Pattern Recognition 61 (2017) 394–404

Contents lists available at ScienceDirect

Pattern Recognition

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

Designing multi-label classifiers that maximize *F* measures: State of the art

Ignazio Pillai, Giorgio Fumera *, Fabio Roli

Dept. of Electrical and Electronic Eng., University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy

ARTICLE INFO

Article history:
Received 22 February 2016
Received in revised form
15 July 2016
Accepted 10 August 2016
Available online 13 August 2016

Keywords:
Multi-label classification
F measure
Learning algorithms
Empirical utility maximization
Decision-theoretic approach

ABSTRACT

Multi-label classification problems usually occur in tasks related to information retrieval, like text and image annotation, and are receiving increasing attention from the machine learning and pattern recognition fields. One of the main issues under investigation is the development of classification algorithms capable of maximizing specific accuracy measures based on precision and recall. We focus on the widely used *F* measure, defined for binary, single-label problems as the weighted harmonic mean of precision and recall, and later extended to multi-label problems in three ways: macro-averaged, micro-averaged and instance-wise. In this paper we give a comprehensive survey of theoretical results and algorithms aimed at maximizing *F* measures. We subdivide it according to the two main existing approaches: empirical utility maximization, and decision-theoretic. Under the former approach, we also derive the optimal (Bayes) classifier at the population level for the instance-wise and micro-averaged *F*, extending recent results about the single-label *F*. In a companion paper we shall focus on the micro-averaged *F* measure, for which relatively fewer solutions exist, and shall develop novel maximization algorithms under both approaches.

1. Introduction

Multi-label (M-L) classification problems, like document categorization, and image and video annotation, usually occur in the design of information retrieval (IR) systems. They consist of deciding whether an instance (e.g., a document) is relevant or not to a given set of queries, which can be viewed as non-mutually exclusive labels. An instance can thus be assigned more than one label. Over the past ten years, M-L classification problems have received an increasing attention from the pattern recognition and machine learning research communities (see, e.g., [33,36]). One of the main topics under investigation is the development of learning algorithms tailored to specific M-L accuracy measures. Such measures are mostly based on precision and recall, which are the main metrics used for evaluating the performance of IR systems. They are different from the ones used in single-label (S-L) problems, like the misclassification probability.

In this work we focus on the widely used *F* measure. It has been originally proposed to evaluate IR systems in [30,34], and is defined as the weighted harmonic mean of precision and recall. It is also used to evaluate the accuracy of S-L binary classifiers aimed at discriminating instances relevant to a query from non-relevant ones.[1]

Three different versions of the *F* measure have subsequently been defined for M-L problems: instance-wise, macro- and micro-averaged. Under the viewpoint of the target accuracy measure, the existing approaches to M-L classifier design can be subdivided into two groups. Works in the first group (including most of the earlier ones) do not focus on a specific measure; they use S-L learning algorithms, and deal with multiple labels per sample using *problem transformation* or *algorithm adaptation* strategies (see the surveys of [33,36]). Among the former, the simplest one is *binary relevance* (BR), which consists of independently learning a binary classifier for each label, disregarding label correlation; other approaches have been proposed to attain a trade-off between taking into account label dependencies and keeping computational complexity low. Works in the second group focus on developing algorithms to maximize a specific accuracy measure, most often one of the M-L *F* measures. Maximizing the *F* measures (including the S-L one) is however particularly difficult since, contrary to S-L

* Corresponding author.
E-mail addresses: pillai@diee.unica.it (I. Pillai), fumera@diee.unica.it (G. Fumera), roli@diee.unica.it (F. Roli).
URL: http://pralab.diee.unica.it (G. Fumera).

[1] The S-L *F* measure is also used in binary problems not related to IR, but characterized by relevant class imbalance. In this case the misclassification probability is not a suitable performance measure, since a classifier that always predicts the majority class attains an accuracy equal to the corresponding prior.

http://dx.doi.org/10.1016/j.patcog.2016.08.008
0031-3203/

measures like accuracy, they do not decompose either over samples, or over labels, or both. Two different approaches for maximizing the S-L and M-L F measures have been considered, in turn [19]. The empirical utility maximization (EUM) approach aims at finding the decision rule which maximizes the chosen F measure on a finite sample of *labelled* instances; this approach has been used to develop several learning algorithms. The decision-theoretic approach (DTA) aims instead at finding the label assignments that maximize the expected value of the chosen F measure on a fixed set of *unlabeled* instances, with respect to their joint label-conditional probability; in practice, this probability is estimated from training data, whereas the unlabeled instances correspond to testing data.

In the present paper we give a comprehensive survey of existing algorithms for maximizing the F measures, which is still lacking in the literature. Nearly all works published in pattern recognition venues follow the EUM approach. Both EUM and DTA have been considered in machine learning venues, instead, where different EUM algorithms have been proposed, and the optimal (Bayes) classifier at the population level has also been recently derived for some of the F measures. Moreover, most of the earlier works focused on a single version of the F measure, and only recently (since [5]) the distinction between the S-L and M-L F, and between the three M-L versions, was clearly pointed out. Our survey can be useful for further developments in this field, especially for the pattern recognition community. As a by-product, we also derive the optimal classifier at the population level for the M-L instance-wise and micro-averaged F, under the EUM approach, extending recent results about the S-L F.

In a companion paper [28] we shall focus on the M-L micro-averaged F, for which relatively fewer solutions exist, and shall develop both learning algorithms based on EUM and an inference algorithm based on DTA.

The rest of this paper is structured as follows. After giving a formal definition of the F measures in Section 2, in Section 3 we describe EUM and DTA. We then survey existing works based on such approaches, respectively in Sections 4 and 5, for each of the F measures (including the S-L F).

Table 1
Notation used in this paper.

$\mathcal{X}$	instance space (e.g., a vector space)
$\mathbf{x} \in \mathcal{X}$	a single instance (e.g., a feature vector)
m	number of labels
$\mathcal{Y} = \{0, 1\}^m$	label space
	$\mathbf{y} = \{y_1, \ldots, y_m\} \in \mathcal{Y}$
label vector of an instance (ground truth)	
$y_i \in \{0, 1\}$	i-th label of an instance, $i = 1, \ldots, m$
n	number of instances in a given data set
$(\mathbf{x}^j, \mathbf{y}^j)$	j-th instance of a given data set, $j = 1, \ldots, n$
$h(\cdot; \theta)$: $\mathcal{X} \mapsto \mathcal{Y}$	multi-label classifier with parameters θ
	$h_i(\cdot; \theta)$: $\mathcal{X} \mapsto \{0, 1\}$
i-th label assigned by classifier $h(\cdot; \theta)$, $i = 1, \ldots, m$	
h_i, h_i^j	short-hand notation respectively for $h_i(\cdot; \theta)$ and $h_i(\mathbf{x}^j; \theta)$
$\mathbf{h}$, $\mathbf{h}^j \in \{0, 1\}^m$	label vectors, respectively $(h_1, \ldots, h_m)$ and $(h_1^j, \ldots, h_m^j)$
F_β^{b}	S-L, binary F_β (Eq.(4))
F_β^{i}	M-L, instance-wise F_β (Eq.(5))
F_β^{M}	M-L, macro-averaged F_β (Eq.(6))
F_β^{m}	M-L, micro-averaged F_β (Eq.(7))

2. Definition of F measures

In Tables 1 and 2 we summarize respectively the notation and the abbreviations used in this paper. We shall use upper-case letters to denote random variables, and the corresponding lower-case letters to denote their values.

For a given M-L problem, let m denote the number of labels, $\mathcal{X}$ the input space (e.g., a feature vector space), $\mathbf{x} \in \mathcal{X}$ an instance (e.g., a feature vector), and $\mathbf{y} \in \mathcal{Y} = \{0, 1\}^m$ the corresponding label vector, where $y_i = 1(0)$ means that $\mathbf{x}$ is (not) relevant to the i-th label. A M-L classifier is commonly formalized as a function

$$h(\mathbf{x}; \theta) = (h_1(\mathbf{x}; \theta), \ldots, h_m(\mathbf{x}; \theta)) \in \mathcal{Y}, \tag{1}$$

where $h_i(\mathbf{x}; \theta) = 1$ (0) means that $\mathbf{x}$ is deemed as (non-)relevant to the i-th label, and θ denotes the parameter vector to be set by the learning algorithm. Precision (p) and recall (r) are the main measures used for evaluating the quality of the results produced by IR systems, in terms of the "degree of matching" between the true and the estimated relevance to a given query. They are defined respectively as the probability that a retrieved instance (e.g., a document) is relevant, and as the probability of retrieving a relevant instance, which are complementary aspects of an IR system's performance. Let $S = \{(\mathbf{x}^j, y^j)\}_{j=1}^n$ be a set of instances, where $y^j \in \{0, 1\}$ denotes the relevance of $\mathbf{x}^j$ to the considered query, and let $h^j \in \{0, 1\}$ denote the estimated relevance. Let TP, FP and FN denote the corresponding number of *true positive* (when $h^j = y^j = 1$), *false positive* ($h^j = 1$, $y^j = 0$) and *false negative* ($h^j = 0$, $y^j = 1$) decisions. Precision and recall can be estimated on the finite sample S as:

$$p = \frac{TP}{TP + FP} = \frac{\sum_{j=1}^n y^j h^j}{\sum_{j=1}^n h^j}, \tag{2}$$

$$r = \frac{TP}{TP + FN} = \frac{\sum_{j=1}^n y^j h^j}{\sum_{j=1}^n y^j}. \tag{3}$$

Single-label, binary F measure: The F measure has been originally proposed for IR systems, to combine p and r into a scalar [30,34]. Based on principled arguments, it is defined as the weighted harmonic mean of p and r. It is also often used to evaluate the accuracy of S-L, binary classifiers ($m = 1$) whose goal is to discriminate between relevant instances to a given query and non-relevant ones. The S-L F is defined on a finite sample as (the superscript 'b' stands for 'binary'):

$$F_\beta^{\mathrm{b}} = \frac{1 + \beta^2}{\frac{1}{p} + \beta^2 \frac{1}{r}} = \frac{(1 + \beta^2) TP}{(1 + \beta^2) TP + \beta^2 FN + FP} = \frac{(1 + \beta^2) \sum_{j=1}^n y^j h^j}{\beta^2 \sum_{j=1}^n y^j + \sum_{j=1}^n h^j}, \tag{4}$$

where $\beta \in [0, +\infty)$ controls the trade-off between p and r. Note that $F_0^{\mathrm{b}} = p$ and $F_{+\infty}^{\mathrm{b}} = r$. For $\beta = 1$ one obtains the unweighted harmonic mean: $F_1^{\mathrm{b}} = \frac{2}{1/p + 1/r}$.

