Classifier Selection Approaches for Multi-label Problems

Ignazio Pillai, Giorgio Fumera, and Fabio Roli

Deparment of Electrical and Electronic Engineering, Univ. of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{pillai,fumera,roli}@diee.unica.it

Abstract. While it is known that multiple classifier systems can be effective also in multi-label problems, only the classifier fusion approach has been considered so far. In this paper we focus on the classifier selection approach instead. We propose an implementation of this approach specific to multi-label classifiers, based on selecting the outputs of a possibly *different* subset of multi-label classifiers for each class. We then derive static selection criteria for the macro- and micro-averaged F measure, which is widely used in multi-label problems. Preliminary experimental results show that the considered selection strategy can exploit the complementarity of an ensemble of multi-label classifiers more effectively than selection approaches analogous to the ones used in single-label problems, which select the outputs of the *same* classifier subset for all classes. Our results also show that the derived selection criteria can provide a better trade-off between the macro- and micro-averaged F measure, despite it is known that an increase in either of them is usually attained at the expense of the other one.

Keywords: Multi-label classification, Multiple classifier systems, Classifier selection

1 Introduction

In multi-label classification problems each sample can belong to more than one class, contrary to traditional, single-label problems. Multi-label problems occur in several applications related to retrieval tasks, like text categorisation, image annotation, protein function classification and music classification, and are receiving an increasing interest from the pattern recognition and machine learning literature. So far, several works have shown that multiple classifier systems (MCSs) can be effectively exploited to improve classification performance also in multi-label problems [7, 6, 9, 10]. In [9] it was also claimed that MCSs can be useful to deal with imbalanced class distribution, which often occurs in multi-label problems.

To our knowledge, all previous works considered the classifier *fusion* approach, which consists in the combination of the outputs of *all* the available classifiers. In this work we focus on the classifier *selection* approach instead. We

first present in Sect. 2 a selection approach specific to multi-label classifiers. It is based on the selection of the output of a possibly *different* subset of multi-label classifiers for each class. This potentially allows one to better exploit the complementarity between the available multi-label classifiers on the different classes. We then focus in Sect. 4 on static selection, and develop selection criteria based on the macro- and micro-averaged Van Rijsbergen's F measure, which is a widely used performance measure in multi-label problems [2, 8, 9, 11]. The F measure is described in detail in Sect. 3. From the analysis of the proposed selection criteria, we argue that our method may also provide a good trade-off between the macro- and micro-averaged F measure. This is an interesting result, as it is known that an improvement in the macro-averaged F measure can be usually be attained only at the expense of the micro-averaged one, and vice versa [13]. In Sect. 5 we give an empirical evaluation of the proposed static classifier selection approach on three multi-label data sets related to text categorisation and gene functionality classification tasks. Conclusions are finally drawn in Sect. 6.

2 A classifier selection approach for multi-label problems

In the following we denote with $\Omega = \{\omega_1, \ldots, \omega_N\}$ the set of classes of a given problem, and the feature vector of a sample as $\mathbf{x} \in X \subseteq \mathbb{R}^n$, where n is the size of the feature space X. In a single-label problem each sample belongs to exactly one class, and a classifier implements a decision function $f : X \to \Omega$.

Given an ensemble of single-label classifiers $f^1, \ldots, f^L$, the rationale of selection methods is that different classifiers can be more accurate than others in different regions of the feature space [12, 4]. Accordingly, given a testing sample $\mathbf{x}$, these methods aim to select one of the classifiers that correctly classify $\mathbf{x}$ (if any). This can be done "statically", by defining at the design phase the so-called region of competence of each classifier in the feature space. Each testing sample is then labelled by the classifier associated to the region in which it falls. Classifier selection can also be done "dynamically". In this case, the classifier which exhibits the highest accuracy in a neighbourhood of the testing sample is selected. Such "local" accuracy is estimated online, at the classification phase. To the scope of this work, a somewhat related approach is the selection of a *subset* of classifiers to be fused, out of a larger ensemble. It is based on a similar rationale as the one above: a different subset of classifiers can be more effective than the whole ensemble, on different testing samples [14]. This approach is usually implemented statically.

