web content, pre-fetching and caching, enhancing Web site design, and

* Corresponding author. Tel.: +44 7760712738.
E-mail address: stephen.matthews@bristol.ac.uk (S.G. Matthews).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.09.003

customer relationship management in e-commerce [13]. Recent work has also applied similar techniques to those in this paper. GAs have mined sequence rules in Web log data [31] and have also performed subgroup discovery [6]. Fuzzy sets have been used to represent the time spent viewing Web pages for fuzzy association rules [33] and fuzzy sequence rules [19]. The temporal and fuzzy features of association rules that are mined in this paper are now reviewed.

The term *temporal* is ambiguous, because it can have different interpretations in temporal data mining [26]. In this paper, a temporal association rule expresses associations between items from the same transaction, and that association is repeated (occurs frequently) in multiple transactions of a subset of a dataset. For example, a rule may be present in several transactions, and that rule may occur more frequently on a Friday than any other day of that week. Exhibition periods [21] are temporal patterns that take into consideration the time when items were introduced into the dataset, e.g., new publications in a publications database. Cyclic patterns [27] have rules that occur more frequently in regular periods, such as a rule that occurs every weekend. Temporal patterns with partial periodicity [15] relax the regularity of cyclic patterns, so the rule may not be present in some cycles of the temporal pattern. These types of temporal association rules are *intra*-transactional, which is different to *inter*-transactional where rules contain items from several transactions spread over a period of time, such as sequence rules [2].

Quantitative association rule mining extends Boolean association rule mining by discovering rules in quantitative attributes [29]. For example, the time spent viewing a Web page, or the quantities of items sold in a shopping basket. Quantitative association rule mining discretises quantitative attributes into bins. Quantitative association rules suffer from the crisp boundary problem, so fuzzy association rules better deal with unnatural boundaries of crisp intervals [20] and inaccuracies with physical measurements [7]. Fuzzy sets [34] allow the quantities to be described with linguistic terms [35], such as *low* and *high*.

The temporal property of not discovering rare fuzzy itemsets [32] is different to our research, because we focus on how the fuzzy sets are defined instead of only the temporal property. Au and Chan [5] also mine fuzzy association rules in temporal partitions of the dataset, and they follow the same two-step process, which can lose rules.

3. Temporal fuzzy association rule mining

Two approaches for mining temporal fuzzy association rules were run on the United States Environmental Protection Agency (EPA) dataset. The purpose is to demonstrate how the flexibility of the 2-tuple linguistic representation approach can help to discover rules on real-world data that a traditional approach cannot. The two approaches are described here, and enhancements to the GA-based approach are explained in Section 4.

3.1. FuzzyApriori

FuzzyApriori [18] is an extension to the Apriori algorithm [1] that uses a breadth-first search. FuzzyApriori uses fuzzy sets to express quantities of items with linguistic terms, but it does not consider any temporal pattern. So, the dataset is partitioned according to its temporal dimension, such as by hour, and FuzzyApriori is executed on each dataset partition separately. The systematic search of the temporal dimension allows for the discovery of temporal features of fuzzy association rules. This is similar to the first approach for mining cyclic association rules [27] where the dataset is also partitioned according to the temporal dimension. The rules mined from each dataset partition are aggregated into a final rule set, which is the end result.

Due to the static nature of membership functions in existing approaches, not all temporal fuzzy association rules can be discovered, hence some are lost. Au and Chan [5] also mine fuzzy association rules in temporal partitions of the dataset, which has been discussed in Section 2. Au and Chan [5] use a different search method in the two-step process, but in theory the same problem of losing rules exists, because the fuzzy sets are defined first and they are static. For this reason, a method based on the seminal Apriori algorithm is only compared, i.e., FuzzyApriori.

3.2. CHC with 2-tuple linguistic representation

The GA-based approach was first described in Matthews et al. [25], so an overview is given before introducing enhancements in Section 4. The pseudocode of the algorithm is described in Section A. The GA-based approach by Matthews et al. [25] is not considered to be traditional like FuzzyApriori, because it is not an exhaustive search method. Instead, a stochastic search method is applied – a GA called Cross-generational elitist selection, Heterogeneous re-combination, and Cataclysmic mutation (CHC) [12]. The contextual change of meaning for linguistic labels is modelled with the 2-tuple linguistic representation, which is a flexible representation. The crucial difference from other temporal fuzzy association rule mining approaches is that Matthews et al. [25] simultaneously search for membership function parameters and the items in the rule, as well as the temporal period when the rule occurs. This overcomes the problem of membership functions remaining the same when there is a contextual change in the meaning of linguistic labels. Alternative GA-based approaches that simultaneously search for fuzzy sets and rules do exist, but they perform different tasks, i.e., control [17], classification [36], and fuzzy modelling [11]. The GA-based approach uses Iterative Rule Learning (IRL) [14]. IRL represents one rule in a chromosome. One rule is used from the final population of a GA. More rules are learnt by repeating the GA and penalising previously learnt rules in the fitness function.

4. Enhanced temporal fuzzy association rule mining

The GA-based approach is extended with an enhanced fitness function. A weight in the fitness function provides a preference-based multi-objective model to overcome confidence dominating the fitness [25]. Previous approaches also use Pareto-based multi-objective models [24], however, selecting a single rule from the Pareto front (for IRL) is a challenging problem. A chromosome C has mixed types, and is defined as $C = (e_l, e_u, i_1, s_1, \alpha_1, a_1, \ldots, i_k, s_k, \alpha_k, a_k)$ where the lower temporal endpoint is e_l (start of time window), the upper temporal endpoint is e_u (end of time window), i is the uniform resource locator (URL), s is the linguistic label expressing the page view time for that URL (e.g., *medium*), α is the lateral displacement of that linguistic label, a determines the antecedent/consequent part, and k is the number of URLs in a rule. For example, a chromosome (807127200, 807130800, ''/*Rules.html*'', ''*medium*'', −0.42, *antecedent*, ''/'', ''*medium*'', 0.31, *consequent*) represents the rule "IF *view time of/Rules.html* is (*medium*, −0.42) THEN *view time of/* is (*medium*, 0.31) during the period from 807127200 to 807130800" (unixtime). A single rule is represented and extracted from a chromosome, because the lateral displacements of a fuzzy set are specific to each rule.

The fuzzy support count of a chromosome C in a single transaction t_j is defined from Hong et al. [18] as

$$\mathrm{FuzSupTran}(C^{(t_j)}) = \min_{n=1}^{k} \mu_{(s_n,\alpha_n)}\left(t_j^{(i_n)}\right), \tag{1}$$