Table 2
List of the abbreviations used in this paper.

IR	Information retrieval
S-L, M-L	Single-label, multi-label
BR	Binary relevance
EUM	Empirical utility maximization
DTA	Decision-theoretic approach

Multi-label F measures: Three different M-L versions of the F measure have been defined. The *instance-wise F* views instances as queries, whose relevant labels have to be retrieved. It is thus defined for a single instance $(\mathbf{x}, \mathbf{y})$ as:

$$F_\beta^{\mathrm{i}} = \frac{(1+\beta^2)\sum_{i=1}^m y_i h_i}{\beta^2 \sum_{i=1}^m y_i + \sum_{i=1}^m h_i}. \tag{5}$$

The macro-averaged F is computed on a set of instances; it is defined as the average of the S-L F measures computed for each label, and gives the same weight to each label:

$$F_\beta^{\mathrm{M}} = \sum_{i=1}^m \frac{(1+\beta^2)TP_i}{(1+\beta^2)TP_i + \beta^2 FN_i + FP_i} = \sum_{i=1}^m \frac{(1+\beta^2)\sum_{j=1}^n y_i^j h_i^j}{\beta^2 \sum_{j=1}^n y_i^j + \sum_{j=1}^n h_i^j}. \tag{6}$$

The micro-averaged F is computed after pooling the labels of all instances of a given set, and gives equal weight to each labeling decision:

$$F_\beta^{\mathrm{m}} = \frac{\sum_{i=1}^m (1+\beta^2)TP_i}{\sum_{i=1}^m [(1+\beta^2)TP_i + \beta^2 FN_i + FP_i]} = \frac{(1+\beta^2)\sum_{j=1}^n \sum_{i=1}^m y_i^j h_i^j}{\beta^2 \sum_{j=1}^n \sum_{i=1}^m y_i^j + \sum_{j=1}^n \sum_{i=1}^m h_i^j}. \tag{7}$$

To simplify the notation, from now on we will omit the subscript β in the symbols denoting the F measures, when it is not necessary.

Choice between the multi-label F measures: The three M-L F measures evaluate different aspects of classifier performance, and thus the choice between them is application-dependent. With regard to the problem of designing classifiers that maximize the M-L F measures, quoting from [5]: "One should carefully distinguish these versions, as algorithms optimized with a given objective are usually performing sub-optimally for other (target) evaluation measures." An empirical evidence of this fact was formerly reported in [6], where it was observed that tuning the decision thresholds of a classifier to maximize F^{M} can decrease the corresponding F^{m}. In particular, it is known that the differences between F^{M} and F^{m} can be large on data sets with rare labels [16]: since the F measures disregard true negatives (i.e., instance-label pairs such that $y_i^j = h_i^j = 0$) and their magnitude is mostly determined by the number of true positives, frequent labels dominate rare ones in F^{m}, whereas F^{M} is much more sensitive to rare labels. Further insights have been given in [15]: for a rare label, a perfect classifier only marginally improves F^{m} over a (trivial) classifier that labels all instances as non-relevant; moreover, for rare labels with an "uninformative predictive model" (i.e., a classifier which outputs the same score for all instances), F^{m} and F^{M} are maximized by classifying all instances respectively as non-relevant and as relevant.

Maximizing the F measures: Under the viewpoint of classifier design, maximizing the S-L and M-L F measures is more difficult than maximizing traditional S-L measures based on the 0–1 loss function and the corresponding misclassification probability, or their variants. The latter are *uni-variate* measures, i.e., they decompose over instances. This means that the optimal label assignment to any given instance is independent of other instances. On the contrary, F^{M} (as well as F^{b}) does not decompose over instances; F^{i} does not decompose over labels; and F^{m} does not decompose over either. Therefore, F^{M} and F^{m} (as well as F^{b}) are multi-variate, which implies that the optimal label assignments to a given instance depend also on the assignments to the *other* instances on which these measures are computed. Additionally, in the case of F^{m} the different label assignments, even for different instances, influence each other. Accordingly, the maximization of these measures is in principle computationally demanding, or even infeasible. Moreover, it fits only batch or off-line settings; in on-line settings one should, e.g., classify the incoming samples in batches, or consider a subset of the previously processed instances when labeling an incoming one [14]. Similarly, although F^{i} is univariate, its maximization requires in principle to consider all possible 2^m label assignments, which is feasible only when the number of labels is small.

3. Approaches to F measure maximization

As mentioned in Section 1, two approaches for maximizing the F measures, both in S-L (F^{b}) and in M-L classification problems (F^{i}, F^{M} and F^{m}), have been proposed so far: EUM and DTA [4,19]. The existing maximization algorithms are surveyed in the next two sections, and are summarized in Table 3. We point out that, with the only exception of [29], all works published in pattern recognition venues follow the EUM approach.

The EUM approach consists of learning a classifier of the form $h(\cdot;\theta)$: $\mathcal{X} \mapsto \mathcal{Y}$ that maximizes the chosen F measure on a given training set of labelled instances $S = \{(\mathbf{x}^j, \mathbf{y}^j)\}_{j=1}^n$; the learnt classifier is then used to predict the label assignments of testing data. In principle, this requires one to jointly evaluate all possible label assignments to S, which amount to 2^n for F^{b}, $n \times 2^m$ for F^{i}, $m \times 2^n$ for F^{M}, and 2^{mn} for F^{m}. Learning algorithms based on EUM have been developed for all F measures, except F^{m}, and the consistency of several learning algorithms has also been investigated. In some of the most recent works, the optimal (Bayes) classifier at the population level has also been derived for the S-L F (which also applies to the M-L, macro-averaged F); it has also been shown that all F measures but the instance-wise can be maximized by reduction to a cost-sensitive problem.

The DTA (also called plug-in rule approach in [4]) focuses instead on a *fixed*, unlabeled sample (testing data) $S = \{\mathbf{x}^j\}_{j=1}^n$ ($n=1$

Table 3
Summary of existing EUM- and DTA-based methods (described respectively in Sections 4 and 5) for maximizing the S-L F^{b} measure and the three M-L F measures.

Empirical utility maximization approach (Section 4)

Works	Measure	Main characteristics
[9,17,20]	F^{b}	Non-convex optimization
[11,18]	F^{b}	SVM-like classifier, convex objective function
[13,13,15,19,23,37]	F^{b}	Optimal classifier, reduction to cost-sensitive problem
[22,13,12]	F^{b}	Consistency analysis of maximization algorithms
[32,25,24]	F^{i}, F^{M}	SVM-like classifier, convex objective function
[6,26,27]	F^{m}	Tuning of binary classifiers' thresholds
[13,23]	F^{M}, F^{m}	Optimal classifier, reduction to cost-sensitive problem

Decision-theoretic approach (Section 5)

Measure	Works and main characteristics
F^{b}	[14]: $O(n2^n)$ complexity, approximate solution
	[1] $O(n^3)$, [10] $O(n^4)$, [19] $O(n^2)$ complexity, exact solution
F^{M}	Same algorithms for F^{b}, independently for each label
F^{i}	[5,2,35]: $O(m^3)$ complexity
F^{i}	[29] (limited to a specific decision rule): $O(m^3)$ complexity

in the case of F^{i}), and predicts through an inference procedure the label assignments that maximize the expectation of the chosen F measure on S, with respect to the joint label-conditional probability distribution $\mathbb{P}(\mathbf{Y}^1, \ldots, \mathbf{Y}^n|\mathbf{x}^n, \ldots, \mathbf{x}^n)$. In practice, this distribution is estimated from training data. The corresponding maximization problem is computationally very demanding as well, since the expectation has to be computed over all possible combinations of true and assigned labels. The number of such combinations is 2^{2n} for F^{b}, 2^{2m} for F^{i}, $m2^{2n}$ for F^{M}, and 2^{2mn} for F^{m}. Maximization algorithms based on DTA have been proposed so far for F^{b} (they also apply to F^{M}) and F^{i}, but not for F^{m}. The consistency of DTA has also been investigated in recent works.

EUM and DTA have been compared in [19], focusing on the S-L F^{b}. These approaches were found to be equivalent asymptotically (i.e., for large training and test sets), provided that the underlying models are accurate. An empirical analysis also provided evidence that EUM is more robust against model misspecification; on the other hand, if an accurate model is chosen, DTA was found to be better in the presence of rare classes, as well as in the common domain adaptation scenario where $\mathbb{P}(\mathbf{X})$ changes while $\mathbb{P}(Y|\mathbf{X})$ remains constant.

A comparison between EUM and DTA focused on M-L problems has later been carried out in [4], limited to the instance-wise F^{i}. In this comparison the EUM framework for structured loss minimization of [32] was considered, together with two specific implementations based on surrogate, convex loss functions [24,25] (see Section 4.2). The analysis of the infinite sample case showed that the DTA is consistent, i.e., it converges to the Bayes optimal classifier for the F^{i} measure, whereas the considered EUM algorithms are not. A further analysis on finite data sets was carried out in [4], by comparing the exact DTA-based inference algorithms for the two cases of conditionally independent and conditionally dependent labels (see Section 5.2), and the EUM-based learning algorithms mentioned above. DTA-based algorithms were found to be more effective than EUM-based ones; they also exhibited a higher efficiency in the training step and for parameter tuning, but a lower efficiency in the inference step.