As in multi-label problems each sample can belong to more than one class, a multi-label classifier implements a decision function $f : X \to \{+1, -1\}^N$, where the value $+1$ (-1) in the k-th element of the vector $f(\mathbf{x})$ means that the sample $\mathbf{x}$ is labelled as (not) belonging to ω_k. The classifier (subset) selection approaches described above can be used with multi-label classifiers as well. In addition, for multi-label classifiers it is possible to implement a different selection approach. Consider an ensemble of multi-label classifiers $f^1, \ldots, f^L$, and let us denote with $f_k^i(\mathbf{x})$ the output of the i-th classifier for the k-th class. Instead of

selecting the same subset of L' classifiers ($1 \leq L' < L$) to label a testing sample $\mathbf{x}$, as happens in single-label problems, one can also select a *different* subset of L' classifiers for each class. In other words, the decision whether $\mathbf{x}$ belongs or not to ω_k is taken by combining the outputs $f_k^{i_1(k)}, f_k^{i_2(k)}, \ldots, f_k^{i_{L'}(k)}$, where a different subset $\{i_1(k), i_2(k), \ldots, i_{L'}(k)\} \subset \{1, \ldots, L\}$ can be chosen for each ω_k. We will denote this approach as "Hybrid Selection of Multi-label classifiers" (HSM), where "hybrid" refers to the selection a possibly different classifier subset for each class.

To implement the HSM approach, a seemingly reasonable criterion is to select for each class the subset of L' classifiers which exhibit the highest classification performance *on that class* (either a local performance measure around a given testing sample, in the case of dynamic selection, or the performance in predefined regions of the feature space, in the case of static selection). However, as we will show in the next sections, it turns out that this criterion is not always suitable to maximise the *overall* classification performance of a multi-label classifier evaluated on all classes. This is due to the particular performance measures usually used in multi-label problems, which are based on precision and recall. In particular, the above criterion is suitable for macro-averaged performance measures, but not for micro-averaged ones. This implies that, to maximise a micro-averaged performance measure, an exhaustive search over all the possible $\binom{L}{L'}^N$ choices of L' out of L classifiers for each of the N classes is required, which is clearly impractical. To address this issue, we first describe in detail performance measures based on precision and recall in Sect. 3, and then derive a suboptimal selection criterion for micro-averaged measures in Sect. 4.

3 Performance measures for multi-label classifiers

The performance of multi-label classifiers is measured in terms of precision and recall, as multi-label problems usually occur in retrieval tasks. In the field of information retrieval, precision and recall are respectively defined as the probability that a retrieved document is relevant to a given query or topic, and as the probability that a relevant document is retrieved. In a multi-label classification problem, each class corresponds to a distinct topic. Accordingly, precision and recall for the k-th class are defined respectively as $p_k = \mathrm{P}(\mathbf{x} \in \omega_k \mid f_k(\mathbf{x}) = 1)$, and $r_k = \mathrm{P}(f_k(\mathbf{x}) = 1 \mid \mathbf{x} \in \omega_k)$. Ideally, both measures should equal 1. However, in practice a higher precision can be attained only at the expense of a lower recall, and vice versa. In practice, they can be estimated from a multi-label data set as:

$$\hat{p}_k = \frac{TP_k}{TP_k + FP_k}, \hat{r}_k = \frac{TP_k}{TP_k + FN_k}, \tag{1}$$

where TP_k (true positive) is the number of samples that are correctly labelled as belonging to ω_k, while FP_k (false positive) and FN_k (false negative) are defined analogously.

To obtain a scalar performance measure, the Van Rijsbergen's F measure is often used. For a class ω_k it is defined as:

$$\hat{F}_{\beta,k} = \frac{1+\beta^2}{\beta^2/\hat{p}_k + 1/\hat{r}_k}, \tag{2}$$

where $\beta \in [0, +\infty]$ allows to weigh the relative importance of precision and recall. In particular, for $\beta = 1$ the F measure equals their harmonic mean.

