Information Sciences 325 (2015) 466–483

Contents lists available at ScienceDirect

Information Sciences

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

A direct measure of discriminant and characteristic capability for classifier building and assessment

Giuliano Armano*

DIEE – Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi 09123, Cagliari, Italy

ARTICLE INFO

Article history:
Received 6 September 2014
Revised 5 April 2015
Accepted 6 July 2015
Available online 15 July 2015

Keywords:
Classifier performance measures
Feature ranking/selection
Confusion matrices

ABSTRACT

Performance measures are used in various stages of the process aimed at solving a classification problem. Unfortunately, most of these measures are in fact *biased*, meaning that they strictly depend on the class ratio – i.e. on the imbalance between negative and positive samples. After pointing to the source of bias for the best known measures, novel unbiased measures are defined which are able to capture the concepts of discriminant and characteristic capability. The combined use of these measures can give important information to researchers involved in machine learning or pattern recognition tasks, in particular for classifier performance assessment and feature selection.

1. Introduction

How to assess the performance of a classifier and the importance of features are key issues in the process of classifier building and assessment. Although framed in the same process, in principle (and in practice) these research topics follow different perspectives. For this reason, short summaries concerning the corresponding issues and proposals will be separately illustrated.

1.1. Assessing classifier performance

As for classifier performance assessment, there are a number of measures which are well known in the machine learning and pattern recognition communities. Let us recall, in particular: accuracy (strictly related to the error rate), precision, sensitivity (also called recall), specificity, F_1 and Matthews Correlation Coefficient (*MCC*). These measures share a common source, as they are all derived from confusion matrices (also called contingency tables). Other well known measures include Mean Square Error [6], cross-entropy [20], and AUC (e.g., [17]). Relevant work devised to shed light on the characteristics of the above measures include [34], [22], [39], and [24]. There are also many graphical representations and tools for model evaluation, such as ROC curves, a 2D visual environment widely acknowledged as the default choice for assessing the intrinsic behavior of a classifier (see, for instance, [5] and [13]), ROC isometrics [16], and cost curves [11]. More information on graphical methods for classifier performance assessment can be found in [32].

A further category of measures is aimed at assessing to which extent the classifier at hand is "keen" to classify inputs as belonging to the main or alternate category (the bias) and to which extent a classifier varies its performance depending on the datasets used for testing (the variance). The interested reader can consult, for instance, the work of Domingos [10] for more information on the issues related to bias and variance.

* Tel.: +39 706755758.
E-mail address: armano@diee.unica.it, armanodiee@gmail.com

http://dx.doi.org/10.1016/j.ins.2015.07.028
0020-0255/© 2015 Elsevier Inc. All rights reserved.

Relevant issues arise also when tests occur in the presence of imbalance between negative and positive samples. In that case, measuring the overall accuracy (or the overall error) gives poor information about the underlying process enforced by the classifier at hand, any such measure being typically affected by the imbalance between data (the more the imbalance is, the less significant the measure is). This fact may be further worsened by a lack of statistical significance of experimental results, which may hold for minority test samples. While no practical solution exists which is able to contrast the latter issue, the former is usually dealt with by adopting a combination of measures, usually a pair, devised to assess the performance beyond the fact that test data are unbalanced.[1] Precision and recall, on one hand, and specificity and sensitivity, on the other hand, are typical examples of this strategy. Also ROC diagrams follow this approach, the default choice for their axes being false positive rate (i.e., $1 -$ specificity) on the x axis and true positive rate (i.e., sensitivity) on the y axis. A further strategy for assessing classifiers, often adopted in the presence of unbalanced data samples, consists of defining a single compound measure defined on top of other ones. F_1 and MCC are both examples of this strategy.

Unfortunately, regardless of the adopted strategy, most of the existing measures are in fact *biased*, meaning that they strictly depend on the class ratio – i.e. on the imbalance between positive and negative samples. However, the adoption of biased measures can only be recommended when the statistics of input data is available. In the event one wants to assess the *intrinsic* properties of a classifier, or other relevant aspects in the process of classifier building and assessment, the adoption of biased measures may not be a reliable choice. For this reason, in the literature, some proposals have been made to introduce unbiased measures – see in particular the work of Flach [16].

1.2. Ranking/selecting features

Several techniques have been devised to support the process of ranking/selecting features according to their importance. Besides classical approaches, e.g., Fisher linear discriminant analysis [15] and Pearson correlation coefficient [14], a number of proposals have been made over time. With the goal of providing a better understanding of the underlying mechanisms, let us summarize the proposals according to two different perspectives.

Starting from the definition given in [4], Guyon and Elisseeff [21] divide feature selection methods in three broad groups: filter, wrapper and embedded methods. The corresponding definitions concentrate on the dependence between the selection method and the underlying learning algorithm L: (i) filters are used independently from L, (ii) embedded methods are used inside L, and (iii) wrappers use L as a black box. Note that the ordering adopted while recalling these groups is not accidental. In fact, filter methods are usually the fastest, as they consist of proper pre-processing activities, while wrapper methods are expected to be the most expensive in terms of computing resources, as the learning algorithm is repeatedly called with the aim of scoring (subsets of) features according to their predictive power. Embedded methods lay in between, being slower than filters but faster than wrappers.

An alternative grouping strategy can be found in the work of Dash and Liu [8]. Borrowing relevant categories from [3] and [4], the authors divide evaluation functions in five groups: distance, information-theoretic, dependence, consistency, and classifier error rate. Also known as separability, divergence, or discrimination measures, distance measures (and their counterpart, i.e., similarity measures) suggest that a feature should be preferred to another when it induces a greater difference between class conditional probabilities. Notable examples in this category are LDA [19] and ICA [26]. Similarity/dissimilarity measures can also be used to encode the sample space with the aim of imposing some useful constraints therein. Recent work in this category includes LFDA [37], sparse LDA [33] and CDA [30]. Information-theoretic measures, typically aimed at evaluating the entropy or the mutual information of features, have been adopted in various feature ranking or feature subset assessment methods (see, for instance, [2], [27], [12], [35] and [7]). Dependence/correlation measures quantify the ability to predict the value of one variable from the value of another variable. They may be used to measure the dependence between/among input features (typically feature pairs) and to identify whether a correlation exists between a feature/a set of features and the desired output. Beyond the classical correlation measures (e.g., Pearson correlation coefficient or cosine similarity), more recent proposals in this field include CCA (e.g., [23]) and dCOV/dCOR [38]. Consistency measures evaluate the distance of a feature subset from the consistent state. As a subset of features may be more or less distant from the consistent state, a consistency based algorithm typically accepts or rejects a feature subset depending on an inconsistency threshold – usually set by the user. Among other proposals made in this field, let us recall [9], [40], [1] and [36]. Classifier error rate delegates to a classifier the assessment of feature subsets. Relevant proposals that fall in this group are [25], [31] and [29].

It is worth pointing out that a bridge between the cited classifications of feature selection methods can be easily found. Indeed, the first four types of evaluation measures are typically framed according to a *filter* perspective, due to their potential independence from the classifier at hand. However, nothing prevents from devising heuristics *embedded* by a learning algorithm that make use of any of those methods (e.g. the information gain heuristics adopted by decision trees). Classifier error rate measures (the fifth group) coincide "de facto" with the *wrappers* category.

[1] Note that getting information about the intrinsic performance of a classifier could be an issue also in presence of complete knowledge about the statistics of data the classifier is expected to handle. In fact, high imbalance in a dataset may not allow to check whether the adopted classifier is in fact able to discriminate or simply puts into practice a "dummy" strategy – recognizing all or the majority of samples as belonging to the most populated category.