4. Empirical utility maximization approach

In this section we describe learning algorithms developed for the S-L and M-L F measures, and then summarize recent theoretical results about the EUM approach. We finally complement such results by deriving the optimal classifier at the population level for the micro-averaged and the instance-wise F.

Learning algorithms proposed so far can be subdivided into four categories: variants of the SVM learning algorithm (based on the maximum-margin approach) [18,11,32,24,25], whose objective function is (except for [18]) a convex approximation of an F measure; optimization algorithms whose objective function is a non-convex approximation [9,17,20]; algorithms that tune the decision thresholds of binary classifiers [6,26,27,22,13,12]; and cost-sensitive algorithms [13,23].

4.1. Single-label F measure

The first learning algorithm was proposed in [18], as a modification of the SVM learning algorithm. The objective function of the latter includes a penalty term which upper bounds the number of misclassified training instances. This term was replaced by the following approximation of $2(1/F_1^{\mathrm{b}} - 1)$, which is a possible loss function corresponding to the use of F_1^{b} as the accuracy measure:

$$\frac{\sum_{j=1}^{n}(1 - \exp(\alpha\xi_j))_+}{n_+ - \sum_{j=1}^{n}\mathbb{I}[y^j = 1](1 - \exp(\alpha\xi_j))_+}, \tag{8}$$

where $\mathbb{I}[a] = 1$ (0) if a= true (false), $x_+ = x$ (0) if $x \geq 0$ (< 0), n_+ is the number of instances with label 1, and α is a positive constant. However, Eq. (8) is non-convex: finding the global minimum of the resulting objective function is not guaranteed, and the optimization problem exhibits a much higher computational complexity than the one of SVMs. Another interesting result was given in [18], related to a different, heuristic modification to the SVM penalty term, formerly proposed by other authors for balancing precision and recall. It consists of assigning different weights to misclassified instances of the two classes:

$$C_+ \sum_{j=1}^{n}\mathbb{I}[y^j = 1]\xi_j + C_- \sum_{j=1}^{n}\mathbb{I}[y^j = 0]\xi_j, \tag{9}$$

where ξ_j is the hinge loss for the j-th training instance. The solution of the corresponding learning problem turned out to approximate the one obtained using (8), for suitable values of C_+ and C_-. In Section 4.3 we shall see that recent theoretical results have proven the equivalence between maximizing F^{b} at the population level and minimizing the expected error with suitable asymmetric misclassification costs.

In [11] an extension of the SVM learning algorithm to performance measures that do not decompose into expectations over instances, including F^{b}, was proposed. It minimizes a *convex* upper bound of the corresponding loss function, and uses a multi-variate decision function which jointly labels *all* training instances (the class labels are conveniently denoted here as -1 and $+1$):

$$h(\mathbf{x}^1, \ldots, \mathbf{x}^n; \mathbf{w}) = \arg \max_{h^1,\ldots,h^n \in \{-1,+1\}^n} \left\langle \mathbf{w}, \sum_{j=1}^{n} h^j \mathbf{x}^j \right\rangle, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. The learning problem is:

$$\begin{aligned} \min_{\mathbf{w},\xi \geq 0} \quad & \frac{1}{2} \| \mathbf{w} \|^2 + C\xi \\ \text{s. t.} \quad & \forall (h^1, \ldots, h^n) \in \{-1, +1\}^n \setminus \{(y^1, \ldots, y^n)\}: \\ & \left\langle \mathbf{w}, \left(\sum_{j=1}^{n} y^j \mathbf{x}^j - \sum_{j=1}^{n} h^j \mathbf{x}^j \right) \right\rangle \\ & \geq \Delta(h^1, \ldots, h^n, y^1, \ldots, y^n) - \xi \end{aligned} \tag{11}$$

where Δ denotes the loss function. If the performance measure is F^{b}, then $\Delta = 1 - F^{\mathrm{b}}$. In principle, Eq. (10) requires one to evaluate 2^n different label assignments; moreover, the learning problem (11) has $2^n - 1$ constraints. Nevertheless, since (10) is a linear function, its maximum can be computed by independently considering each of the n assignments $(h^1, \ldots, h^n)$. Moreover, problem (11) can be solved with $O(n^2)$ computational complexity, thanks to the properties of F^{b}, using an optimization strategy proposed in [31]. SVMs turns out to be a particular case of the above classifier, when the error rate is used in (11) as the loss function.

In [9,17] learning algorithms that maximize continuous but non-convex approximations of F^{b} were proposed, using numerical optimization techniques. In [9] the linear discriminant function of logistic regression classifiers was used, and F^{b} is approximated similarly to Eq. (8). To deal with non-convexity, the optimization algorithm was run several times, starting from randomly chosen parameter values. In [17] the class-conditional distribution $\mathbb{P}(\mathbf{X}|Y)$ is first estimated, then the TP, FP and FN counts are approximated, for a given discriminant function, by integrating $\mathbb{P}(\mathbf{X}|Y)$ in the corresponding decision regions. The parameters of the discriminant function that maximize F^{b} are finally estimated by an optimization algorithm.

4.2. Multi-label F measures

In the following we review existing EUM-based learning algorithms, separately for each of the three M-L F measures.

4.2.1. Instance-wise F

In [32] a SVM-like classifier was proposed for structured-output problems with instance-wise performance measures, including F^{i}. The proposed discriminant function exploits the structure and dependencies within the output values:

$$h(\mathbf{x};\mathbf{w}) = \arg\max_{\mathbf{h}\in\mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{h})\rangle, \tag{12}$$

where $\Psi(\mathbf{x},\mathbf{h})$ is a feature mapping (a combined feature representation of inputs and outputs), and $\langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{h})\rangle$ measures how "compatible" a pair $(\mathbf{x},\mathbf{h})$ is. The learning problem is:

$$\begin{aligned} &\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{n}C\sum_{j=1}^{n}\xi_j \\ &\text{s. t.} \quad \forall\, \mathbf{h}\in\mathcal{Y}\backslash\mathbf{y}^j,\ j=1,\ldots,n: \\ &\qquad \left\langle \mathbf{w}, \left(\Psi\left(\mathbf{x},\mathbf{y}^j\right) - \Psi\left(\mathbf{x},\mathbf{h}\right)\right)\right\rangle \geq \Delta\left(\mathbf{y}^j,\mathbf{h}\right) - \xi_j,\ \ \xi_j \geq 0. \end{aligned} \tag{13}$$

When the performance measure is F^{i}, then $\Delta = 1 - F^{\mathrm{i}}$. An efficient optimization algorithm was also developed, that explicitly examines only a small subset of the constraints in (13), which are $n\times(2^m-1)$ in the case of F^{i}.[2]

A similar approach was proposed in [25], which explicitly models the dependencies (only the positive correlations) between pairs of labels. The decision function is defined as:

$$h(\mathbf{x};\theta) = \arg\max_{\mathbf{h}\in\mathcal{Y}} \mathbf{h}^{\top}A\mathbf{h}, \tag{14}$$

where A is an $m\times m$ upper-triangular matrix defined as $A_{ii} = \langle \mathbf{x},\theta_i\rangle$, and $A_{ik} = C_{ik}\theta_{ik}$, $i\neq k$; the parameter vector θ_i weighs the features for the i-th class; C_{ij} is the normalized counts of co-occurrence of labels i and j in training instances; and θ_{ik} is a scalar parameter which is forced to be non-negative, implying that $A_{ik}>0$ for $i\neq k$, which allows (14) to be efficiently solved. The parameter θ in the left-hand side of (14) is defined as $(\theta_1,\ldots,\theta_m,\theta_{1,2},\theta_{1,3},\ldots,\theta_{m-1,m})$. The learning problem and the proposed optimization strategy are similar respectively to (13) and to the one of [32], to efficiently handle the constraints:

$$\begin{aligned} &\min_{\theta,\xi} \quad \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{j=1}^{n}\xi_j \\ &\text{s. t.} \quad \forall\ \mathbf{h}\in\mathcal{Y}\backslash\mathbf{y}^j,\ j=1,\ldots,n: \\ &\qquad (y^j)^{\top}A\mathbf{y}^j - \mathbf{h}^{\top}A\mathbf{h} \geq \Delta(\mathbf{y},\mathbf{h}) - \xi_j,\ \ \xi_j\geq 0. \end{aligned} \tag{15}$$

Finally, the specific setting in which an ensemble of independently trained binary classifiers are used for each label, and their scores are linearly combined, was considered in [8]. A non-convex approximation of F^{i} was devised, and an algorithm for maximizing it with respect to the combination weights was developed. We shall describe it in Section 4.2.3, since it was applied also to F^{m}.

4.2.2. Macro-averaged F

In [24] a SVM-like approach similar to the one of [32] was proposed for M-L loss functions that decompose over labels, including F^{M}. The classification problem is formulated as a *reverse prediction*: given a set of instances $\{(\mathbf{x}^j,\mathbf{y}^j)\}_{j=1}^n$, the m labels are considered as the set of *inputs*, and the instances that are relevant to a label are considered as the corresponding *output*. The input value corresponding to the i-th label is encoded as an m-dimensional vector $\mathbf{a}^i\in\{0,1\}^m$, with $a^i_i=1$, and $a^i_k=0$ for $k\neq i$; the corresponding output values are encoded as $\mathbf{b}^i\in\{0,1\}^n$, with $b^i_j=1$ (0) if the j-th instance is (not) relevant to the i-th label. A given data set is then transformed into a set of m instances $\{(\mathbf{a}^i,\mathbf{b}^i)\}_{i=1}^m$ made up of all possible input values and the corresponding output vectors. The decision function for the i-th label is defined as:

$$\bar{\mathbf{b}}^i = \arg\max_{\mathbf{b}\in\{0,1\}^n} \langle \phi(\mathbf{a}^i,\mathbf{b}),\theta\rangle, \tag{16}$$

where

$$\phi(\mathbf{a}^i,\mathbf{b}) = \sum_{j=1}^{n} b_j(\mathbf{x}^j\otimes\mathbf{a}^i) \in \mathbb{R}^{d\times m}, \tag{17}$$

d is the dimensionality of $\mathcal{X}$, and $\theta\in R^{d\times m}$ is a parameter matrix. Similarly to [25], the learning problem is:

$$\begin{aligned} &\min_{\theta,\xi} \quad \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ &\text{s. t.} \quad \forall\ \mathbf{b}\in\{0,1\}^n\backslash\mathbf{b}^i,\ i=1,\ldots,m: \\ &\qquad \langle\phi(\mathbf{a}^i,\mathbf{b}^i),\theta\rangle - \langle\phi(\mathbf{a}^i,\mathbf{b}),\theta\rangle \geq \Delta(\mathbf{b}^i,\mathbf{b}) - \xi_i,\ \ \xi_i\geq 0, \end{aligned} \tag{18}$$

where the loss function is defined as $\Delta(\mathbf{b}^i,\mathbf{b}) = 1 - F^{\mathrm{c},i}$. The term $\frac{1}{m}\sum_{i=1}^m\xi_i$ in the objective function is a convex upper bound on Δ. An efficient, $O(n^2)$ optimization algorithm was developed for solving problem (18). It was also shown that the decision function (16) can be computed in $O(n)$ time.

Since the M-L F^{M} measure is the average of the corresponding S-L F^{b} measures, it is pertinent to investigate the relationship between the maximum-margin approach of [24] (described above) and the one formerly developed in [11] (Section 4.1), aimed at maximizing respectively F^{M} and F^{b}. No comparison between these approaches was reported in [24]. As a contribution of this paper, here we show that these approaches are equivalent, as stated in the following Proposition:

Proposition 1. For $C=\frac{1}{4\lambda m}$, the M - L decision function (16) obtained by solving the learning problem (18) of [24] coincides with the set of decision functions (10) of independently trained binary classifiers (i.e., using BR) obtained by solving the learning problem (11) of [11].

Proof. We first prove that their decision functions are equivalent. Since $\mathbf{a}^i$ in [24] is defined as an m-dimensional column vector in which the i-th element is 1 and all the other ones equal 0, it follows that $\mathbf{x}^j\otimes\mathbf{a}^i$ in Eq. (17) is a $d\times m$ matrix in which the i-th column equals $\mathbf{x}^j$, and all the other elements are zero. Therefore, also $\phi(\mathbf{a}^i,\mathbf{b})$ in Eq. (17) is a $d\times m$ matrix, in which the i-th column equals $\sum_{j=1}^n b_j\mathbf{x}^j$ and all the other elements are zero. The argument of the arg max in (16) can thus be rewritten as:

$$\langle\phi(\mathbf{a}^i,\mathbf{b}),\theta\rangle = \left\langle \sum_{j=1}^{n} b_j\mathbf{x}^j,\theta_i\right\rangle. \tag{19}$$

This means that the assignment for the i-th label depends only on θ_i. We can thus rewrite the decision function (16) for the i-th label as:

$$\bar{\mathbf{b}}^i = \arg\max_{\mathbf{b}\in\{0,1\}^n} \left\langle \sum_{j=1}^{n} b_j\mathbf{x}^j,\theta_i\right\rangle. \tag{20}$$

We now make the following change of variables:

[2] An alternative formulation was also proposed, in which the right-hand side of each constraint is $1-\xi_j/\Delta(\mathbf{y},\mathbf{h})$, as well as equivalent formulations in which quadratic terms ξ_j^2 are used in the objective function for penalizing margin violations.

$$\bar{h}_i^{\,j} = 2\bar{b}_j^i - 1, \quad h^j = 2b_j - 1, \quad \mathbf{w}_i = \frac{1}{2}\theta_i. \tag{21}$$

Note that this implies that $h^j \in \{-1, +1\}$. The decision function (20) for the i-th label can be rewritten as:

$$\begin{aligned}\bar{h}_i^{\,1}, \ldots, \bar{h}_i^{\,n} &= \arg\max_{h^1,\ldots,h^n \in \{-1,1\}^n} \left\langle \sum_{j=1}^{n} \left(\frac{h^j + 1}{2} \right) \mathbf{x}^j, 2\mathbf{w}_i \right\rangle \\ &= \arg\max \left(\left\langle \sum_{j=1}^{n} h^j \mathbf{x}^j, \mathbf{w}_i \right\rangle + \left\langle \sum_{j=1}^{n} \mathbf{x}^j, \mathbf{w}_i \right\rangle \right).\end{aligned} \tag{22}$$

The last term $\langle \sum_{j=1}^{n} \mathbf{x}^j, \mathbf{w}_i \rangle$ is constant with respect to $h^1, \ldots, h^n$, which makes the decision function (22) identical to (10).

We now prove that the learning problems are equivalent, for a proper choice of their parameters λ and C. The objective function of problem (18) can be rewritten by explicitly indicating the Frobenius norm of the parameter matrix θ as a function of the 2-norm of its columns, denoted as θ_i, $i = 1, \ldots, m$:

$$\min_{\theta,\xi} \frac{\lambda}{2} \sum_{i=1}^{m} \| \theta_i \|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i. \tag{23}$$

Using (19), the constraints of (18) can be rewritten as:

$$\begin{aligned}&\forall \ \mathbf{b} \in \{0,1\}^n \backslash \mathbf{b}^i, \quad i = 1, \ldots, m: \\ &\left\langle \left(\sum_{j=1}^{n} b_j^i \mathbf{x}^j - \sum_{j=1}^{n} b_j \mathbf{x}^j \right), \theta_i \right\rangle \geq (1 - F_\beta^{c,i}) - \xi_i, \ \ \xi_i \geq 0.\end{aligned} \tag{24}$$

It is now evident that minimizing (23) under constraints (24) amounts to solving the following m *independent* optimization problems, one for each label:

$$\begin{aligned}\min_{\theta_i, \xi_i \geq 0} \quad & \frac{\lambda}{2} \| \theta_i \|^2 + \frac{1}{m} \xi_i \\ \text{s. t.} \quad & \forall \ \mathbf{b} \in \{0,1\}^n \backslash \mathbf{b}^i : \left\langle \left(\sum_{j=1}^{n} b_j^i \mathbf{x}^j - \sum_{j=1}^{n} b_j \mathbf{x}^j \right), \theta_i \right\rangle \\ & \geq (1 - F_\beta^{c,i}) - \xi_i.\end{aligned} \tag{25}$$

We now make another change of variables:

$$y_i^j = 2b_j^i - 1. \tag{26}$$

Together with (21), this allows us to rewrite the constraints of (25) as:

$$\begin{aligned}&\forall \ (\mathbf{h}^1, \ldots, \mathbf{h}^j) \in \{-\mathbf{1}, \mathbf{1}\}^n \backslash \{(\mathbf{y}_i^1, \ldots, \mathbf{y}_i^n)\} \\ &: \left\langle \left(\sum_{j=1}^{n} \left(\frac{y_i^j + 1}{2} - \frac{h^j + 1}{2} \right) \mathbf{x}^j \right), 2\mathbf{w}_i \right\rangle \\ &\geq (1 - F_\beta^{c,i}) - \xi_i \equiv \left\langle \left(\sum_{j=1}^{n} (y_i^j - h^j) \mathbf{x}^j \right), \mathbf{w}_i \right\rangle \\ &\geq (1 - F_\beta^{c,i}) - \xi_i,\end{aligned} \tag{27}$$

which are identical to the constraints of (11), for the i-th label. Finally, using (21), the objective function of (25) becomes:

$$\frac{\lambda}{2} \| 2\mathbf{w}_i \|^2 + \frac{1}{m} \xi_i = \frac{1}{2}(4\lambda) \| \mathbf{w}_i \|^2 + \frac{1}{m} \xi_i. \tag{28}$$

The solution of the corresponding learning problem does not change by rescaling the objective function (28); dividing it by 4λ, it becomes identical to the objective function of (11) when $C = \frac{1}{4\lambda m}$, which completes our proof. □.

4.2.3. Micro-averaged F

F^{m} is the most challenging measure, since it does not decompose over instances nor over labels. Existing EUM-based approaches consist of using a M-L decision function defined as $h_i(\mathbf{x}) = \mathrm{sign}[f_i(\mathbf{x}) - \theta_i]$, $i = 1, \ldots, m$, where $f_i(\mathbf{x})$ are real-valued discriminant functions obtained by independently training one binary classifier for each label (using any performance measure), whereas $\theta_i \in \mathbb{R}$ are decision thresholds that are tuned afterwards (i.e., keeping fixed the $f_i(\cdot)$'s) to maximize F^{m} on validation data.[3] Let $F^{\mathrm{m}}(\theta_1, \ldots, \theta_m; S)$ denote the value of F^{m} computed on a given data set S (e.g., a validation set) as a function of the decision thresholds. The optimal threshold values are the solution of the following optimization problem:

$$\theta_1^*, \ldots, \theta_m^* = \arg\max_{\theta_1,\ldots,\theta_m} F^{\mathrm{m}}(\theta_1, \ldots, \theta_m; S). \tag{29}$$

This approach was first proposed in [6], where a heuristic optimization procedure shown as Algorithm 1 was developed. Algorithm 1 consists of iteratively updating a *single* threshold at each step by maximizing the corresponding F^{m}, while keeping all the other thresholds at their current values, until some stopping criterion is met. Since $F^{\mathrm{m}}(\theta_1, \ldots, \theta_m; S)$ can attain up to $|S| + 1$ distinct values with respect to any single threshold, the corresponding maximization step (the arg max step of Algorithm 1) can be solved by a simple line search with complexity $O(|S|)$. This approach was proposed in [6] without theoretical support nor optimality guarantees.

Algorithm 1. F^{m} maximization algorithm of [6].

Input: m trained binary classifiers f_i, a data set S, a constant $\epsilon > 0$
Output: m decision thresholds

$\theta_1^{(0)} \leftarrow 0, \ldots, \theta_m^{(0)} \leftarrow 0$, $F^{(0)} \leftarrow F^{\mathrm{m}}(\theta_1^{(0)}, \ldots, \theta_m^{(0)}; S)$, $t \leftarrow 1$
repeat
  for $i = 1, \ldots, m$ **do**
    $\theta_i^{(t)} \leftarrow \mathrm{argmax}_\theta F^{\mathrm{m}}(\theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta, \theta_{i+1}^{(t-1)}, \ldots, \theta_m^{(t-1)}; S)$
  end for
  $F^{(t)} \leftarrow F^{\mathrm{m}}(\theta_1^{(t)}, \ldots, \theta_m^{(t)}; S)$
until $\frac{F^{(t)} - F^{(t-1)}}{F^{(0)}} < \epsilon$
return $\theta_1^{(t)}, \ldots, \theta_m^{(t)}$

In [26,27] we analyzed the optimization problem (29), by studying the behavior of F^{m} as a function of $\theta_1, \ldots, \theta_m$ on a given sample S. Our main result was the following proposition (reported in [27] as Property 1):

Proposition 2. Consider any given value $\theta'_1, \ldots, \theta'_m$ of the decision thresholds, and the corresponding value $F^{\mathrm{m}}(\theta'_1, \ldots, \theta'_m; S)$. If no higher value of F^{m} can be attained by changing any single threshold, while keeping all the other $m - 1$ ones at their current value, then $F^{\mathrm{m}}(\theta'_1, \ldots, \theta'_m; S) = \max_{\theta_1,\ldots,\theta_m} F^{\mathrm{m}}(\theta_1, \ldots, \theta_m; S)$.

Proposition 2 allows the *exact* solution of (29) to be found with low computational complexity. Indeed, it implies that the global maximum of F^{m} can be attained by starting from any threshold values, and iteratively updating one threshold at a time to *any* value that increases F^{m} (if any), until no further increase of F^{m} can be achieved. As a by-product, Algorithm 1 of [6] turns out to be one possible implementation of our optimization strategy above, provided that no early stopping condition is used, i.e., if the `repeat-until`. loop ends only when $F^{(t)} = F^{(t-1)}$. We also proved

[3] Recently, it has been shown that the optimal solution can also be obtained by solving a cost-sensitive problem with respect to the 2 m error counts FP_i and FN_i (see Eq. (7)) [23], but no algorithm has been developed so far to implement it. This approach will be described in Section 4.3.

that, if each threshold is initially set to $-\infty$, [4] then the exact solution of (29) is attained by considering at each step (e.g., in the argmax step of Algorithm 1) only higher values of each threshold than the current one [27]; this reduces the computational complexity to no more than $O(m^2n^2)$.

For the sake of completeness, we finally mention a similar approach that was considered in [8] (we mentioned it also in Section 4.2.1). It consists of independently learning an ensemble of K binary classifiers which output a real-valued score for each label, $f_{i,k}: \mathcal{X} \mapsto \mathbb{R}$, $i = 1, \ldots, m$, $k = 1, \ldots, K$. These classifiers are then linearly combined: $f_i(\mathbf{x}) = \sum_{k=1}^{K} w_k f_{i,k}(\mathbf{x}) + w_0$. In [8], F^{m} was maximized with respect to the combination weights, that do not depend on the label. To this aim, a non-convex approximation of all three M-L F measures was defined, by approximating the TP, FP and FN counts, on a given data set, using a logistic function of the scores f_i; a quasi-Newton optimization algorithm was then used.

4.3. Recent theoretical results about the single-label F measure

During the past two years several works have theoretically investigated the F measure maximization problem under the EUM approach, and have derived the optimal (Bayes) solution for the S-L F^{b}, either on a finite sample or at the population level. Novel maximization algorithms have also been developed, some of them based on the above mentioned theoretical results, and their consistency has been analyzed.

The optimal classifier at the population level has been derived in [19,37,13,15]. The corresponding expression of F^{b} can be obtained by replacing the TP, FP and FN counts in (4) with the corresponding probabilities, denoted as tp, fp and fn, and given by:

$$tp = \mathbb{P}(h(\mathbf{X}) = 1, Y = 1) = \mathbb{P}(Y = 1) \int_{\mathbf{x}: h(\mathbf{x})=1} p(\mathbf{x}|Y = 1)\mathrm{d}\mathbf{x} \tag{30}$$

$$fp = \mathbb{P}(h(\mathbf{X}) = 1, Y = 0) = \mathbb{P}(Y = 0) \int_{\mathbf{x}: h(\mathbf{x})=1} p(\mathbf{x}|Y = 0)\mathrm{d}\mathbf{x} \tag{31}$$

$$fn = \mathbb{P}(h(\mathbf{X}) = 0, Y = 1) = \mathbb{P}(Y = 1) \int_{\mathbf{x}: h(\mathbf{x})=0} p(\mathbf{x}|Y = 1)\mathrm{d}\mathbf{x} \tag{32}$$

Accordingly, $F^{\mathrm{b}}_{\beta} = \frac{(1+\beta^2)tp}{(1+\beta^2)tp + \beta^2 fn + fp}$. The optimal classifier h^* consists of thresholding the posterior probability:

$$h^*(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1|\mathbf{x}) \geq \theta^* \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

where $\theta^* = \frac{F^{\mathrm{c}*}_{\beta}}{1+\beta^2}$, and $F^{\mathrm{c}*}_{\beta}$ is the maximum S-L F. Since $F^{\mathrm{c}*}_{\beta}$ is unknown in practice, also θ^* is unknown. Note also that θ^* is a population-dependent value, i.e., the optimal decision whether labeling any instance as relevant or non-relevant depends not only on that instance, but also on all the other instances on which F^{b} is computed, as already pointed out in Section 2.[5] Actually, this is another way to express the fact that F^{b} does not decompose over instances.

In practice, in the above mentioned works the optimal decision function (33) was approximated by first estimating the posterior $\mathbb{P}(Y = 1|\mathbf{x})$, and then tuning the decision threshold on validation data. The results in [7] allow to approximate it using a different procedure based on the ROC curve of an underlying binary classifier. It amounts to thresholding $\frac{\mathbb{P}(Y=1|\mathbf{x})}{F^{\mathrm{c}*}_{\beta}}$ at $\frac{1}{1+\beta^2}$, where the calibrated estimate of $\mathbb{P}(Y = 1|\mathbf{x})$ and the estimate of $F^{\mathrm{c}*}_{\beta}$ can be obtained from the ROC convex hull; in this case, the threshold depends only on β.[6]

Note that all the above results also apply to F^{M}, whose optimal classifier is obtained by independently using (33) for each label; in this case, the optimal threshold can be different for each label.

An alternative solution was obtained in [23]: it was shown that the optimal classifier, both at the population level or on a finite sample, can be obtained by reduction to a cost-sensitive problem. Such a problem consists of minimizing the expected weighted error given by a linear combination of the fp and fn probabilities of each label (or the corresponding FP and FN counts), for suitable costs. Analogously to rule (33), such costs depend on the maximum F^{b}, and thus are unknown in practice. This implies that the optimal solution can be obtained by wrapping a cost-sensitive classification algorithm in an inner loop by an outer loop that sets the appropriate costs [23]. Although this requires in principle to solve an infinite series of cost-sensitive problems, it was shown that the cost space can be discretized to approximate the optimal solution with a desired accuracy level, by choosing the costs that provide the maximum F^{b} value a posteriori. Interestingly, similar results were derived in [23] for the M-L F^{M} and F^{m}.

We finally summarize recent results about algorithms for maximizing F^{b} (and thus also the M-L F^{M}).

The theoretical result of [23] mentioned above was applied in the same work to existing cost-sensitive algorithms for binary problems. Interestingly, their results apply also to the M-L F^{m}; however, exploiting them to develop specific cost-sensitive algorithms for this measure is not straightforward, since the FP and FN counts of each class are *simultaneously* involved, and was left in [23] as a future work.

In [13,22] the consistency of "plug-in" algorithms for maximizing F^{b}, consisting of thresholding an estimate of the posterior $\mathbb{P}(Y = 1|\mathbf{x})$, and of empirically computing the threshold value, was investigated. In [13] a different two-step approach was also considered ("Weighted Empirical Risk Minimization"), based on a theoretical result analogous to the one of [23]. In the first step a classifier with real-valued predictions $f(\mathbf{x})$ is learnt by minimizing a surrogate weighted loss with label-dependent costs, defined as

$$\ell(f(\mathbf{x}), y) = (1 - \delta)\mathbb{I}[y = 1]\ell(f(\mathbf{x}), 1) + \delta\mathbb{I}[y = 0]\ell(f(\mathbf{x}), 0), \tag{34}$$

which is known to be consistent with the (ideal) classifier given by $\mathrm{sign}\big(\mathbb{P}[Y = 1|\mathbf{X}] - \delta\big)$. In the second step the empirical F^{b} is maximized with respect to δ. This algorithm is computationally less demanding than the one of [23], since it only requires a single loop to scan the values of δ.

In [12] a similar two-step approach as the above Weighted Empirical Risk Minimization was investigated. Different possible surrogate loss functions were considered to learn the classifier at the first step, among strongly proper composite loss functions, such as logistic, squared-error, and exponential loss. The results provided in [12] are not limited to the consistency of the considered approach, as in [13], but are valid also for finite samples; in particular, it was shown that the regret of the considered classifier, measured with respect to the target metric, is upper bounded by the regret of the score function $f(\cdot)$ measured with respect to the surrogate loss.

A different algorithm was developed in [20], based on point-based stochastic updates, and in particular on stochastic alternate maximization. For the sake of completeness we also mention that in [21] some algorithms were developed for maximizing versions

[4] In practice, if $\theta_i^{(0)} < \min_{\mathbf{x} \in \mathcal{S}} f_i(\mathbf{x})$.

[5] In [14] it had been already shown that the rule $\mathrm{sign}[\mathbb{P}(Y = 1|\mathbf{x}) - \theta]$, where θ is any *fixed* threshold value, can not be optimal.

[6] This result has been suggested by one of the reviewers.

of the macro- and micro-averaged F defined for multi-class S-L problems, which are different from the M-L versions considered in this paper. In particular, we point out that if the micro-averaged F of Eq. (7) (which is different from the one considered in [21]) is used in a multi-class S-L problem, it reduces to classification accuracy [16].

Finally, it is worth pointing out that most of the above results apply to broad classes of performance measures based on ratios of TP, FN and FP counts, beside the F measures.

4.4. Optimal classifier for the multi-label micro-averaged and instance-wise F

Here we show that the above mentioned results of [19,37,13,15] on the S-L F^{b} can be exploited to derive the optimal classifier at the population level, under the EUM approach, also for the M-L F^{m}.[7] To this aim, we follow an analogous proof procedure as the one in [15]. We then derive also the optimal classifier for F^{i}. As mentioned above, whereas the optimal classifier for the S-L F^{b}, and for the M-L F^{M} and F^{m}, can also be obtained by reduction to cost-sensitive problems, no analogous solution is known for the F^{i} [23].

Micro-averaged *F:* Our result is given by the following proposition.

Proposition 3. The optimal classifier at the population level for F_β^{m} consists of deciding $h_i(\mathbf{x}) = 1$, if and only if:

$$\mathbb{P}(Y_i = 1|\mathbf{x}) \geq \frac{F_\beta^{*\mathrm{m}}}{1+\beta^2}, \tag{35}$$

where $F_\beta^{*\mathrm{m}}$ is the optimal value of F_β^{m}.

Proof. Assume that the optimal decisions for all labels have already been found on the whole instance space $\mathcal{X}$, except for the k-th label in a region $\Delta \subset \mathcal{X}$ around a given $\mathbf{x}^*$. Now we write the F_β^{m} at the population level, by separating the contribution of the decision $h_k(\mathbf{x})$ on Δ. Using Eq. (30), the term at the numerator of the empirical F_β^{m} of Eq. (7) corresponding to $\sum_{i=1}^m TP_i$, minus the contribution of $h_k(\mathbf{x})$ on Δ, is given by the following expression, which we denote again as tp for the sake of simplicity:

$$\begin{aligned} tp = &\sum_{i=1, i\neq k}^m \mathbb{P}(Y_i = 1) \int_{\mathbf{x}\in\mathcal{X}: h_i(\mathbf{x})=1} p(\mathbf{x}|Y_i = 1)\mathrm{d}\mathbf{x} \\ &+ \mathbb{P}(Y_k = 1) \int_{\mathbf{x}\in\mathcal{X}-\Delta: h_k(\mathbf{x})=1} p(\mathbf{x}|Y_k = 1)\mathrm{d}\mathbf{x}. \end{aligned} \tag{36}$$

The terms corresponding to $\sum_{i=1}^m FP_i$ and $\sum_{i=1}^m FN_i$ in Eq. (7) can be written similarly, using Eqs. (31) and (32); we denote them respectively as fp and fn. To keep the following expressions simple, we also write:

$$\begin{aligned} &b_k = \mathbb{P}(Y_k = 1), \\ &P_{1k}(\Delta) = \int_{\mathbf{x}\in\Delta} p(\mathbf{x}|Y_k = 1)\mathrm{d}\mathbf{x}, \\ &P_{0k}(\Delta) = \int_{\mathbf{x}\in\Delta} p(\mathbf{x}|Y_k = 0)\mathrm{d}\mathbf{x}. \end{aligned} \tag{37}$$

The value of F_β^{m} can now be written by considering the two possible choices for $h_k(\mathbf{x})$, $\mathbf{x} \in \Delta$. By choosing $h_k(\mathbf{x}) = 1$, we get:

$$F_\beta'^{\mathrm{m}} = \frac{(1+\beta^2)\big[\, tp + b_k P_{1k}(\Delta)\big]}{(1+\beta^2)\big[\, tp + b_k P_{1k}(\Delta)\big] + \beta^2 fn + fp + (1-b_k)P_{0k}(\Delta)}. \tag{38}$$

By choosing $h_k(\mathbf{x}) = 0$, instead, we get:

$$F_\beta''^{\mathrm{m}} = \frac{(1+\beta^2)tp}{(1+\beta^2)tp + \beta^2\big[fn + \beta^2 b_k P_{1k}(\Delta)\big] + fp}. \tag{39}$$

Accordingly, the optimal decision rule for the k-th label in Δ is $h_k(\mathbf{x}) = 1$, if and only if $F_\beta'^{\mathrm{m}} \geq F_\beta''^{\mathrm{m}}$. After some algebraic manipulations, this amounts to:

$$\frac{b_k P_{1k}(\Delta)}{(1-b_k)P_{0k}(\Delta)} \geq \frac{tp}{\beta^2 tp + \beta^2 fn + fp + \beta^2 b_k^2 P_{1k}^2(\Delta)}. \tag{40}$$

Let us now take the limit $\Delta \to \{\mathbf{x}^*\}$. The left-hand side of inequality (40) becomes (see also Eq. (37)):

$$\begin{aligned} \lim_{\Delta\to\{\mathbf{x}^*\}} \frac{b_k P_{1k}(\Delta)}{(1-b_k)P_{0k}(\Delta)} &= \lim_{\Delta\to\{\mathbf{x}^*\}} \frac{\int_{\mathbf{x}\in\Delta}\mathbb{P}(Y_k = 1|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}}{\int_{\mathbf{x}\in\Delta}\mathbb{P}(Y_k = 0|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}} \\ &= \frac{\mathbb{P}(Y_k = 1|\mathbf{x}^*)}{\mathbb{P}(Y_k = 0|\mathbf{x}^*)}. \end{aligned} \tag{41}$$

Since $\lim_{\Delta\to\{\mathbf{x}^*\}} P_{1k}^2(\Delta) = 0$, for the right-hand side of inequality (40) we get:

$$\lim_{\Delta\to\{\mathbf{x}^*\}} \frac{tp}{\beta^2 tp + \beta^2 fn + fp + \beta^2 b_k^2 P_{1k}^2(\Delta)} = \frac{tp^*}{\beta^2 tp^* + \beta^2 fn^* + fp^*}, \tag{42}$$

where tp^* denotes the value of Eq. (36) computed in the whole instance space $\mathcal{X}$ (except for the zero-measure element $\mathbf{x}$), corresponding to the optimal micro-averaged F, and similarly for fn^* and fp^*. Finally, taking into account that $\mathbb{P}(Y_k = 0|\mathbf{x}) = 1 - \mathbb{P}(Y_k = 1|\mathbf{x})$, after some algebraic manipulations on Eqs. (41) and (42) we obtain the claimed optimal decision rule:

$$\mathbb{P}(Y_k = 1|\mathbf{x}^*) \geq \frac{tp^*}{(1+\beta^2)tp^* + \beta^2 fn^* + fp^*} = \frac{F_\beta^{*\mathrm{m}}}{1+\beta^2}. \tag{43}$$

Instance-wise *F:* In this case the optimal classifier is given by the following proposition. □.

Proposition 4. For a given instance $\mathbf{x}$, the optimal classifier at the population level for F_β^{i} consists of deciding $h_i(\mathbf{x}) = 1$ for the m^* labels exhibiting the highest posteriors, and $h_i(\mathbf{x}) = 0$ to the remaining ones, where $0 \leq m^* \leq m$ is given by:

$$m^* = \arg\max_{k\in\{0,\ldots,m\}} \frac{(1+\beta^2)\sum_{i=0}^k \mathbb{P}(Y_{(i)} = 1|\mathbf{x})}{k + \beta^2 \sum_{i=1}^m \mathbb{P}(Y_i = 1|\mathbf{x})}, \tag{44}$$

where we write $\mathbb{P}(Y_{(0)} = 1|\mathbf{x}) = 0$, and $Y_{(1)}, \ldots, Y_{(m)}$ denote the labels sorted for decreasing values of the posteriors $\mathbb{P}(Y_i = 1|\mathbf{x})$.

