Applied Soft Computing 34 (2015) 705–720

Contents lists available at ScienceDirect

Applied Soft Computing

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

Conception of a dominance-based multi-objective local search in the context of classification rule mining in large and imbalanced data sets

Julie Jacques[a,b,c,*], Julien Taillard[c], David Delerue[c], Clarisse Dhaenens[a,b], Laetitia Jourdan[a,b]

[a] LIFL, Université Lille 1, Bât. M3, 59655 Villeneuve d'Ascq cedex, France
[b] INRIA Lille Nord Europe, 40 Av. Halley, 59650 Villeneuve d'Ascq, France
[c] Société ALICANTE, 50 Rue Philippe de Girard, 59113 Seclin, France

ARTICLE INFO

Article history:
Received 4 July 2013
Received in revised form 2 June 2015
Accepted 2 June 2015
Available online 10 June 2015

Keywords:
Partial classification
Imbalanced data
Multi-objective
Local search

ABSTRACT

Classification on medical data raises several problems such as class imbalance, double meaning of missing data, volumetry or need of highly interpretable results. In this paper a new algorithm is proposed: MOCA-I (Multi-Objective Classification Algorithm for Imbalanced data), a multi-objective local search algorithm that is conceived to deal with these issues all together. It is based on a new modelization as a Pittsburgh multi-objective partial classification rule mining problem, which is described in the first part of this paper. An existing dominance-based multi-objective local search (DMLS) is modified to deal with this modelization. After experimentally tuning the parameters of MOCA-I and determining which version of DMLS algorithm is the most effective, the obtained MOCA-I version is compared to several state-of-the-art classification algorithms. This comparison is realized on 10 small and middle-sized data sets of literature and 2 real data sets; MOCA-I obtains the best results on the 10 data sets and is statistically better than other approaches on the real data sets.

1. Motivations

Classification on real data sets raises several challenges, especially when dealing with medical data sets. One common issue is class imbalance, where the class to predict is underrepresented among the observations of the data set. Not so uncommon repartitions are 100:1 or even 10,000:1. As an example in hospital data, stroke – a frequent disease – will concern at best 1% of the hospital stays. Most classification algorithms build their classifiers using *Accuracy*, which measures the percentage of well-classified observations. However, this is ineffective with imbalance data: in the stroke example, a dummy classifier labeling each stay as "no stroke" will have a 99% classification *Accuracy*, while being totally useless to predict stroke. Some approaches have been proposed to overcome this problem, as detailed in the review of He et al. [1] and will be more developed further in this paper.

Another challenge comes from the absence of real negation in medical files, which brings uncertainty. The absence of some information in the patient medical file has a double meaning. In most cases when information about a disease is missing, it means the patient does not suffer from the disease. In other cases, the patient may have the disease but is not diagnosed yet, or this information has not been entered in the system. Moreover, in some medical coding, such as Anatomical Therapeutic Chemical (ATC) Classification System,[1] a same information can have several encodings, depending on the context: a same procedure or diagnose may be coded differently depending on the healthcare professional. In the presence of these "yes"/"no∨unknown" binary values, it is not reliable to predict patients having class="no∨unknown". This kind of problem is particularly indicated to partial classification, which focuses only on predicting a subset of the population, for example only the patients having class="yes". The amount of hospital data available raises another challenge. More than 50,000 medical procedures and diseases can be entered in patient data through ICD-10 encoding, which is available in all French hospitals. Since a classifier is a combination of tests on patient information, classification can be seen as a combinatorial problem. Operational research and meta-heuristics are indicated to solve this kind of problems. In their review, Corne et al. explain how these techniques can be applied to data mining [2]. In this context, many multi-objective optimization algorithms

* Corresponding author at: Société ALICANTE, 50 Rue Philippe de Girard, 59113 Seclin, France. Tel. +33 328559250.

[1] Available on the World Health Organisation's website: http://www.whocc.no/atc_ddd_index/

http://dx.doi.org/10.1016/j.asoc.2015.06.002
1568-4946/

have been proposed for rule mining, most of them are detailed in the review of Srinivasan and Ramkrishnan [3].

Another issue is the simplicity to understand and use the algorithm. Indeed, the predictions will be a part of a medical decision aiding tool, the OPCYCLIN project – an industrial project dedicated to decision aid for clinical trials, involving Alicante Company, hospitals and academics as partners. Thus, the generated classifiers must provide a good interpretability, allowing the users assessing the validity of the predictions. Besides, the main features of the data sets depend on the hospital under study [4] or on the patient information to predict: depending on the disease, class imbalance represents from 1% to 20% of the stays or patients under study. This requires setting the parameters depending on the data set under study. However, users often do not have sufficient knowledge in data mining to parameter the algorithms [5]. Robust approaches able to give good results on most data sets will be preferred.

