Information Sciences 180 (2010) 4496–4513

Contents lists available at ScienceDirect

Information Sciences

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

G3P-MI: A genetic programming algorithm for multiple instance learning

Amelia Zafra, Sebastián Ventura *

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

ARTICLE INFO

Article history:
Received 27 January 2009
Received in revised form 18 July 2010
Accepted 27 July 2010

Keywords:
Multiple instance learning
Grammar-Guided Genetic Programming
Rule learning

ABSTRACT

This paper introduces a new Grammar-Guided Genetic Programming algorithm for resolving multi-instance learning problems. This algorithm, called G3P-MI, is evaluated and compared to other multi-instance classification techniques in different application domains. Computational experiments show that the G3P-MI often obtains consistently better results than other algorithms in terms of accuracy, sensitivity and specificity. Moreover, it makes the knowledge discovery process clearer and more comprehensible, by expressing information in the form of IF-THEN rules. Our results confirm that evolutionary algorithms are very appropriate for dealing with multi-instance learning problems.

1. Introduction

Multiple instance learning (MIL), introduced by Dietterich et al. [21], is a generalization of traditional supervised learning. In MIL, training patterns called bags are represented as a set of feature vectors called instances. Each bag contains a number of non-repeated instances and each instance usually represents a different view of the training pattern attached to it. There is information about the bags and each one receives a special label, although the labels of instances are unknown. The problem consists of generating a classifier that will correctly classify unseen bags of instances. The key challenge in MIL is to cope with the ambiguity of not knowing which instances in a positive bag are actually positive examples, and which ones are not. In this sense, a multiple instance learning problem can be regarded as a special kind of supervised learning problem with incomplete labeling information.

This learning framework is receiving growing attention in the machine learning community because numerous real-world tasks can be represented very naturally as multiple instance problems. These tasks include text categorization [3], content-based image retrieval [68,43,70], image annotation [67,24], drug activity prediction [40,80], web index page recommendation [77,69] and semantic video retrieval [13]. In order to resolve these problems, the literature abounds in a great number of method proposals that range from algorithms designed specifically to solve multi-instance problems to extensions of classical algorithms adapted to the MIL scenario. The latter include *k*-nearest neighbors, decision trees, rule based systems, support vector machines, neural networks, inductive logic programming and ensembles.

This paper presents a framework for modifying and extending evolutionary algorithms (EAs) to deal with MIL problems. One extension of these algorithms, based on Grammar-Guided Genetic Programming (G3P), is designed and implemented; the resulting algorithm is called G3P-MI. It is evaluated, analyzed and compared to other MIL classification techniques in a collection of multi-instance datasets. Experimental results show that G3P-MI often obtains consistently better results than other algorithms in several applications. Moreover, it adds comprehensibility and clarity to the knowledge discovery process,

* Corresponding author. Address: Department of Computer Science and Numerical Analysis, University of Cordoba, Campus Universitario Rabanales, Edificio Einstein, Tercera Planta, 14071 Cordoba, Spain. Tel.: +34 957212218; fax: +34 957218630.
E-mail addresses: azafra@uco.es (A. Zafra), sventura@uco.es (S. Ventura).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.07.031

expressing information in the form of IF-THEN prediction (classification) rules. These results show that EAs are a promising tool for solving MIL problems.

In summary, the key points of this paper are: to introduce evolutionary algorithms to solve MIL; to study the effectiveness of our algorithm (G3P-MI) in obtaining classification accuracy and a high degree of discovered knowledge (comprehensible rules); to present an empirical comparison between different data sets using different algorithms; and to provide data sets to facilitate future comparisons in this area. The last two points are very relevant to forward real progress in MIL because they allow different methods to be compared under the same conditions. Currently, this is the greatest weakness found in studies on MIL because, although a wide range of problem domains have been researched using a broad spectrum of approaches, extensive comparisons of algorithms are infrequent; most studies empirically compare only a few approaches and use only a few data sets to do so (often only one).

The paper is structured as follows. Section 2 briefly reviews advances in the area of multiple instance learning. Section 3 describes the proposed algorithm. Section 4 evaluates and compares our algorithm to other techniques implemented in the WEKA tool in ten datasets. Finally, Section 5 draws some conclusions and raises several issues for future work.

2. Antecedents

This section gives a definition and a notation of MIL and reviews the most important developments in MIL in recent years.

2.1. Definition and notation of multiple instance learning

MIL is designed to solve the same problems as single-instance learning: learning a concept that correctly classifies training data as well generalizing unseen data. Although the actual learning process is quite similar, the two approaches differ in the class labels provided which are what they learn from. In a traditional machine learning setting, an object m_i is represented by a feature vector v_i, which is associated with a label $f(m_i)$. However, in the multiple instance setting, each object m_i may have V_i various instances denoted $m_{i,1}, m_{i,2}, \ldots, m_{i,v_i}$. Each of these variants will be represented by a (usually) distinct feature vector $V(m_{i,j})$. A complete training example is therefore written as $(\{V(m_{i,1}), V(m_{i,2}), \ldots, V(m_{i,v_i})\}, f(m_i))$.

The goal of learning is to find a good approximation to the function $f(m_i)$, $\hat{f}(m_i)$, analyzing a set of training examples and labeled as $f(m_i)$. To obtain this function Dietterich defines a hypothesis that assumes that if the result observed is *positive*, then at least one of the variant instances must have produced that positive result. Furthermore, if the result observed is *negative*, then none of the variant instances could have produced a positive result. This can be modelled by introducing a second function $g(V(m_{i,j}))$ that takes a single variant instance and produces a result. The externally observed result, $f(m_i)$, can then be defined as follows:

$$f(m_i) = \begin{cases} 1 & \text{if } \exists\, j | g(V(m_{i,j})) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

In the early years of multi-instance learning research, all multi-instance classification work was based on this assumption, that is, that MIL formulations explicitly or implicitly encoded the assumption that a bag was positive if and only if at least one of its instances was positive. This formulation is known as the standard multi-instance hypothesis or Dietterich hypothesis. More recently, generalized multi-instance learning models have been formalized where 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. Firstly, Weidmann et al. [59] defined three kinds of generalized multi-instance problems, based on employing different assumptions of how the classification of instances classifications determine their bag label. These definitions are *presence-based MI*, *threshold-based MI*, and *count-based MI*. *Presence-based MI* is defined in terms of the presence of at least one instance of each concept in a bag (note that the standard hypothesis is a special case of this assumption which considers just one underlying concept); *threshold-based MI* requires a certain number of instances of each concept to be present simultaneously and *count-based MI* requires a maximum as well as a minimum number of instances of a certain concept in a bag. Independently, Scott et al. [51] defined another generalized multi-instance learning model in which a bag label is not based on the proximity of one single instance to one 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.

Similarly, there has also been an extension of the definition of multi-instance problems which use outputs with discrete values (in particular, binary). The first proposal to use real-valued labels was introduced by Amar et al. [2]. These problems, known as multi-instance regression (MIR) problems, have drawn wide-spread attention. Ray and Page [48] modified regression for MIL by assuming that if an algorithm could identify the best instance in each bag, then straightforward regression could be used to learn a labeling method for each instance (and then for each bag); Ramon and De Raedt [46] adapted internal labeling methods for neural networks to learn a function that used real-valued instances; Dooly et al. [22] presented extensions of *k*-nearest neighbors (*k*-NN), Citation-kNN, and the diverse-density algorithm for the real-valued setting and studied their performance on boolean and real-valued data.