European Journal of Operational Research 250 (2016) 389–399

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Discrete Optimization

Integer programming models for feature selection: New extensions and a randomized solution algorithm

P. Bertolazzi [a], G. Felici [a,*], P. Festa [b], G. Fiscon [a,c], E. Weitschek [a,d]

[a] Institute of Systems Analysis and Computer Science, National Research Council, Via dei Taurini 19, Rome, 00185 Italy
[b] Department of Mathematics and Applications "R. Caccioppoli", University of Napoli Federico II, Compl. MSA, Via Cintia, Napoli, 80126 Italy
[c] Department of Computer, Control and Management Engineering, Sapienza University, Via Ariosto, 25, Rome, 00185 Italy
[d] Department of Engineering, Uninettuno International University, Corso Vittorio Emanuele II, 39, Rome, 00186 Italy

ARTICLE INFO

Article history:
Received 25 April 2014
Accepted 24 September 2015
Available online 9 October 2015

Keywords:
Data mining
Heuristics
Integer programming

ABSTRACT

Feature selection methods are used in machine learning and data analysis to select a subset of features that may be successfully used in the construction of a model for the data. These methods are applied under the assumption that often many of the available features are redundant for the purpose of the analysis. In this paper, we focus on a particular method for feature selection in supervised learning problems, based on a linear programming model with integer variables. For the solution of the optimization problem associated with this approach, we propose a novel robust metaheuristics algorithm that relies on a Greedy Randomized Adaptive Search Procedure, extended with the adoption of short memory and a local search strategy. The performances of our heuristic algorithm are successfully compared with those of well-established feature selection methods, both on simulated and real data from biological applications. The obtained results suggest that our method is particularly suited for problems with a very large number of binary or categorical features.

© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

1. Introduction

Feature Selection (FS) addresses a class of methods used to extract relevant information from data. FS has always been a central topic in Multivariate Statistics and Data Analysis, but has received important contributions also from mathematicians and computer scientists; at the same time, the ever increasing amount of data that are being collected in many real world applications, jointly with the evolution of technology, poses new challenges for FS methods. An example of such challenges can be found in the study of biological and genomic data, where interesting data sets may be composed of a few hundreds of tissue samples on which the activity of tens of thousands of genes is measured. The analysis of such data requires to identify a limited number of genes (i.e., features) able to identify an interesting model. Similarly, new data collection techniques based on cheap sensors and on internet activity are creating very large repositories where precious information may be hidden and needs to be mined out.

In the general setting, FS can be described as follows: given a data matrix defined by a finite set of features measured on a finite number of objects, select a subset of the feature set that is particularly relevant, with respect to all the other possible subsets, for the analysis that is to be conducted on the data under study. In this work, we focus on *supervised learning* (i.e., classification), where data are analysed to identify a model able to predict if an observation belongs to one of two or more classes, based on the values of its features. In supervised learning, FS operates to select a relevant – and possibly small – subset of features to be used by the classifier.

We propose a method designed to treat directly integer or binary features, keeping in mind that discretization methods are often used to transform continuous measures into discrete or binary ones; such a process is adopted in many settings to control noise, to ease the interpretation of the classification model, and, last but not least, to apply specific logic-based classifiers also in the presence of continuous features.

The proposed approach is based on an optimization problem derived from the data matrix, where each feature is associated with a binary variable. Such an approach is not new in the literature; it stems from the *minimum test collection* originally described in Garey and Johnson (1979). We show that such optimization problems are still not tractable – even with state-of-the-art mixed integer solvers – and propose a new heuristic algorithm for their solution.

The performances of our method are tested on different data sets, and compared with other established FS methods, in combination

* Corresponding author. Tel. +39 0649937117; fax: +39 0649937106.
E-mail addresses: paola.bertolazzi@iasi.cnr.it (P. Bertolazzi), giovanni.felici@iasi.cnr.it (G. Felici), paola.festa@unina.it (P. Festa), fiscon@dis.uniroma1.it (G. Fiscon), emanuel.weitschek@iasi.cnr.it (E. Weitschek).

http://dx.doi.org/10.1016/j.ejor.2015.09.051
0377-2217/© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

with classifiers of different nature. The tests are run both on simulated data sets, composed of binary features, and on two real genomic data sets, composed of continuous variables. For the latter, we adopt a simple discretization procedure. The results appear to be very satisfactory both from the standpoint of solution quality and of solution time, particularly when applied to data sets of large dimension.

The paper is organized as follows. Section 2 provides a brief introduction to the different approaches to FS and the related literature. Integer programming models for FS are treated in Section 3. In Section 4 we describe the new Greedy Randomized Adaptive Search Procedure with memory proposed for the solution of the optimization problem associated with FS. The comparison of all the FS methods and their performances on real and simulated data sets are treated in Section 5 and its subsections, jointly with the description of the classifiers that we use to compare the different FS methods, and the main motivations of our experiments. Conclusions and future lines of work are drawn in Section 6.

2. Methods for feature selection

FS can be looked at from different angles. One may simply evaluate the features according to their individual merit, order them accordingly, and then select the desired number of them, possibly controlling the quality of the solution when the number of selected features increases. Such an approach is the one adopted by *Ranker* methods (Kira & Rendell, 1992a). Conversely, one may want to evaluate a subset of the features according to their integrated contribution, and thus is faced with a more complex subset selection problem, which has an intrinsic combinatorial nature and is recognized to be a complex problem. In the latter case, some methods are designed to construct a solution set by adding features iteratively, paying attention to evaluate the feature to be added conditionally to those that are already in the set; such a *forward* selection approach is paired with a *backward* approach, where features are, iteratively, eliminated from the current set.