Many recent contributions have been proposed to deal with some of these issues, for example [6–9]. As far as we know, an approach overcoming all these issues at the same time had not been proposed yet. In this paper a new algorithm is introduced: MOCA-I (Multi-Objective Classification Algorithm for Imbalanced data), which is an optimization algorithm able to generate partial classification rules in large and imbalanced data sets. This paper aims to find the better parameters to use with MOCA-I and then will compare it to algorithms of literature. Section 2 describes more deeply the partial classification rule mining problem. Then Section 3 presents MOCA-I – our implementation as a multi-objective problem – while first explaining notions about multi-objective algorithms. Section 4 contains a deep study of MOCA-I algorithm, assessing the best parameters to use, such as size of rules, archive size, etc. to ensure the best results over most data sets. Section 5 compares the results to those obtained by state-of-the-art algorithms, both on benchmark data sets from literature and real data sets. Finally, Section 6 gives the conclusion and perspectives.

2. Context

This section first describes the partial classification problem and how to evaluate the performance of a classifier. Then it describes the problems risen by our real data: class imbalance and high volume of data.

2.1. Partial classification

The classification task aims to predict a fact – called a class, for example "flu?yes/no" – on unknown observations. Observations depend on the domain of application and can be of various forms like bills, patients, events, etc. For each observation several information are available, which are called attributes. In the flu example, each observation is a patient; attributes are a list of symptoms that were observed; and each patient may or may not have presented each symptom. The classification task will generate a classifier that describes how to determine the class – here "flu" – by using the attributes – here the symptoms. A classifier is a combination of attributes tests (AT). Each represents a test on an attribute, for example "age>25". Partial classification is a subclass of classification, which interests only in predicting observations matching a subset of the class: observations not matching the subset class are not predicted. An example of partial classification task could be to predict flu = yes on unknown patients, while having no interest to find healthy patients. Several kinds of classifiers can be extracted to predict the class. The most common are trees and rules, which consist of conjunctions or disjunctions of attributes. As an example: *cough = yes* and *fever = yes* and *musclepain = yes* $\Rightarrow$ *flu*. Less interpretable classifiers exist such as support vector machines or neural networks but they are not the object of this paper.

Table 1
Confusion matrix.

	P	$\overline{P}$	
C	TP	FP	
$\overline{C}$	FN	TN	
			N

2.2. Evaluation of the performance of a classifier

More than 40 metrics have been proposed in the literature to assess the efficiency of these classifiers [10,11]. Most of them are based on the confusion matrix given in Table 1. Given a rule classifier of the form $C \Rightarrow P$, it counts the number of observations well classified – true positives (TP) and true negatives (TN) – as well as the wrongly classified ones – false negatives (FN) and false positives (FP). As a rule of thumb, classifiers are often evaluated on both known and unknown data, to assess the capacity of the classifier to deal with unknown data. In order to do so, the data are split as training and test data sets. The training data set is used to build the classifier, while the test data set is used to evaluate it.

2.3. Impact of class imbalance and volume

Previously it has been mentioned that medical data can bring several problems. One of them is uncertainty: the absence of information in the patient file can have several meanings. Thus, $\bar{P}$ observations cannot be completely relied upon: a small part of them (up to 15%) may be in fact P observations. The use of metrics adapted to partial classification – such as *F-measure* [1] – allows dealing with such data by not focusing too much on $\bar{P}$.

Another major problem is imbalance data. When dealing with imbalance data, $|P| << |\bar{P}|$: positive observations are less available than the other observations. Metrics based on counting the number of well classified observations (both P and $\bar{P}$), such as *Accuracy* or the number of wrongly classified observations will tend to encourage the classification of $\bar{P}$ observations. This is especially the case with high imbalance (class to predict is less than 1% of observations), where the cost on a "misclassified" observation will be minimal. Several solutions have been proposed to solve this problem. Most of them are detailed in the review of He et al. [1]. Cost sensitive methods set weights on P and $\bar{P}$ to force metrics to deal with the class to predict, while Boosting methods set weights directly on observations and work on several iterations. Each iteration, weights on the misclassified observations are increased, to force the classifier to deal better with them. These weight-based approaches are often used to enhance the results of some basic classifiers such as the famous *C4.5* algorithm [12]. Hence, Ting et al. proposed a cost-sensitive version of *C4.5* named *C4.5-CS* [13], while *AdaC2* and *DataBoost-IM* use it within a boosting algorithm [14,15]. However, the weights used in these approaches can be hard to set. The object of this paper is to build a classifier system able to deal natively with class imbalance, to avoid dealing with more parameters due to weighting. Thus a metric well-adapted to deal with imbalance will be chosen instead of dealing with weights. However boosting or cost-sensitive methods can probably improve the results and may be the object of further works. Other methods focus on resampling the data to obtained well-balanced data, adding new observations (over-sampling), for example by generated new observations like the SMOTE method [16] or by removing observations (under-sampling) like the ACOSampling method [17]. Since