Proof. Since F^{i} is computed on a single instance $\mathbf{x}$ (see Eq. (5)), its probabilistic definition involves only the posteriors $\mathbb{P}(Y_i|\mathbf{x})$. For ease of notation, let $P = \{i\colon h_i(\mathbf{x}) = 1\}$, $N = \{i\colon h_i(\mathbf{x}) = 0\}$, $P_{i1} = \mathbb{P}(Y_i = 1|\mathbf{x})$, and $P_{i0} = \mathbb{P}(Y_i = 0|\mathbf{x})$. We then have:

$$F_\beta^{\mathrm{i}} = \frac{(1+\beta^2)\sum_{i\in P} P_{i1}}{(1+\beta^2)\sum_{i\in P} P_{i1} + \beta^2 \sum_{i\in N} P_{i1} + \sum_{i\in P} P_{i0}}. \tag{45}$$

Since $\sum_{i\in P}\big(P_{i1} + P_{i0}\big) = |P|$, and $\sum_{i\in P} P_{i1} + \sum_{i\in N} P_{i1} = \sum_{i=1}^m P_{i1}$, we get:

$$F_\beta^{\mathrm{i}} = \frac{(1+\beta^2)\sum_{i\in P} P_{i1}}{|P| + \beta^2 \sum_{i=1}^m P_{i1}}. \tag{46}$$

Note that, for any given $|P| > 0$, Eq. (46) is maximized by deciding $h_i(\mathbf{x}) = 1$ for the $|P|$ labels exhibiting the highest posteriors P_{i1}, and $h_i(\mathbf{x}) = 0$ for the remaining labels. It immediately follows that F_β^{i} is maximized by the decision rule claimed above. □

5. Decision-theoretic approach

F measure maximization algorithms based on DTA have been

[7] This result has been suggested by one of the reviewers of a previous version of this paper.

proposed so far for F^{b} and F^{i}. We point out that no specific algorithm for F^{M} has been developed under this approach, since the same algorithms for F^{b} can be applied, independently for each label (under the usual assumption of i.i.d. instances). No algorithm based on the DTA has been developed yet for F^{m}, instead; we shall fill this gap in our companion paper [28].

5.1. Single-label F

The label assignment that maximizes the expected value $\mathbb{E}[F^{\mathrm{b}}_{\beta}]$ on a given set of instances $\mathbf{x}^1, \ldots, \mathbf{x}^n$ is given by:

$$(h^{*1}, \ldots, h^{*n}) = \arg \max_{(h^1,\ldots,h^n)\in\{0,1\}^n} \sum_{y^1,\ldots,y^n\in\{0,1\}^n} \mathbb{P}(y^1, \ldots, y^n|\mathbf{x}^1, \ldots, \mathbf{x}^n) \frac{(1+\beta^2)\sum_{j=1}^n y^j h^j}{\beta^2 \sum_{j=1}^n y^j + \sum_{j=1}^n h^j}. \tag{47}$$

If the labels are conditionally independent, i.e., $\mathbb{P}(Y^1, \ldots, Y^n|\mathbf{x}^1, \ldots, \mathbf{x}^n) = \prod_{j=1}^n \mathbb{P}(Y^j|\mathbf{x}^j)$, then only up to $n2^n$ combinations of true and assigned labels need to be evaluated, out of all possible $2^{2\,n}$ combinations [14]. This is because each summand in (47) (i.e., the value of F^{b}_{β} for fixed $y^1, \ldots, y^n$) is maximized by one of the n label assignments in which the label 1 is given to the $n' \le n$ instances exhibiting the n' highest posteriors $\mathbb{P}(Y^j = 1|\mathbf{x}^j)$, for some $n' \in \{0, \ldots, n\}$ [14].

Exact inference algorithms with lower computational complexity, under the assumption of conditionally independent labels, were subsequently derived in [1], with $O(n^3)$ complexity; in [10], with $O(n^4)$ complexity; and in [19], with $O(n^3)$ complexity, which reduces to $O(n^2)$ time and $O(n)$ space complexity when β^2 is rational. We report this latter procedure as Algorithm 2, as it is the one with lowest computational complexity. It provides the $n+1$ values of the expected F^{b}_{β}, denoted as $f_{\beta,0}, \ldots, f_{\beta,n}$, corresponding to assigning $h_j{=}1$ to the n' instances exhibiting the highest posteriors, for $n' \in \{0, \ldots, n\}$. The optimal label assignment is the one corresponding to the highest $f_{\beta,n'}$.

Algorithm 2. Inference algorithm for maximizing $\mathbb{E}[F^{\mathrm{b}}_{\beta}]$ for a rational β^2 [19].

Input: p and q, where $\beta^2 = p/q$; the posteriors $p_j := \mathbb{P}(Y^j = 1|\mathbf{x}^j)$, $j = 1, \ldots, n$
Output: the values $f_{\beta,1}, \ldots, f_{\beta,n}$
  for $0 \le j \le n$, set $C[j]$ as the coefficient of z^j in the polynomial $[p_1 z + (1-p_1)]\ldots[p_n z + (1-p_n)]$
  $S[j] \leftarrow q/j$, $j = 1, \ldots, (q+r)n$
  for $n' = n$ to 1 **do**
    $f_{\beta;n'} \leftarrow \sum_{k_1=0}^{n} (1 + r/q) k_1 C[k_1] S[rk + qk_1]$
    for $j = 1$ to $(q+r)(n'-1)$ **do**
      $S[j] \leftarrow (1 - p_{n'})S[j] + p_{n'} S[j+q]$
    end for
  end for
  return $f_{\beta,1}, \ldots, f_{\beta,n}$

In the most general case when the independence assumptions does not hold, one can use the exact inference algorithm developed in [5] for F^{i} (described in Section 5.2), with $O(n^3)$ complexity. This is possible because the expression of F^{b} (4), and thus problem (47), are formally identical respectively to the expression of F^{i} (see Eq. (5) and to problem (48)).

5.2. Instance-wise F

The label assignment $\mathbf{h}^*$ that maximizes $\mathbb{E}[F^{\mathrm{i}}_{\beta}]$ for a given instance $\mathbf{x}$ is given by:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}\in\mathcal{Y}} \sum_{\mathbf{y}\in\mathcal{Y}} \mathbb{P}(\mathbf{y}|\mathbf{x}) \frac{(1+\beta^2)\sum_{i=1}^m y_i h_i}{\beta^2 \sum_{i=1}^m y_i + \sum_{i=1}^m h_i}. \tag{48}$$

This problem has the same form as (47), but in this case the assumption of conditionally independent labels is not realistic, and thus the inference algorithms of [14,1,10,19] do not provide the exact solution. An exact solution with $O(m^3)$ complexity was derived in [5,4,2,35]. It does not require the knowledge of the full distribution $\mathbb{P}(\mathbf{Y}|\mathbf{x})$, but only of the $m^2 + 1$ probabilities $\mathbb{P}(\mathbf{Y} = \mathbf{0}|\mathbf{x})$ and $\mathbb{P}(Y_i = 1, S_{\mathbf{Y}} = s|\mathbf{x})$, $i, s = 1, \ldots, m$, where $\mathbf{0} = \{0\}^m$, and $s_{\mathbf{Y}} = \sum_{i=1}^m Y_i$ is the number of classes $\mathbf{x}$ is relevant to. The inference algorithm of [5] consists of two nested maximization steps,[8] and is reported as Algorithm 3, where P and W denote the $m \times m$ matrices defined as:

$$P_{i,s} = \mathbb{P}(Y_i = 1, S_{\mathbf{Y}} = s|\mathbf{x}), \quad i, s = 1, \ldots, m, \tag{49}$$

$$W_{i,k} = \frac{1}{\beta^2 i + k}, \quad i, k = 1, \ldots, m, \tag{50}$$

and F denotes the matrix PW. It is worth noting that, when the assumption of conditionally independent labels does not hold, the difference between the exact solution provided by Algorithm 3 and the ones provided by algorithms based on such an assumption can theoretically become arbitrarily large [5].

Algorithm 3. The inference algorithm of [5] for maximizing $\mathbb{E}[F^{\mathrm{i}}]$

Input: the matrix P of Eq.(49) and the value of $\mathbb{P}(\mathbf{Y} = \mathbf{0}|\mathbf{x})$
Output: the label assignment of Eq.(48)
  compute the matrix W defined by Eq.(50)
  $F \leftarrow PW$, $\mathbf{h}^{(0)} \leftarrow \mathbf{0}$, $E_0 \leftarrow \mathbb{P}(\mathbf{Y} = \mathbf{0}|\mathbf{x})$
  for $k = 1$ to m **do**
    set $\mathbf{h}^{(k)}$ such that $h_i{=}1$ for the top-k elements $F_{i,k}$ in the k-th column of F,
      and $h_i{=}0$ for the other elements
    $E_k \leftarrow 2\sum_{i=1}^m h_i^{(k)} F_{i,k}$
  end for
  $q \leftarrow \mathrm{argmax}_{k=0,\ldots,m} E_k$
  return $\mathbf{h}^{(q)}$

In [5,2,35], $\mathbb{P}(\mathbf{Y} = \mathbf{0}|\mathbf{x})$ and P are estimated by sampling from the distribution $\mathbb{P}(\mathbf{Y}|\mathbf{x})$. The latter is in turn estimated using the Probabilistic Classifier Chains (PCC) method [3], which exploits the product rule of probability:

$$\mathbb{P}(\mathbf{Y}|\mathbf{x}) = \prod_{i=1}^m \mathbb{P}(Y_i|\mathbf{x}, y_1, \ldots, y_{i-1}), \tag{51}$$

and learns m probabilistic classifiers that independently estimate each of the m terms in the right-hand side of Eq. (51). To this aim, linear regularized logistic regression was used in [2,5]. In [4,35] $\mathbb{P}(\mathbf{Y} = \mathbf{0}|\mathbf{x})$ and P were estimated using a reduction approach, by independently solving the following $m+1$ S-L multi-class probability estimation problems. For a given $i \in \{1, \ldots, m\}$, $\mathbb{P}(Y_i = 1, S_{\mathbf{Y}}|\mathbf{x})$ can be rewritten as $\mathbb{P}(Y'_i|\mathbf{x})$ by defining a new

[8] This approach had already been proposed in [10], but their inference algorithm exhibits a much higher computational complexity.

random variable $Y'_i = \mathbb{I}[Y_i = 1] \times S_{\mathbf{Y}} \in \{0, \ldots, m\}$; then $\mathbb{P}(Y'_i|\mathbf{x})$ can be estimated, e.g., using multinomial regression. Similarly, $\mathbb{P}(\mathbf{Y} = \mathbf{0}|\mathbf{x})$ is obtained by a reduction to a binary problem associated to a random variable $Y' = \mathbb{I}[\mathbf{Y} = \mathbf{0}] \in \{\mathbf{0}, \mathbf{1}\}$, by estimating $\mathbb{P}(Y'|\mathbf{x})$. On the one hand, the latter approach avoids a computationally demanding sampling step; on the other hand, it produces non-calibrated probabilities; a post-processing step is thus required, or additional constraints have to be included in the above learning problems [4].