The overall precision and recall on all categories can be computed either by macro- or micro-averaging the class-related values, depending on application requirements [8]. We will denote macro- and micro-averaged values respectively with the superscripts 'M' and 'm'. Macro-averaging simply consists in averaging the category-related values. The corresponding F measure is:

$$\hat{F}_{\beta}^{\mathrm{M}} = \frac{1}{N}\sum_{k=1}^{N}\hat{F}_{\beta,k} = \frac{1}{N}\sum_{k=1}^{N}(1+\beta^2)/\left((1+\beta^2) + \frac{FP_k + \beta^2 FN_k}{TP_k}\right). \tag{3}$$

The micro-averaged precision and recall are instead defined as:

$$\hat{p}^{\mathrm{m}} = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N}(TP_k + FP_k)},\ \hat{r}^{\mathrm{m}} = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N}(TP_k + FN_k)}, \tag{4}$$

and the corresponding F measure is defined as $\hat{F}_{\beta}^{\mathrm{m}} = \frac{1+\beta^2}{\beta^2/\hat{p}^{\mathrm{m}} + 1/\hat{r}^{\mathrm{m}}}$ [13], which after some algebra leads to:

$$\hat{F}_{\beta}^{\mathrm{m}} = (1+\beta^2)/\left((1+\beta^2) + \frac{\sum_{k=1}^{N}(FP_k + \beta^2 FN_k)}{\sum_{k=1}^{N} TP_k}\right). \tag{5}$$

It is known that macro-averaged measures are dominated by the performance on rare classes, while the opposite happens in the case of micro-averaging [8, 13]. Furthermore, an improvement on one of them can be usually attained only at the expense of the other one, especially when there are rare categories [2]. In the rest of this paper we will consider only the F measure, as it is widely used in multi-label tasks, is easier to handle being a scalar measure, and can be used to find a trade-off between precision and recall [13].

4 Criteria based on the *F* measure for static multi-label classifier selection

In this section we discuss selection criteria for the static implementation of the HSM method, when classification accuracy is evaluated in terms of the F measure. We consider first the case of $L' = 1$. In this case, for each class ω_k we want to select a single, possibly different classifier $f^{i(k)}$, $i(k) \in \{1, \dots L\}$. What we obtain can be seen as a new, single multi-label classifier whose outputs for the N classes are given by $f_1^{i(1)}(\mathbf{x}), f_2^{i(2)}(\mathbf{x}), \dots, f_N^{i(N)}(\mathbf{x})$. The final goal is to maximise the overall F measure (either macro- or micro-averaged).

From Eq. (2) it can be seen that maximising the macro-averaged F measure $\hat{F}_\beta^{\mathrm{M}}$ amounts to independently maximise the $\hat{F}_{\beta,k}$ measure of each class. This implies that the classifier for each class can be chosen independently on the other classes. It is also easy to see that to maximise $\hat{F}_{\beta,k}$ one should choose the classifier $f^{i(k)}$ such that:

$$i(k) = \arg \min_{i \in \{1,\ldots,L\}} \frac{FP_k^i + \beta^2 FN_k^i}{TP_k^i} \ , \tag{6}$$

where FP_k^i denotes the number of samples erroneously labelled as belonging to ω_k by f^i, and similarly for FN_k^i and TP_k^i. Obviously, these terms have to be estimated from validation samples. We name the above selection criterion $\mathrm{HSM}^{\mathrm{M}}$ (the superscript 'M' stands for "macro-averaging").