Another way of looking at FS methods is to distinguish them according to how the feature sets are evaluated and used in data analysis. This defines *Filter*, *Wrapper*, and *Embedded* methods (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013; Forman, 2003). Methods of the first group select features according to a score function; methods of the second group iteratively test feature sets performing data analysis, until a satisfactory result is obtained; to the third group belong those methods that automatically select the features that appear to be good for the purpose of their analysis. As recognized in Bolón-Canedo et al. (2013), filters are often to be preferred for their stand-alone nature and their speed when compared to wrappers. Indeed, analysing the performances of different methods on several synthetic data sets, the authors of Bolón-Canedo et al. (2013) conclude that filter methods seem to perform better. Also in Forman (2003), where the analysis is restricted to text classification problems, filter methods stand out – in particular, the Bi-Normal Separation proposed by the authors.

FS problems of large size can be solved efficiently also with embedded methods; among the most successful ones are Support Vector Machines (SVM; Cristianini and Shawe-Taylor, 2000), where some proper modifications of the underlying optimization model can efficiently combine the choice of the separating hyperplane with the selection of good features (see, among others, Carrizosa, Martin-Barragan, & Morales, 2008; Maldonado, Pérez, Weber, & Labbé, 2014). For an additional overview of FS, the interested reader may refer to Guyon and Elisseeff (2003), John, Kohavi, and Pfleger (1994), Kira and Rendell (1992b), Liu and Motoda (2000), Liu, Li, and Wong (2002) and Swiniarski and Skowron (2003); a more specific analysis of FS methods for data mining is presented in Piramuthu (2004). As far as FS applications are concerned, a very actual battlefield is to be found in medical and bioinformatics data analysis, where supervised learning problems with very large number of features abound; here, data mining applications strongly rely on FS methods – some examples are in Dagliyan, Uney-Yuksektepe, Kavakli, and Turkay (2011), Lan and Vucetic (2013) and Peter and Somasundaram (2012).

Particularly relevant for the scope of this paper are the methods that adopt a mathematical formulation of the FS problem based on integer variables, able to exploit its combinatorial nature. The most representative and seminal work in this line of research is the minimum test collection problem, stated in Garey and Johnson (1979), based on a Set Covering formulation where binary variables are associated with the features, and a covering constraint is defined for each pair of elements that belong to different classes. In these constraints the feature variable is present only if it exhibits a different value in the two addressed elements.

Also in embedded methods, mathematical optimization is largely used. In Rubin (1990), the solution to a linear program is used to find a separating hyperplane between two sets of points; the linear program is then augmented with binary variables associated with features, resulting in a difficult problem for which several heuristics have been proposed. Similarly, in Bradley and Mangasarian (1998) linear separating hyperplanes are derived via linear programming, and then developed into the well-established theory of the already mentioned SVM (Cristianini & Shawe-Taylor, 2000). Iannarilli and Rubin (2003) adopt an optimization model, where additional packing constraints on binary variables control the dimension of the feature set, while the objective function takes care of maximizing a quality measure of the features based on the Kullback–Leiber divergence.

In this paper, we propose a method based on some variants of the minimum test collection problem, that is guaranteed to provide a separation between the classes, but does not rely on the choice of a specific classification method. A similar approach is used, among others, in Boros, Ibaraki, and Makino (1999) and in previous applications to biological and genomic data (Bertolazzi, Felici, Festa, & Lancia, 2008; Weitschek et al., 2012; Weitschek, Velzen, Felici, & Bertolazzi, 2013). Such an approach is substantially different from methods based on the search of separating hyperplanes such as Bradley and Mangasarian (1998), Carrizosa et al. (2008), Iannarilli and Rubin (2003), Maldonado et al. (2014) and Rubin (1990).

The adoption of a model where integer variables are associated with the choice of a feature sets results in computationally challenging problems, that become intractable for general purpose solvers when the dimensions of the problem increase. We thus propose a properly designed greedy randomized adaptive heuristic, usually referred to as GRASP (Feo & Resende, 1989; 1995), as a viable strategy to obtain good solutions for large FS problems that arise in supervised learning. As already mentioned above, the adoption of properly designed heuristics is frequent in FS problems: a similar GRASP approach is proposed, in a different framework, in Bermejo, Gámez, and Puerta (2011) to control the choice of the feature sets evaluated by a wrapper method; in Unler and Murat (2010) the importance of good heuristics for large sized FS problems is acknowledged, proposing a particle swarm optimization algorithm, while in Meiri and Zahavi (2006) simulated annealing is used to deal with FS problems arising in marketing applications.

According to the distinction of FS into filter, wrapper, and embedded approaches, the method proposed in this work can be considered as a filter method, and therefore the main filter FS algorithms will be taken into account for a computational assessment of the quality of the results of our method. A more detailed description of these methods – namely, Relief (Kira & Rendell, 1992a), Las Vegas Filter (LVF) (Liu & Setiono, 1996), FOCUS (Almuallim & Dietterich, 1994), Correlation-based Feature Selection (CFS) (Hall, 1999), Sequential Forward (Backward) Selection (Elimination) SFS (SBE) (Devijver & Kittler, 1982), and Information Gain (InfoGain) (Hall & Smith, 1998) – is provided in Section 5.1.

Following, we describe the integer programming models (Section 3) and the algorithm designed for their solution 4.

دریافت فوری متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات