Applied Soft Computing 11 (2011) 93–102

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Multiple Instance Learning with Multiple Objective Genetic Programming for Web Mining

Amelia Zafra*, Eva L. Gibaja, Sebastián Ventura

*Department of Computer Science and Numerical Analysis, University of Cordoba, Spain*

ARTICLE INFO

*Article history:*
Received 30 October 2008
Received in revised form 6 September 2009
Accepted 18 October 2009
Available online 24 October 2009

*Keywords:*
Multi-instance learning
Multi-objective learning
Genetic programming
Web Mining

ABSTRACT

This paper introduces a multi-objective grammar based genetic programming algorithm, MOG3P-MI, to solve a Web Mining problem from the perspective of multiple instance learning. This algorithm is evaluated and compared to other algorithms that were previously used to solve this problem. Computational experiments show that the MOG3P-MI algorithm obtains the best results, adds comprehensibility and clarity to the knowledge discovery process and overcomes the main drawbacks of previous techniques obtaining solutions which maintain a balance between conflicting measurements like sensitivity and specificity.

## 1. Introduction

In recent years there has been a growing interest in applying web-usage mining techniques to build web recommender systems (RSs) [1]. These systems use information filtering technology to anticipate the needs of web users and provide recommendations that predict whether a particular user will like a specific item, or identify a set of items that will be of interest to a certain user. Most of these systems involve analyzing the behaviour of users, identifying patterns of user behaviour, and predicting their subsequent behaviour or interests. Recommender systems have been used in a number of different applications, such as recommending such products as movies [2], news [3], or vacations [4], identifying web-pages that will be of interest, or suggesting alternate ways of searching for information. In this paper, we focus our attention on the application of web index page recommendation from a multi-instance perspective [5]. Web index pages are pages that contain references or brief summaries of other pages. The goal is to identify whether a new web index page will interest a user or not by analyzing the web index pages that the user has browsed. The difficulty entailed in this learning lies in the fact that the information available about the user is whether or not he or she is interested in an index page, instead of specifying the concrete links that this person is really interested in.

This problem has been resolved from a traditional perspective with several techniques such as *k*-nearest neighbour [6] and inverse document frequencies [7], as well as from a multiple instance perspective adapting a *k*-nearest neighbour algorithm [5] and a grammar guided genetic programming algorithm [8,9]. The main characteristic which is shared by all the methods previously described in the literature is that a single objective function is used which combines different measurements, such as accuracy, recall and precision by means of using linear combinations of them or adding constraints. The problem emerges because these measurements often conflict with each other (if we optimize one of them, the values of the others decrease) and it is very difficult to obtain a trade-off between them. So the user profiles generated are not of high quality, because by optimizing one measurement without taking the decrease of the others into account, we end up by recommending pages to users that contain no pertinent topics of interest for this person or vice versa. The relevance of the recommendations depends to a great extent on the adequate optimization of these contradictory factors, because it is just as important to recommend everything that will be of interest to the user as it is to reject whatever holds no interest whatsoever. That is why the generation of user models is a task which is well suited to a multi-objective (MO) metaheuristic approach that seeks all optimal solutions and it allows us to select a solution which reconciles the two measurements.

In this study, we propose a novel multi-objective grammar guided genetic programming algorithm [10], MOG3P-MI, to recommend any particular item that could interest the user by using

* Corresponding author at: Department of Computer Science and Numerical Analysis, University of Cordoba, Campus Universitario Rabanales, Edificio Marie Curie, Planta Baja, 14071 Cordoba, Spain. Tel.: +34 957212172; fax: +34 957218630.
*E-mail addresses:* azafra@uco.es (A. Zafra), egibaja@uco.es (E.L. Gibaja), sventura@uco.es (S. Ventura).

1568-4946/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2009.10.021

the optimization of two objectives, sensitivity and specificity. There are two main motivations for introducing multi-objective genetic programming into this field. First of all, grammar guided genetic programming (G3P) is considered to be a robust tool for classification in noisy and complex domains that overcomes the drawbacks of *k*-nearest neighbour (k-NN) algorithms. Although k-NN algorithms have been used extensively and have achieved an important acknowledgment in this area, these algorithms get harder to scale for a large number of items, while still maintaining reasonable prediction performance and accuracy. This is because they require computations that grow linearly with the number of items. Moreover the discovered knowledge is not easily understandable, they give no information about the user preferences. In this regard, G3P not only obtains competitive results, but also adds comprehensibility and clarity to the knowledge discovery process, expressing the information in the form of IF–THEN prediction (classification) rules. The second reason to introduce our proposal is because genetic programming with multi-objective strategy allows us to obtain a set of optimal solutions that represent a trade-off between different rule quality measurements, where none can be considered to be better than any other with respect to all objective functions. Therefore, from this set of optimal solutions we could introduce preference information to select that solution which offers the best classification guarantee with respect to new data sets.

