Expert Systems With Applications 88 (2017) 205–216

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Improved multiclass feature selection via list combination

Javier Izetta, Pablo F. Verdes, Pablo M. Granitto*

CIFASIS, French Argentine International Center for Information and Systems Sciences, UNR–CONICET, Bv. 27 de Febrero 210 Bis, 2000 Rosario, Argentina

ARTICLE INFO

Article history:
Received 8 March 2017
Revised 30 June 2017
Accepted 30 June 2017
Available online 6 July 2017

Keywords:
Feature selection
Multiclass problems
Support vector machine

ABSTRACT

Feature selection is a crucial machine learning technique aimed at reducing the dimensionality of the input space. By discarding useless or redundant variables, not only it improves model performance but also facilitates its interpretability. The well-known Support Vector Machines–Recursive Feature Elimination (SVM-RFE) algorithm provides good performance with moderate computational efforts, in particular for wide datasets. When using SVM-RFE on a multiclass classification problem, the usual strategy is to decompose it into a series of binary ones, and to generate an importance statistics for each feature on each binary problem. These importances are then averaged over the set of binary problems to synthesize a single value for feature ranking. In some cases, however, this procedure can lead to poor selection. In this paper we discuss six new strategies, based on list combination, designed to yield improved selections starting from the importances given by the binary problems. We evaluate them on artificial and real-world datasets, using both One–Vs–One (OVO) and One–Vs–All (OVA) strategies. Our results suggest that the OVO decomposition is most effective for feature selection on multiclass problems. We also find that in most situations the new K-First strategy can find better subsets of features than the traditional weight average approach.

1. Introduction

Many important problems in Machine Learning, as well as in-silico Chemistry (Raies & Bajic, 2016), Biology, "high-throughput" technologies (Golub et al., 1999; Leek et al., 2010) or text processing (Forman, 2003; Uysal, 2016), share the property of involving much more features than measured samples are available (Guyon & Elisseeff, 2003). The datasets associated to these problems are, unsurprisingly, called "wide". Usually, most of these variables carry a relatively low importance for the problem at hand. Furthermore, in some cases they interfere with the learning process instead of helping it, a scenario usually referred to as "curse of dimensionality".

Feature selection is an important pre–processing technique of Machine Learning aimed at coping with this curse (Kohavi & John, 1997). Its main goal is to find a small subset of the measured variables that improve, or at least do not degrade, the performance of the modeling method applied to the dataset. But feature selection methods do not only avoid the curse of dimensionality: they also allow for a considerable reduction in model complexity, an easier visualization and, in particular, a better interpretation of the data under analysis and the developed models (Liu et al., 2005).

Several methods have been introduced in recent years, from general ones like Wrappers (Kohavi & John, 1997) and filters (Kira & Rendell, 1992) to very specific ones developed for SVM (Nguyen & De la Torre, 2010; Weston et al., 2000) and RVM (Mohsenzadeh, Sheikhzadeh, & Nazari, 2016; Mohsenzadeh, Sheikhzadeh, Reza, Bathaee, & Kalayeh, 2013) classifiers. Amongst other methods in the field (Hua, Tembe, & Dougherty, 2009), the well-known Recursive Feature Elimination (RFE) algorithm provides good performance with moderate computational efforts (Guyon, Weston, Barnhill, & Vapnik, 2002) on wide datasets. The original and most popular version of this method uses a linear Support Vector Machine (SVM) (Vapnik, 2013) to select the candidate features to be eliminated. According to the SVM–RFE algorithm, the importance of an input variable i is directly correlated with the corresponding component (w_i) of the vector defining the separating hyperplane ($\mathbf{w}$). The method is widely used in Bioinformatics (Guyon et al., 2002; Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005). Alternative RFE methods using other classifiers have also been introduced in the literature (Granitto, Furlanello, Biasioli, & Gasperi, 2006; You, Yang, & Ji, 2014).

Typical feature selection algorithms are designed for binary classification problems, as the original version of RFE. Multiclass problems have received much less attention because of their increased difficulty. Also, because some classifiers involved in the se-

* Corresponding author.
E-mail addresses: izetta@cifasis-conicet.gov.ar (J. Izetta), verdes@cifasis-conicet.gov.ar (P.F. Verdes), granitto@cifasis-conicet.gov.ar (P.M. Granitto).

http://dx.doi.org/10.1016/j.eswa.2017.06.043
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

lection process are designed to solve binary problems. Most methods available for feature selection on multiclass problems are simple extensions of base methods. For example, RFE can be associated to a multiclass classifier like Random Forest (Breiman, 2001; Granitto et al., 2006).

Although SVM was originally developed to deal only with binary problems, it was extended to directly solve multiclass problems in different manners (Crammer & Singer, 2001; Hsu & Lin, 2002; Weston & Watkins, 1999), but with a modest success attributed mainly to the increased complexity of the solutions. On the other hand, in the last years several methods were developed to solve a multiclass problem using an appropriate combination of binary classifiers (Allwein, Schapire, & Singer, 2000; Hsu & Lin, 2002). The most usually followed strategy for multiclass SVM is known as "One–vs–One" (OVO). According to this approach, a classification problem with c classes is replaced with $M = c(c-1)/2$ reduced binary ones, each one of them consisting of discriminating a pair of classes. In order to classify a new example, it is passed through all binary classifiers and the most voted class is selected. Another useful strategy is "One–vs–All" (OVA). In this second case, a problem with c classes is replaced with $M = c$ reduced binary problems, each one of them consisting of discriminating a single class from all remaining ones.

Therefore, the most usual approach to implement a multiclass SVM–RFE method is to directly apply the RFE algorithm over an OVO or OVA multiclass SVM (Duan, Rajapakse, & Nguyen, 2007; Ramaswamy et al., 2001; Zhou & Tuck, 2007). The pioneering work of Ramaswamy et al. (2001) proposed the OVA solution, but also compared results with the OVO strategy. Duan et al. (2007) and Zhou and Tuck (2007) developed slight variations of the method, always considering both OVA and OVO implementations. Zhou and Tuck (2007) also considered solutions to the RFE problem using a direct multiclass implementation.

Interestingly, the solutions to the multiclass SVM–RFE problem that we have just described involve an important decision about the feature selection process which is usually neglected: they rank features by simply averaging components over the binary problems. For an input variable i they use $<|w_{ij}|>_j$, the mean importance over all binary problems j, as the corresponding importance. As we discuss in the next section, this strategy can lead to sub–optimal selections in many cases. Once the original multiclass problem has been divided into multiple binary ones, the feature selection problem can be treated in a similar way. Then, a possible solution is to cast the multiclass feature selection problem as the problem of selecting candidate features from multiple lists (Jurman et al., 2008), each list corresponding to a different binary sub-problem.

Similar solutions have been studied in related fields. In Bioinformatics, for example, Haury, Gestraud, and Vert (2011) discussed the combination of multiple lists of genes from bootstraps of the same gene-expression dataset. Zhou and Dickerson (2014) and Zhou and Wang (2016) proposed the use of class–dependent features (different features for each binary problem) for biomarker discovery. Dittman, Khoshgoftaar, Wald, and Napolitano (2013) showed that combining multiple lists in binary classification problems can improve the feature selection results. In a short work in text categorization, Neumayer, Mayer, and Nørvåg (2011) suggested that the combination of rankings generated by diverse methods can improve the results of using a single method. Kanth and Saraswathi (2015) used class–dependent features for speech emotion recognition, but using independent features for each class, not a final unique list.

In this work we discuss in depth the use of combination of multiple lists in feature selection for multiclass classification problems. We first introduce a simple mathematical framework for multiple lists. Using this framework, we propose diverse strategies to produce improved selection of feature subsets with SVM-RFE. Also, we use some specifically–designed artificial datasets and real–world examples to evaluate them extensively, using both the OVO and OVA strategies.

The rest of this article is organized as follows: in Section 2, we describe the feature selection methods introduced in this work. In Section 3 we evaluate these methods on diverse datasets and experimental setups. Finally, we draw our conclusions in Section 4.

2. List combination methods for SVM–RFE

The RFE selection method is a recursive process that ranks variables according to a given importance measure. At each iteration of the algorithm, the importance of each feature is calculated and the less relevant one is removed —in order to speed up the process, not one but a group of low relevance features is usually removed. Recursion is needed because the relative importance of each feature can change substantially when evaluated over a different subset of features during the stepwise elimination process, in particular for highly correlated features. The inverse order in which features are eliminated is used to create a final ranking. Then, the feature selection process itself is reduced to take the first n features from this ranking.

In the original binary version of SVM–RFE (Guyon et al., 2002), the projection of **w** (the normal vector to SVM's decision hyperplane) in the direction of feature i, w_i, is used as the importance measure. The method was efficiently extended to multiclass problems, employing the well-known OVO or OVA strategies to decompose the multiclass problem into a series of related binary ones (Duan et al., 2007; Ramaswamy et al., 2001; Zhou & Tuck, 2007). In both cases a set of M related binary problems is generated, each one solved by a vector $\mathbf{w_j}$. For each binary problem j, the importance of feature i is given by the corresponding component, w_{ij}.

In order to obtain a unique importance for each feature in this setup, the simplest solution is to average the absolute value of the components $|w_{ij}|$ over all related binary problems. We will call this method "Average" in the following. The Average solution is implemented, to the best of our knowledge, in all available RFE software packages, including the most popular amongst researchers (MATLAB, R and PYTHON platforms).

However, the only real advantage of the Average strategy is its simplicity. Two main drawbacks of this approach should be taken into consideration but are usually ignored:

1. The first issue can be called the *flattening* problem. Consider, for example, a feature e which is able to separate class j from all remaining classes, but is uninformative in other cases. Component w_{ej} will be large, but components w_{ek} with $k \neq j$ will be small, giving a low value for $<w_{ej}>_j$. Consider now another feature d which can give a modest help in separating any class from the others, obtaining always moderate values of w_{dj}, and therefore giving a medium value for $<w_{dj}>_j$. The Average strategy will clearly rank the latter over the former, but in most scenarios it will be desirable to keep the first variable over the second.
2. The second issue with the Average solution refers to *relative scales*. The length of vector $\mathbf{w_j}$ is different for each binary problem, as it depends on the margin of the solution, which can change considerably for classes that are relatively close or far away in feature space. Averaging components of vectors of different lengths can lead to the selection of sub–optimal subsets.

New strategies for feature selection able to overcome these drawbacks are needed. Here we propose to cast the problem as a selection of candidate features from multiple ranking lists (Jurman et al., 2008). We start by decomposing the multiclass problem into a set of M related binary problems (through the OVA or