In the case of the micro-averaged $\hat{F}_\beta^{\mathrm{m}}$ measure instead, Eq. (5) shows that it can not be maximised by independently considering the contribution of the different classes. This can be seen by noting that maximising $\hat{F}_\beta^{\mathrm{m}}$ amounts to minimise the ratio between the two summands in Eq. (5). Let us rewrite this term by separating the contribution of any class ω_k:

$$\frac{(FP_k + \beta^2 FN_k) + \sum_{j \neq k}(FP_j + \beta^2 FN_j)}{TP_k + \sum_{j \neq k} TP_j} \ . \tag{7}$$

It is clear that the contribution of the terms related to ω_k, $(FP_k + \beta^2 FN_k)$ and TP_k, to the overall value of expression (7) is not independent on the contribution of the remaining terms, related to all the other classes. It follows that, to select the classifiers $f_1^{i(1)}, f_2^{i(2)}, \ldots, f_N^{i(N)}$ which maximise $\hat{F}_\beta^{\mathrm{m}}$, an exhaustive search over all the possible L^N choices is required, which is clearly impractical.

Nevertheless, the analysis of expression (7) reveals that, under some conditions on the values of the four terms $(FP_k + \beta^2 FN_k)$, $\sum_{j \neq k}(FP_j + \beta^2 FN_j)$, TP_k and $\sum_{j \neq k} TP_j$, the contribution of the terms related to ω_k *is* independent on that of the remaining terms. Under such conditions, it turns out that the classifier $f_k^{i(k)}$ that maximises $\hat{F}_\beta^{\mathrm{m}}$ can be chosen independently on the other ones, using the following criterion:

$$i(k) = \arg \min_{i \in \{1,\ldots,L\}} \frac{(FP_k^i + \beta^2 FN_k^i) + A_k}{TP_k^i + B_k} \ , \tag{8}$$

where A_k and B_k are two arbitrary, positive constants. It also turns out that, under slightly stricter conditions, both A_k and B_k can be zero. Due to the lack of space, the proof is not reported in this paper.[1]

According to the above result, when the micro-averaged F measure is used we propose to use the criterion (8) to select a classifier for each class. We will call this criterion $\mathrm{HSM}^{\mathrm{m}}$ (the superscript stands for "micro-averaging"). Clearly, this is a suboptimal choice, as the conditions under which (8) is the optimal criterion may not hold for all classes simultaneously, and anyway in practice one can not

[1] The proof can be found at `http://prag.diee.unica.it/pra/bib/pillai_mcs2011`

know whether they hold or not, for any class. The choice of the constants A_k and B_k can be made in such a way to limit the consequences of the non-optimality of HSM^{m}. To this aim, we propose to set A_k and B_k to a value that approximates the corresponding terms $\sum_{j\neq k}(FP_j + \beta^2 FN_j)$ and $\sum_{j\neq k} TP_j$ in Eq. (7), which can be estimated from validation data together with the terms FP_k^i, FN_k^i and TP_k^i of (8).

Consider now the case of $L' > 1$, namely when two or more classifiers have to be selected for each class. To maximise $\hat{F}_\beta^{\text{M}}$, the best subset of L' classifiers can be chosen independently for each class, for the same reason explained above. The corresponding criterion is the same as HSM^{M} of(6), where the FP, FN and TP values now refer to the combination of L' classifiers instead of a single classifier. However, this requires to evaluate all possible $\binom{L}{L'}$ combinations of classifiers, which may be impractical.

Similar considerations apply to the case of the $\hat{F}_\beta^{\text{M}}$ measure. Even in the conditions under which the criterion HSM^{m} of (8) is optimal for all classes (where the FP, FN and TP values now refer to an ensemble of L' classifiers as above), all possible $\binom{L}{L'}$ combinations of classifiers must be evaluated for each class. In principle, in the worst case when such conditions do not hold for any class, the number of classifier ensembles to evaluate becomes $\binom{L}{L'}^N$.

To keep computational complexity low when $L' > 1$, in this paper we will consider the simplest sub-optimal criterion for both $\hat{F}_\beta^{\text{M}}$ and $\hat{F}_\beta^{\text{m}}$. It consists in selecting the top L' classifiers for each class, ranked in terms of the corresponding HSM^{M} or HSM^{m} criterion of (6) and (8).

We finally discuss an interesting by-product of the above results. We mentioned above that under the conditions when HSM^{m} is optimal, the positive constants A_k and B_k of (8) can be arbitrarily small. Accordingly, as A_k and B_k approach zero, HSM^{m} tends to the HSM^{M} criterion of (6). This leads us to argue that HSM^{M} may also provide a good micro-averaged F measure. On the other hand, HSM^{m} requires to maximise for each class a quantity that does not depend on the other classes, analogously to HSM^{M}. It may thus provide in turn also a good macro-averaged F measure. In other words, we argue that both HSM^{m} and HSM^{M} may provide a good trade-off between the macro- and micro-averaged F measure, with respect to the performance of the individual multi-label classifiers.

5 Experimental evaluation

The experiments presented in this section are aimed at investigating whether, given a set of multi-label classifiers, the HSM method can outperform a "standard" selection method which selects the same classifier subset for each class, as well as the fusion of all the available classifiers.

The experiments have been carried out on three widely used benchmark data sets: the "ModApte" version of "Reuters 21578";[2] the Heart Disease subset of the

[2] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

Data set	Reuters	Ohsumed	Yeast
N. of training samples	7769	12775	1500
N. of testing samples	3019	3750	917
Feature set size	18157	17341	104
N. of classes	90	99	14
Distinct sets of classes	365	1392	164
N. of labels per sample (avg./max.)	1.234 / 15	1.492 / 11	4.228 / 11
N. of samples per class (min./max.)	1.3E-4 / 0.37	2.4E-4 / 0.25	0.07 / 0.75

Table 1. Characteristics of the three data sets used in the experiments.

Ohsumed data set [5]; and the Yeast data set.[3] Reuters and Ohsumed are text categorisation tasks, while Yeast is a gene functionality classification problem. Their main characteristics are reported in Table 1.

For Reuters and Ohsumed we used the *bag-of-words* feature model with the term frequency-inverse document frequency (tf-idf) kind of feature [8]. A feature selection step has also been carried out through a four-fold cross-validation on training samples, by applying stemming, stop-word removal and the information gain criterion. A feature set of 15,000 features was obtained for both data sets.

We implemented multi-label classifiers using the well known *binary relevance* (BR) approach. It consists in independently training N two-class classifiers (one per class) using the one-vs-all strategy: each classifiers independently decides whether labelling an input sample as belonging or not to the corresponding class [6, 8, 11]. We used as base two-class classifier a support vector machine (SVM) with a linear kernel for Reuters and Ohsumed (as it is considered the state of the art classifier for text categorisation tasks) and a SVM with a radial-basis function (RBF) kernel for Yeast. We used the `libsvm` software to implement SVMs [1]. The C parameter of the SVM learning algorithm was set to the `libsvm` default value of 1. The σ parameter of the RBF kernel, defined as $K(\mathbf{x}, \mathbf{y}) = \exp\left(-||\mathbf{x} - \mathbf{y}||^2/2\sigma\right)$, was set to 1 according to a four-fold cross-validation on training samples.

Since the output of a SVM is a real number, a threshold has to be set to decide whether labelling or not an input sample as belonging to the corresponding class. The threshold values can be set to optimise the considered performance measure. To maximise the macro-averaged F measure, it is known that the threshold can be set by independently maximising the F measure of each class [13]. No optimal criterion exists for maximising the micro-averaged F measure instead. To this aim we used a sub-optimal algorithm proposed in [2]. In both cases we estimated the thresholds through a five-fold cross-validation on training samples.

All the quantities involved in the selection criteria $\mathrm{HSM}^{\mathrm{M}}$ and $\mathrm{HSM}^{\mathrm{m}}$ were estimated through a five-fold cross-validation on training samples. Classification performance was evaluated using the F_1 measure, namely $\beta = 1$.

To generate an ensemble of multi-label classifiers for each data set, we used the random subspace method of [3]: each individual multi-label classifier was

[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html`

trained on the whole set of training samples, by using a different, randomly chosen feature subset. We used a fraction of 3/4 of the original feature set, and set the ensemble size to 10. We used as combining rules majority voting and simple averaging. For the latter, we also estimated the decision thresholds on the outputs of the fused classifiers, using again a five-fold cross-validation on training samples.