Experimental results for solving this problem show that this approach obtains the best results in terms of accuracy, sensitivity and specificity, demonstrating the effectiveness and applicability of the proposed method. MOG3P-MI allows to discover user preferences and generates a simple rule based model that increases generalization ability and includes interpretability and clarity in the discovered knowledge by providing information about the user's interests and classifying new examples quickly.

The rest of this paper is organized as follows. Section 2 introduces the multiple instance learning paradigm. Section 3 comments on the web index pages recommendation problem. Section 4 describes the MOG3P-MI proposed algorithm. Section 5 reports on experimental results. Finally, Section 6 presents the conclusions and future research.

## 2. Multiple instance learning

Multiple instance learning (MIL) introduced by Dietterich et al. [11] is a learning framework where the labels of individual objects in the training data, called instances, are not available; instead, the labelled unit is a set of instances called a bag. In other words, in multiple instance learning, a training example is a labelled bag and the labels of the instances are unknown. The goal of learning is to obtain a hypothesis from the training examples that generates labels for unseen bags. In this sense, the multiple instance learning problem can be regarded as a special kind of supervised machine learning problem where the labeling information is incomplete. There are several hypotheses for solving this problem. The most widely used one was proposed by Dietterich et al. [11]. This hypothesis is known as *standard assumption* and sets that a bag is labelled positive if and only if the bag has one or more positive instances, and is labelled negative if and only if all its instances are negative. More recently, we can find other hypothesis categorized as *generalized assumptions*. These assumptions determine that a bag is qualified to be positive if instances in the bag satisfy some sophisticated constraints other than simply having at least one positive instance. A proposal made by Weidmann et al. [12] defined three kinds of generalized multi-instance problems, based on employing different assumptions of how the classifications of instances determine their bag label. Independently, Scott et al. [13] defined another generalized multi-instance learning model in which a bag label is not based on a single instance proximity to a single target point. Rather, a bag is positive if and only if it contains a collection of instances, each near one of a set of target points.

During the last decade, this recent learning framework has found an interested audience in the machine learning community and its research works are drawing wide-spread attention for two basic reasons. Firstly, there have been numerous real-world applications which have found in MIL a natural way of being represented. Among these tasks we can cite text categorization [14], content-based image retrieval [15–17], drug activity prediction [18,19], image annotation [20–22], web index page recommendation [5], stock selection [18], landmark matching [23], computer security [24] and subgoal discovery [25].

The second reason is due to the great amount of new methods of multi-instance learning that have been designed. If we go through the literature, we can find specifically developed algorithms for solving MIL problems, such as APR algorithms [11], Diverse Density (DD) [18], EM-DD [19] and more recently the proposal of Pao et al. [26]. On other hand, we can find algorithms that are adaptations of popular machine learning paradigms, such as, multi-instance lazy learning algorithms [27], multi-instance tree learners [24,28,29], multi-instance rule inducers [28], multi-instance logistic regression methods [30], multi-instance neural networks [31–34], multi-instance kernel methods [14,35–39], and multi-instance ensembles [17,32]. We can see that almost all popular machine learning algorithms have been applied to solve multiple instance problems, but it is remarkable that the first proposals to adapt Evolutionary Algorithms (EAs) to this scenario did not appeared until 2007 [8,10] even though these algorithms have been applied successfully in many problems of supervised learning.

## 3. Web index pages recommendation problem

The recommender system (RS) [1] tries to anticipate the needs of web users and provide them with recommendations for products they might appreciate, taking into account their past rating profile and history of purchase or interest. Due to the rapid growth in the amount of information on the Internet and the increasing need of finding more exactly needed information, these systems have become essential tools to assist users.

The recommendation problem, in its most common formulation, is reduced to the problem of estimating ratings for the items that have not been seen previously seen by a user. This estimation could be used to recommend to the user those items with the highest estimated ratings. The problem discussed in this study, web index pages recommendation, is classified as a recommendation problem, where web index pages are recommended to a user according to his/her preferences. These pages provide titles or brief summaries of other pages. They contain plentiful information regarding references, leaving the detailed presentation to their linked pages. An example of a web index page is the news entry of CNN (http://www.cnn.com) which is shown in Fig. 1. In this figure the web index page is represented in the middle and the content of four of its linked pages is shown in sides of the figure. We can see how a web index page contains a lot of general information. If the user is interested in any particular topic, he/she has to click on the link to reach its content.

Many web index pages can be found on Internet. These pages, like most of the information on the Internet, are largely unstructured, with pages about on diverse domains of topics. Some of these pages may contain issues that interest the web user while some may not. It is very difficult for a web user to find and select only those pages containing interesting topics from among so much information. Therefore, it would be interesting to be able to analyze these pages automatically in order to show the user only those