A different approach was proposed in [29], focused on the following decision rule:

$$h_i = \begin{cases} 1, & \text{if } \mathbb{P}(Y_i = 1|\mathbf{x}) \geq \theta(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases} \qquad i = 1, \ldots, m \tag{52}$$

It labels a sample as relevant to the labels whose marginal posterior exceeds a threshold $\theta(\mathbf{x})$ which depends on the sample itself (equivalently, to the labels exhibiting the top-$k(\mathbf{x})$ values of $\mathbb{P}(y_i|\mathbf{x})$, where the value $k(\mathbf{x})$ depends again on the sample). A dynamic programming strategy with $O(m^3)$ complexity was proposed to find the value of $\theta(\mathbf{x})$ (or $k(\mathbf{x})$) that maximizes the expectation in Eq. (48). Note however that, if the labels are not coernditionally independent, the decision rule (52) is naot guaranteed to provide the global maximum of $\mathbb{E}[F^i]$ [14].

6. Conclusions

We provided a unifying, comprehensive survey of the existing approaches and algorithms aimed at maximizing the F measures in multi-label classification problems. We believe this is a useful contribution for further developments in this field, due to the increasing interest on applications related to information retrieval, and on the corresponding measures of classification accuracy, from both the pattern recognition and machine learning research communities.

Works published over the past few years considerably improved the knowledge about F measures, and provided theoretically-grounded algorithms for their optimization. The optimal (Bayes) classifier at the population level is now known both for the S-L and for all three M-L F measures; in particular, the ones for the M-L micro-averaged and instance-wise F were explicitly derived in this paper. An equivalent solution based on a reduction to cost-sensitive problems is also known, except for the M-L, instance-wise F, and algorithms based on this approach have already been derived for the S-L F and the M-L, macro-averaged F. Different maximization algorithms have also been proposed for all these measures, and the consistency of some of them has been proven. Only for the M-L micro-averaged F relatively fewer solutions are available: under the empirical utility maximization approach, only maximization algorithms that tune the decision thresholds of binary classifiers are known, and no maximization algorithm based on the decision-theoretic approach has been derived so far. We shall fill these gaps in our companion paper [28].

Acknowledgments

This work has been supported by a grant No. CRP-59872 funded by Regione Autonoma della Sardegna, L.R. 7/2007, Bando 2012. We thank the reviewers of a previous version of this paper for their detailed and constructive comments.

References

[1] K.M.A. Chai, Expectation of F-measures: Tractable exact computation and some empirical observations of its properties, in: Proc. Int. ACM SIGIR Conf. Research and Development in Inf. Retr., ACM., 2005, pp. 593–594.
[2] W. Cheng, K. Dembczyński, E. Hüllermeier, A. Jaroszewicz, W. Waegeman, F-measure maximization in topical classification, in: Rough Sets and Current Trends in Computing, LNCS vol. 7413, Springer, 2012, pp. 439–446.
[3] K. Dembczyński, W. Cheng, E. Hüllermeier, Bayes optimal multilabel classification via probabilistic classifier chains, in: Proc. Int. Conf. Machine Learning, 2010, pp. 279–286.
[4] K. Dembczyński, A. Jachnik, W. Kotlowski, W. Waegeman, E. Hüllermeier, Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization, in: Proc. Int. Conf. Machine Learning, 2013.
[5] K. Dembczyński, W. Waegeman, W. Cheng, E. Hüllermeier, An Exact Algorithm for F-Measure Maximization, in: Neural Inf. Proc. Systems, 2011, pp. 1404–1412.
[6] R.-En Fan, C.-J. Lin, A Study on Threshold Selection for Multi-label Classification, Tech. rep., National Taiwan University, 2007.
[7] P. Flach, M. Kull, Precision-Recall-Gain Curves: PR Analysis Done Right, in: Neural Inf. Proc. Systems, 2015, pp. 838–846.
[8] A. Fujino, H. Isozaki, J. Suzuki, Multi-label Text Categorization with Model Combination based on F1-score Maximization, in: Proc. Int. Joint Conf. Natural Language Proc., 2008.
[9] M. Jansche, Maximum expected F-measure training of logistic regression models, in: Proc. Int. Conf. on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 692–699.
[10] M. Jansche, A maximum expected utility framework for binary sequence labeling, in: Proc. Annual Meeting of the Association of Computational Linguistics, 2007, pp. 736–743.
[11] T. Joachims, A Support Vector Method for Multivariate Performance Measures, in: Proceedings of Int. Conf. on Machine Learning, 2005, pp. 377–384.
[12] W. Kotlowski, K. Dembczyński. Surrogate regret bounds for generalized classification performance metrics, CoRR 2015, abs/1504.07272.
[13] O. Koyejo, N. Natarajan, P.K. Ravikumar, I.S. Dhillon, Consistent binary classification with generalized performance metrics, in: Proceedings of the Adv. in Neural Inf. Proc. Systems 27, 2014, pages 2744–2752.
[14] D.D. Lewis, Evaluating and optimizing autonomous text classification systems, in: Proc. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, ACM, 1995, pp. 246–254.
[15] Z.C. Lipton, C. Elkan, B. Narayanaswamy, Optimal thresholding of classifiers to maximize F1 measure, in: Proc. Machine Learning and Knowledge Discovery in Databases – European Conf., ECML PKDD, Part II, LNCS vol. 8725, Springer, 2014, pp. 225–239.
[16] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
[17] M. Di Martino, G. Hernández, M. Fiori, A. Fernández, A new framework for optimal classifier design, Pattern Rec. 46 (8) (2013) 2249–2255.
[18] D.R. Musicant, V. Kumar, A. Ozgur, Optimizing F-measure with Support Vector Machines, in Int. Florida Artif. Int. Research Society Conf., AAAI Press, 2003, pages 356–360.
[19] Ye Nan, K.M.A. Chai, W.S. Lee, H.L. Chieu, Optimizing F-measure: A tale of two approaches, in: Proc. Int. Conf. on Machine Learning, 2012.
[20] H. Narasimhan, P. Kar, P. Jain, Optimizing non-decomposable performance measures: a tale of two classes, in: Proc. Int. Conf. Machine Learning, JMLR Proceedings vol. 37, 2015, pp. 199–208.
[21] H. Narasimhan, H.G. Ramaswamy, A. Saha, S. Agarwal, Consistent multiclass algorithms for complex performance measures, in: Proc. Int. Conf. Machine Learning, JMLR Proceedings vol. 37, 2015, pp. 2398–2407.
[22] H. Narasimhan, R. Vaish, S. Agarwal, On the statistical consistency of plug-in classifiers for non-decomposable performance measures, in: Adv. in Neural Inf. Proc. Syst. 27, 2014, pp. 1493–1501.
[23] S.P. Parambath, N. Usunier, Y. Grandvalet, Optimizing F-measures by cost-sensitive classification, in: Adv. in Neural Inf. Proc. Syst. 27, 2014, pp. 2123–2131.
[24] J. Petterson, T.S. Caetano, Reverse multi-label learning, Adv. Neural Inf. Process. Syst. 23 (2010) 1912–1920.
[25] J. Petterson, T.S. Caetano, Submodular multi-label learning, Adv. Neural Inf. Process. Syst. 24 (2011) 1512–1520.
[26] I. Pillai, G. Fumera, F. Roli, F-Measure Optimisation in Multi-label Classifiers, in Int. Conf. Pattern Recognition, 2012.
[27] I. Pillai, G. Fumera, F. Roli, Threshold optimisation for multi-label classifiers, Pattern Recognit. 46 (7) (2013) 2055–2065.
[28] I. Pillai, G. Fumera, F. Roli, Designing multi-label classifiers that maximize the micro-averaged F measure (Submitted for publication).
[29] J.R. Quevedo, O. Luaces, A. Bahamonde, Multilabel classifiers with a probabilistic thresholding strategy, Pattern Rec. 45 (2) (2012) 876–883.
[30] C.J. Van Rijsbergen, Foundation of evaluation, J. Doc. 30 (4) (1974) 365–373.
[31] I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun, Support vector machine learning for interdependent and structured output spaces, in: Proc. Int. Conf. Machine Learning, 2004, page 104.

[32] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. 6 (2005) 1453–1484.
[33] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data Mining and Knowledge Discovery Handbook, 2010, pp. 667–685.
[34] C.J. van Rijsbergen, Information Retrieval, Butterworth, London, UK, 1979.
[35] W. Waegeman, K. Dembczyński, A. Jachnik, W. Cheng, E. Hüllermeier, On the Bayes-optimality of F-measure maximizers, J. Mach. Learn. Res. 15 (2014) 3333–3388.
[36] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[37] M.-J. Zhao, N.U. Edakunni, A. Pocock, G. Brown, Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications, J. Mach. Learn. Res. 14 (1) (2013) 1033–1090.

Ignazio Pillai is a post-doc at the Pattern Recognition and Applications Lab led by Prof. Roli. His research interests include multimedia document categorization and face verification. He is a member of the IAPR.

Giorgio Fumera is an Associate Professor of Computer Engineering. His research interests are related to statistical pattern recognition, multiple classifier systems and adversarial classification, with applications to document categorization and person re-identification. He is a member of the IEEE and IAPR.

Fabio Roli is a Full Professor of Computer Engineering. His research over the past twenty years addressed the design of pattern recognition systems in real applications. He played a leading role for the research field of multiple classifier systems. He is Fellow of the IEEE and Fellow of the IAPR.