Five runs of the experiments were carried out. The average F_1 values and the standard deviation are shown in Table 2, related to the majority voting rule. The different columns contain the F_1 measure attained by all the considered methods. The two left-most and the two right-most columns correspond to the case when the criteria for classifier selection and for threshold optimisation were based respectively on the micro- and on the macro-averaged F_1. In both cases we report both the resulting macro- and micro-averaged F_1. For each data set we show the performance of the "standard" selection method when one, three and five two-class classifiers are selected for each class, the performance of the HSM method for the same number of selected classifiers, and the performance attained by the fusion of all the multi-label classifiers of the ensemble. The comparison between the "standard" selection method and HSM has to be done being equal the number of selected classifiers, and within the same column.

Table 2 shows that, being equal the number of selected classifiers, HSM almost always performed better or at least as good as the "standard" selection method, both in terms of the macro- and the micro-averaged F_1 measure. This shows that it can be capable to better exploit the complementarity between the multi-label classifiers on the different classes.

HSM also attained a better or at least a very similar performance as the fusion of all the available multi-label classifiers. This shows that, besides being more efficient at the classification phase, HSM can also attain a higher classification performance. On the contrary, the "standard" selection method was almost always outperformed by the fusion of all the available classifiers, with some exceptions on the Yeast data sets only.

Finally, HSM attained a higher macro-averaged F_1 measure than the "standard" selection method, even when the selection criterion based on the micro-averaged F_1 was used (see the second column of each table), and vice versa (third column). Moreover, we observed that, regardless on the selection criterion (either HSM^{M} or HSM^{m}), the resulting macro-averaged F_1 attained by HSM was very similar. The same result was observed in the case of the micro-averaged F_1 (these results are not reported, due to lack of space). The micro-averaged F_1 attained by the "standard" selection method was instead significantly higher than the macro-averaged F_1, if the selection criterion was based on the former measure, and vice-versa. This result provides evidence that, as argued in Sect. 4, both the HSM^{M} and the HSM^{m} criteria can be capable to attain a good trade-off between the macro- and micro-averaged F measure.

Similar results have been obtained using the simple average combining rule, as well as two different base classifiers, k-nearest neighbours and Naive Bayes (these results are not reported here due to lack of space).

	Selection method and ensemble size	Selection based on F_1^{m}		Selection based on F_1^{M}	
		F_1^{m}	F_1^{M}	F_1^{m}	F_1^{M}
Reuters	Best single	0.863 ± 0.005	0.520 ± 0.020	0.261 ± 0.037	0.564 ± 0.018
	HSM 1	0.878 ± 0.002	0.586 ± 0.011	0.427 ± 0.028	0.609 ± 0.010
	Best 3	0.874 ± 0.001	0.534 ± 0.013	0.267 ± 0.030	0.586 ± 0.019
	HSM 3	0.881 ± 0.002	0.588 ± 0.009	0.333 ± 0.030	0.613 ± 0.004
	Best 5	0.878 ± 0.002	0.557 ± 0.011	0.270 ± 0.038	0.587 ± 0.015
	HSM 5	0.882 ± 0.001	0.582 ± 0.005	0.308 ± 0.029	0.617 ± 0.005
	All	0.880 ± 0.002	0.560 ± 0.006	0.274 ± 0.037	0.603 ± 0.010
Ohsumed	Best single	0.667 ± 0.005	0.536 ± 0.021	0.653 ± 0.003	0.573 ± 0.012
	HSM 1	0.683 ± 0.001	0.591 ± 0.010	0.672 ± 0.002	0.612 ± 0.010
	Best 3	0.681 ± 0.003	0.560 ± 0.016	0.672 ± 0.005	0.607 ± 0.003
	HSM 3	0.688 ± 0.001	0.595 ± 0.007	0.682 ± 0.002	0.624 ± 0.010
	Best 5	0.685 ± 0.003	0.576 ± 0.006	0.677 ± 0.001	0.609 ± 0.005
	HSM 5	0.690 ± 0.001	0.596 ± 0.006	0.684 ± 0.002	0.623 ± 0.008
	All	0.684 ± 0.004	0.573 ± 0.005	0.681 ± 0.003	0.617 ± 0.005
Yeast	Best single	0.675 ± 0.004	0.439 ± 0.009	0.618 ± 0.008	0.494 ± 0.002
	HSM 1	0.676 ± 0.002	0.449 ± 0.004	0.642 ± 0.005	0.497 ± 0.004
	Best 3	0.679 ± 0.003	0.443 ± 0.006	0.623 ± 0.005	0.496 ± 0.004
	HSM 3	0.680 ± 0.002	0.449 ± 0.005	0.641 ± 0.003	0.501 ± 0.004
	Best 5	0.680 ± 0.002	0.444 ± 0.003	0.624 ± 0.006	0.498 ± 0.002
	HSM 5	0.680 ± 0.002	0.445 ± 0.003	0.636 ± 0.003	0.499 ± 0.001
	All	0.680 ± 0.002	0.438 ± 0.003	0.631 ± 0.006	0.498 ± 0.002

