Expert Systems with Applications 40 (2013) 5924–5931

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Multi-label classification based on analog reasoning

Ruben Nicolas *, Andreu Sancho-Asensio, Elisabet Golobardes, Albert Fornells, Albert Orriols-Puig

Grup de Recerca en Sistemes Intel.ligents, La Salle – Universitat Ramon Llull, Quatre Camins 2, 08022 Barcelona, Spain

ARTICLE INFO

Keywords:
Multi-label
Classification
Case-Based Reasoning

ABSTRACT

Some of the real-world problems are represented with just one label but many of today's issues are currently being defined with multiple labels. This second group is important because multi-label classes provide a more global picture of the problem. From the study of the characteristics of the most influential systems in this area, MlKnn and RAkEL, we can observe that the main drawback of these specific systems is the time required. Therefore, the aim of the current paper is to develop a more efficient system in terms of computation without incurring accuracy loss. To meet this objective we propose MlCBR, a system for multi-label classification based on Case-Based Reasoning. The results obtained highlight the strong performance of our algorithm in comparison with previous benchmark methods in terms of accuracy rates and computational time reduction.

1. Introduction

Recent progress in machine learning and data mining has led to the application of their techniques in more complex multi-label problems, such as forecasting, where we have data with different features obtained from several stations that could be used in order to predict just one class, e.g., rain probability, but other relevant classes could be analyzed together to provide a more global picture of the forecast. These labels could be temperature, humidity, wind, and so on. This difficulty is also emphasized in other categorization problems such as medicine, emotions, texts, biology, or face verification, among others because they are complex issues that could be analyzed from more than one point of view (Tsoumakas, Katakis, & Vlahavas, 2008). There are two ways to tackle this problem (Tsoumakas & Katakis, 2007): (1) to transform the dataset to single-label and use classical classification algorithms or (2) to modify these classical algorithms to accept multi-label data. In our case we have worked on the second given that the first family is somehow a step backwards towards the single-label classification because these systems lose the possibility of analyzing the problem from different points of view. Within this second group there are several contributions among which MlKnn (Zhang & Zhou, 2005, 2007) and RAkEL (Tsoumakas et al., 2007) are the most noteworthy. These two concrete proposals face the problem effectively in terms of accuracy but they are not efficient time-wise. This paper tackles the difficulty of reducing the computational cost of classification from multi-label data without losing the precision achieved with previous methods. This is really important from the standpoint that nowadays problems can be represented with datasets that are not only rich in labels but also in the number of cases. The increment of instances correlates a direct increase in computational time.

To achieve a reduction in time costs without penalizing the accuracy we propose a Case-Based Reasoning (CBR) (Aamodt & Plaza, 1994) system for multi-label classification based on MlKnn fundamentals. The choice of CBR as the core of our algorithm is based on its main skills: (1) good adaptation to multi-label characteristics; (2) low complexity being a competent method; (3) explicative capability of CBR that is extremely important in problems such as medical prognosis; (4) existences of an active CBR community that is interested in the adaptation to multi-label problems (Brinker & Hüllermeier, 2007); (5) non-existence of an approach to this goal using CBR despite the interest of the researchers in this area. In addition if we consider CBR as an improvement on Knn systems, we should also consider it as an effective approach to enhancing some of the characteristics of MlKnn. Our adaptation of CBR algorithm to multi-label problems has been focused on the retrieval and reuse stages. Results of our proposal are compared with other two competitive multi-label learning systems, MlKnn and RAkEL, using seven synthetic dataset and three other real-world datasets used as benchmark by multi-label classification community (Ávila, Gibaja, & Ventura, 2009). The algorithms are compared with Friedman, Holm and Shaffer statistical tests.

The remainder of this paper is organized as follows: Section 2 summarizes the background information and the related work; Section 3 presents the contribution for multi-label classification; Section 4 describes the experimentation and discusses the results; and finally, Section 5 ends with the conclusions and further work.

* Corresponding author. Tel.: +34 932902451.
E-mail address: rnicolas@salle.url.edu (R. Nicolas).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.05.004

2. Related work

This paper tackles the difficulty of reducing the computational cost of classifying using multi-label data without losing accuracy. Currently the work in multi-label problem solving is divided into two different families. On the one hand, Problem Transformation Methods (PTM) transform the learning task into one or more single-label classification tasks. The main problem of this family is that with the unification of different labels into a single one we may lose information that could be critical in cases such as medical prognosis. In contrast, the positive aspect is the possibility of using existing algorithms without having to modify them. On the other hand, Algorithm Adaptation Methods (AAM) deal with the problem of modifying classical algorithms to work in a multi-label mode. Despite the fact that this second family of methods focuses on not losing information, the adaptation is not trivial and could increase the calculations and, consequently, the time consume.

