A Classification Approach with a Reject Option 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. We investigate the implementation of multi-label classification algorithms with a reject option, as a mean to reduce the time required to human annotators and to attain a higher classification accuracy on automatically classified samples than the one which can be obtained without a reject option. Based on a recently proposed model of manual annotation time, we identify two approaches to implement a reject option, related to the two main manual annotation methods: browsing and tagging. In this paper we focus on the approach suitable to tagging, which consists in withholding either all or none of the category assignments of a given sample. We develop classification reliability measures to decide whether rejecting or not a sample, aimed at maximising classification accuracy on non-rejected ones. We finally evaluate the trade-off between classification accuracy and rejection rate that can be attained by our method, on three benchmark data sets related to text categorisation and image annotation tasks.

Keywords: Multi-label classification, Reject option

1 Introduction

In a multi-label classification problem 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 [13], notably text categorisation [12] and scene categorization [2], and are receiving an increasing interest in the pattern recognition and machine learning literature. Nevertheless, in many tasks automatic classification techniques do not achieve a satisfactory performance yet [14]. As an example in a text categorisation task, the best results obtained through the automatic "Medical Text Indexer" tool at the U.S. National Library of Medicine database (MEDLINE), is a recall of about 0.53 and a precision of about 0.30 [1]. In the recent ImageCLEF 2010 image annotation contest, the best automatic system attained a mean average precision of 0.45 [9]. Therefore, manual categorisation remains the only reliable solutions for many practical applications, although it is a tedious and labour-intensive procedure. This is also confirmed by the proliferation of manual image annotation tools [14], and by the use of "Medical Text Indexer" only as a *recommendation* tool by MEDLINE's human indexers [11].

Based on the above premises, in this paper we investigate a hybrid manual–automatic annotation approach inspired by the *reject option* used in single-label classifiers. The reject option consists in withholding the automatic classification of a sample, if the decision is not considered reliable enough. It is a mean to limit excessive misclassifications, at the expense either of a manual post-processing of rejections, or of their automatic handling by a more accurate but also computationally more costly classifier, and requires therefore a trade-off between the accuracy attainable on non-rejected samples and the amount (cost) of rejections [4, 10]. Analogously, in multi-label problems a classifier with a reject option could automatically take decisions on category assignments deemed reliable for a given sample, and could withheld and leave to a manual annotator only the ones deemed unreliable. This could allow a classifier to attain a high classification performance on non-withheld decisions, which should be traded for the cost of manual annotation of withheld decisions.

However, the theory and implementations of the reject option proposed so far in the pattern recognition literature have been developed only for single-label classifiers and only under the framework of the minimum risk theory. They can not be applied to multi-label classifiers, whose performance measures are based on precision and recall, and do not take into account the cost of correct/incorrect decisions. Therefore, in this paper we will first discuss how a reject option can be implemented in multi-label classifiers, based on the analysis of the cost (time) of manual labelling given in [14]. In Sect. 2 we show that this analysis suggests two possible implementations: rejecting all the category assignments of a sample, or only a subset of them. The latter option has already been proposed in previous works by the authors, although it was not not rigorously motivated [6, 7]. Therefore, in this paper we focus on the former option. In Sect. 3 we discuss how classification accuracy on non-rejected samples can be measured in terms of precisions and recall, and in Sect. 4 we derive two methods to maximise such accuracy for a given fraction of rejected samples, namely a given cost of manual annotation. The trade-off between classification accuracy and the fraction of rejected samples is experimentally evaluated in Sect. 5 on three benchmark data sets related to a text categorisation and to an image annotation task.

2 Rejection criteria for multi-label problems

Single-label classification problems with a reject option were formalized under the framework of the minimum risk theory in [4]. Denoting the costs of correct classifications, rejections, and misclassifications respectively as λ_C, λ_R and λ_E (with $\lambda_C < \lambda_R < \lambda_E$), the expected classification cost is minimized by assigning a sample to the class with the maximum a posteriori probability, if such probability is higher than $(\lambda_E - \lambda_R)/(\lambda_E - \lambda_C)$, and otherwise in rejecting it. This framework does not fit multi-label problems, whose performance measures are given in terms of precision and recall, which are not related to classification costs (see Sect. 3). To devise an implementation of a reject option in multi-label

problems, one issue to address is how to evaluate the cost of manually handling withheld category assignments.

The cost of manual annotation clearly depends on the annotation time. A model of the annotation time has been proposed in [14], for two possible annotation procedures: *tagging*, which consists in labelling a sample according to a given set of categories (keywords or "tags"), and *browsing*, in which the relevance has to be decided for a whole set of samples to one category at a time. According to [14], the annotation time of tagging and browsing is:

$$t_{\text{tagging}} = M \cdot (\overline{K} \cdot t_f + t_s), \qquad t_{\text{browsing}} = \sum_{k=1\ldots N} (M_p^k \cdot t_p + M_n^k \cdot t_n) \ ,$$

where N is the number of categories, M is the number of samples, t_s is the so called "initial setup" time to analyse a sample, t_f is the time to assign one label, $\overline{K}$ is the average number of labels per sample, M_p^k and M_n^k are respectively the number of samples which belong and do not belong to the k-th category, while t_p and t_n are the time for deciding whether or not a sample belongs to a category. The most efficient procedure among tagging and browsing can be made on the basis of the values of the above parameters, according to the task at hand [14].