Table 2. Average F_1 measure and standard deviation attained on the three data sets by the "standard" selection method and by HSM (see text for the details).

6 Conclusions

In this work we proposed a classifier selection approach specific to ensembles of multi-label classifiers, which is based on selecting a possibly different subset of classifiers for each class. This allows in principle to better exploit the complementarity between the multi-label classifiers, on the different classes. Moreover, we developed two static classifier selection criteria based on the macro- and the micro-averaged F measure, which is widely used in multi-label tasks.

Our experimental results provided evidence that the proposed selection approach can be more effective than a "standard" approach based on selecting the same classifier subset for each class, as well as than fusing all the available multi-label classifiers. An interesting by-product is that both the proposed selection can also attain a good trade-off between the macro- and micro-averaged F measure, despite it is known that an increase in either of them is usually attained at the expense of the other one.

In light of these results, it becomes interesting to further investigate the following issues: investigating the characteristics of an ensemble of multi-label classifiers that make the proposed selection approach more effective; evaluating

dynamic selection methods based on this approach; analysing the behaviour of these static and dynamic selection methods as a function of the training set size, as well as the effect of class imbalance, which is a typical problem in multi-label tasks involving a high number of classes.

Acknowledgements This work was partly supported by a grant from Regione Autonoma della Sardegna awarded to Ignazio Pillai, PO Sardegna FSE 2007-2013, L.R.7/2007 "Promotion of the scientific research and technological innovation in Sardinia".

References

1. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
2. Fan, R.E., Lin, C.J.: A study on threshold selection for multi-label. Tech. rep., National Taiwan University (2007)
3. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20(8), 832–844 (1998)
4. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
5. Lewis, D.D., Schapire, R.E., Callan, J.P., Papka, R.: Training algorithms for linear text classifiers. In: SIGIR. pp. 298–306 (1996)
6. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: ECML/PKDD. vol. 5782, pp. 254–269. Springer (2009)
7. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39(2/3), 135–168 (2000)
8. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (Mar 2002)
9. Tahir, M., Kittler, J., Mikolajczyk, K., Yan, F.: Improving Multilabel Classification Performance by Using Ensemble of Multi-label Classifiers. In: Proc. of Multiple Classifier Systems (2010)
10. Tsoumakas, G., Katakis, I.: Multi label classification: An overview. Int. Journal of Data Warehousing and Mining 3(3), 1–13 (2007)
11. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data Mining and Knowledge Discovery Handbook pp. 667–685 (2010)
12. Woods, K., Kegelmeyer, W.P., Bowyer, K.W.: Combination of multiple classifiers using local accuracy estimates. IEEE Trans. Pattern Anal. Mach. Intell. (1997)
13. Yang, Y.: A study of thresholding strategies for text categorization. In: Int. Conf. on Research and development in information retrieval. New York, USA (2001)
14. Z.-H- Zhou, Z., Jianxin, W., Wei, T.: Ensembling neural networks: many could be better than all. Artificial Intelligence 137(1/2), 239–263 (2002)