The most influential works from the AAM family include (Schapire, 2000) which presents the BoosTexter a system that uses boosting algorithms for text-categorization. This platform, designed for automatic call-type identification makes classes for further classification; (Clare & King, 2001) that deal with multi-label biological data and adapt the entropy analysis in order to use classical C4.5 (Quinlan, 1993) algorithm to create a decision tree; and (Elisseeff & Weston, 2001) that focuses its attention on an approach based on a ranking method combined with a predictor of the size of the sets which tries to overcome the difficulties found by previous works when adapting multi-label problems to two classes ones.

The most competent works in PTM for multi-label classification are MlKnn and RAkEL. These two works are recognized by the community as reference algorithms. The first one, MlKnn, is a theoretical approach to multi-label classification that adapts the combination of the *k* recovered cases of classical k nearest-neighbor algorithm (Knn) (Han & Kamber, 2006) to multiple label problems. RAkEL is an ensemble platform that allows the classification of multi-label datasets by dealing with each label separately and combining the single-label results. It can be used with several algorithms as a single-label classifier system (as it is implemented using WEKA (Hall et al., 2009) libraries all its classifiers can be used) but the one used as a common benchmark is C4.5. Both algorithms are publicly available with a standard configuration under the name of MULAN. Although MlKnn and RAkEL are the most competent and commonly used platforms there are other interesting works in this field used by the community such as (Read, 2008) where the authors present a pruned transformation that combines key-points of several previous approaches and (Zhang & Zhou, 2006) that uses neural networks for multi-label classification.

In reference to the characteristics of previous works in multi-label classifications and its shortcomings, we have developed our proposals based on AAM because, attending to the literature, this family reaches better results than PTM. These are described in the following section.

3. Multi-label Case-Based Reasoning Algorithm

Current multi-label classification methods in the AAM family provide competent accuracy results but show high complexity in terms of computation. These systems propose a complex algorithm with a high level of calculus that increases the computational time. Our proposal obtains a system which is as accurate as previous ones but which employs less calculus and is, therefore, less complex. This platform has been named Multi-label Case-Based Reasoning (MlCBR). The most similar work that addresses the use of algorithms with small number of calculations for multi-label classification is MlKnn. This work proposes the adaptation of Knn algorithm to multi-label classification. The changes suggested by Zhang and Zhou (2005) to transform the single-label algorithm into a multi-label approach are the addition of some mathematical calculations after recovering the *k* most similar cases of the case memory. In our case, unlike MlKnn we adapted to multi-label classification by employing CBR method, which is a technique that solves new cases by using others previously solved. In order to achieve this objective, four phases are applied: (1) first of all, the system retrieves the most similar cases from the case memory with the assistance of a similarity function; (2) secondly, it tries to reuse the solutions from the retrieved cases with the aim to solve the present case, (3) then it revises the solution, and (4) finally it retains the useful information of the solved case, if necessary. All the steps are centered on the case memory, which contains the experience of system in terms of cases. A case is an instance of a problem. We have chosen this algorithm because to a certain extent it is an improvement on Knn by the addition of the retaining, revising and reuse phases to the simple retrieval of the other option. Furthermore, the competence of this kind of algorithm is visible in problems related to medicine, semantic web or general purpose classification. The main advantages of CBR that make it perfect for a multi-label transformation are its accredited results of good performance and low complexity, its explicative capacity and the fact that it has an active community working on it which is interested in this specific kind of problems. In this paper we centered our effort on the retrieve and reuse stages of CBR because these are the features that will enable us to meet our objectives, namely the reduction of computational time and maintaining or improving the accuracy. The retrieve stage of Multi-label Case-Based Reasoning Algorithm (MlCBR) algorithm is based on MlKnn where the *k* most similar cases to the case study are recovered of the case memory. As reuse phase two approaches are proposed. Probabilistic Reuse (PR) is the first option where the final classification is made through a voting process which all the recovered cases are equally weighted. In contrast, Probabilistic Reuse based on Experience (PRE) adds the concept of experience to better weight the recovered cases. In the following subsections we detail the algorithms proposed for reuse stage on multi-label classification using CBR: PR and PRE.

3.1. First step: Probabilistic Reuse

Probabilistic Reuse algorithm present probabilistic variations in the classical reuse stage in order to adapt CBR to multi-label classification. Once the system recovers the *k* best cases, they are mixed in order to propose a solution. We consider a voting combination of cases similar to the one proposed by MlKnn but adapted to the CBR idea. MlKnn, in the same way as other single-label Knn algorithms, recovers the *k* best cases of the previously recovered ones and gives a classification result combining the *k* cases. This combination is done through counting the number of recovered instances that predict each label. The platform considers that a label will be set to one if more than a half of the *k* recovered cases have this label with a positive value. In the case of MlCBR with PR reuse, after we recover the *k* best cases in the retrieval stage we combine it in reuse. This reuse step considers the frequency for each label and sets it to positive if the percentage is more than 50%. The mathematical process followed by MlKnn and PR reuse to combine the *k* cases obtained is the same in terms of the final result. The