The above model of manual annotation time suggests two main approaches to implement a reject option in multi-label classifiers, aimed at trading the classification accuracy on automatically assigned labels for the cost of manually processing category assignments withheld by a classifier:

1. In tasks where tagging is used, the manual annotation time of withheld category assignments can be directly controlled by setting a constraint to the number of samples which contain withheld assignments, which corresponds to the term M. Accordingly, in this case it make sense to reject either *all* or none of the assignments of a sample, rather than only a subset of them.
2. In tasks where browsing is used, the manual annotation time of withheld category assignments can be controlled by setting a constraint on the number of samples which contain withheld assignments, *independently* for each category, which correspond to M_p^k and M_n^k. In this case it makes sense to withheld for each sample only a subset of its category assignments (not necessarily all of them), obviously the most unreliable ones.

In the former approach, the objective is clearly to maximise classification accuracy on non-rejected samples, with a constraint on the maximum fraction of rejected samples. In the latter approach, the objective is instead to maximise classification accuracy on non-withheld decisions, with a constraint on the maximum fraction of withheld decisions for each individual category. In both cases, the effectiveness of a reject option has to be evaluated in terms of the attainable trade-off between the accuracy of the classifier on non-withheld category assignments, and the cost (annotation time) of withheld ones, taking into account the application requirements of the task at hand.

An implementation of the latter approach has already been investigated by the authors in [6, 7], although it was not motivated by the above arguments. Therefore, in this paper we focus on the former implementation.

3 Accuracy of multi-label classifiers with a reject option

In this section we discuss how to evaluate the accuracy of a multi-label classifier in presence of withheld category assignments. To this aim we first introduce accuracy measures based on precision and recall.

In the field of information retrieval, *precision* is the probability that a retrieved document is relevant to a given query or topic, while *recall* is the probability that a relevant document is retrieved. In a multi-label classification problem, each class corresponds to a distinct topic. Denoting the set of categories as $\Omega = \{\omega_1, \ldots, \omega_N\}$, 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, 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 $f(\mathbf{x})$ means that the sample $\mathbf{x}$ is labelled as (not) belonging to ω_k. Accordingly, precision for the k-th class, denoted as p_k, is the probability that a sample belongs to ω_k, given that it is labelled as such: $p_k = \mathrm{P}(\mathbf{x} \in \omega_k \mid f_k(\mathbf{x}) = 1)$. Recall (r_k) is the probability that a sample is correctly labelled as belonging to ω_k: $r_k = \mathrm{P}(f_k(\mathbf{x}) = 1 \mid \mathbf{x} \in \omega_k)$. Ideally, both precision and recall should equal 1. However, in practice a higher precision can be attained only at the expense of a lower recall, and vice versa. As limit cases, labelling all samples as belonging to ω_k leads to $p_k = 0$ and $r_k = 1$, why labelling all samples as not belonging to ω_k leads to $p_k = 1$ and $r_k = 0$.

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

$$F_{\beta,k} = \frac{1 + \beta^2}{\beta^2/p_k + 1/r_k}, \tag{1}$$

where the parameter $\beta \in [0, +\infty]$ weigh the relative importance of precision and recall: $\beta < 1$ gives a higher weight to recall, while the opposite happens for $\beta > 1$.

Precision and recall can be estimated from a multi-label data set as:

$$\hat{p}_k = \tfrac{TP_k}{TP_k + FP_k}, \quad \hat{r}_k = \tfrac{TP_k}{TP_k + FN_k}, \tag{2}$$

where TP_k (true positive) and FP_k (false positive) are respectively the number of samples correctly and erroneously labelled as belonging to ω_k, while FN_k (false negative) is the number of samples erroneously labelled as not belonging to ω_k. The F measure can be estimated by replacing the estimates of precision and recall of eq. (2) into eq. (1).

For a multi-label classifier, the global precision and recall over all categories can be computed either by macro- or micro-averaging the class-related values, depending on application requirements [12]. We will denote macro- and micro-averaged values respectively with the superscripts 'M' and 'm'. Macro- and

micro-averaged precision and recall are defined as:

$$\hat{p}^{\mathrm{M}} = \frac{1}{N} \sum_{k=1\ldots N} \hat{p}_k, \qquad \hat{r}^{\mathrm{M}} = \frac{1}{N} \sum_{k=1\ldots N} \hat{r}_k, \tag{3}$$

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

The corresponding F measure is defined as [15]:

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

$$\hat{F}^{\mathrm{m}}_{\beta} = \frac{1+\beta^2}{\beta^2/\hat{p}^{\mathrm{m}} + 1/\hat{r}^{\mathrm{m}}} = (1+\beta^2) / \left((1+\beta^2) + \frac{\sum_{k=1}^{N} (FP_k + \beta^2 FN_k)}{\sum_{k=1}^{N} TP_k} \right). \tag{6}$$

Let us now consider how to extend the above performance measures to a multi-label classifier with a reject option. A withheld decision for a sample $\mathbf{x}$ and a category ω_k can be denoted with the value 0 as the output of $f_k(\mathbf{k})$. In single-label problems the accuracy attained by a classifier with a reject option is evaluated as the conditional probability that a pattern is correctly classified, given that it has not been rejected. Analogously, precision and recall for a given category, when a reject option is used, can be defined only with respect to non-withheld decisions. It is easy to see that their corresponding probabilistic definition remains the standard one given at the beginning of this section, which only considers the case $f_k(\mathbf{k}) = 1$, thus excluding withheld assignments (namely, the case when $f_k(\mathbf{k}) = 0$). Consequently, also the F measure can still be defined as in Eq. (1). The estimate of these measures on a given data set can be obtained using again Eq. (2), but taking into account only non-withheld category assignments in the computation of TP_k, FN_k and FP_k. The micro- and macro-averaged values can be computed in the same way using Eqs. (3)–(6).

In the rest of this paper we will consider only the F measure (both macro- and micro-averaged), 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 [15].

4 Maximising the F measure for a given cost of rejections

In this section we address the issue of how to define a decision function $f(\mathbf{x}) = \{f_1(\mathbf{x}), \ldots, f_N(\mathbf{x})\} \in \{-1, 0, +1\}^N$ for a N-category multi-label classifier with a reject option, with the constraint that either each or none of the $f_k(\mathbf{x})$ equals 0, according to the approach discussed in Sect. 2. As explained in Sect. 2, the goal is to maximise the classification accuracy on non-rejected samples, with the constraint that up to a given fraction of samples can be rejected. We will denote such fraction as $r_{\max}$.

To decide whether a given sample $\mathbf{x}$ has to be rejected or not, by analogy with approaches widely used in single-label problems we would like to define a

measure of "classification reliability" $R(\mathbf{x})$ and a rejection threshold T, such that a sample $\mathbf{x}$ is rejected if $R(\mathbf{x}) < T$, and is automatically classified otherwise. The value of T has to be set according to the desired rejection rate $r_{\max}$, usually from validation data. To define a classification reliability measure, one could estimate the effect of rejecting a sample on the F measure: intuitively, the higher is the F measure obtained after rejecting a given sample $\mathbf{x}$ belonging to any set of samples S, the less reliable is its automatic classification. Formally, the sample $\mathbf{x}^* \in S$ which is classified with the lowest reliability is given by:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in S} \hat{F}_\beta(S - \{\mathbf{x}\}) \ , \tag{7}$$

where $\hat{F}_\beta(A)$ denotes the value of the F measure (either macro- or micro-averaged) evaluated on the set of samples A. Accordingly, $R(\mathbf{x})$ could be defined as a monotonic decreasing function of (an estimate of) $\hat{F}_\beta(S - \{\mathbf{x}\})$: the higher $\hat{F}_\beta(S - \{\mathbf{x}\})$, the less reliable the classification of $\mathbf{x}$.

Consider first the micro-averaged F measure of Eq. (6). Maximising $\hat{F}^{\mathrm{m}}_\beta(S - \{\mathbf{x}\})$ amounts to maximise the term

$$\frac{TP(S) - TP(\mathbf{x})}{(FP(S) + \beta^2 FN(S)) - (FP(\mathbf{x}) + \beta^2 FN(\mathbf{x}))} \ , \tag{8}$$

where $FP(Z)$ denotes the number of false positive errors made by the classifier on the set of samples Z, while the meaning of $TP(Z)$ and $FN(Z)$ is similar. Unfortunately, while in single-label problems the contribution of the classification outcome (either correct or wrong) of a given sample to the expected risk does not depend on the outcome of the other samples, it turns out that this does not hold for multi-label problems when classification performance is evaluated using the F measure. Indeed, it is easy to see that the value of Eq. (8) depends not only on the rejected sample $\mathbf{x}$, but also on all the other samples.

Nevertheless, the analysis of Eq. (8) reveals that, under some conditions on the values of its terms, the contribution of a sample $\mathbf{x}$ *does not* depend on the other samples. In particular, under such conditions it can be shown that the individual sample $\mathbf{x}^*$ whose rejection maximises $\hat{F}^{\mathrm{m}}$ can be found as follows:

$$\mathbf{x}^* = \min_{\mathbf{x}\in S} \frac{TP(\mathbf{x}) + A}{FP(\mathbf{x}) + \beta^2 FN(\mathbf{x}) + B} \ , \tag{9}$$

where A and B are two arbitrary positive constants.[1] Whether or not the conditions mentioned above hold is however unknown in practice. Therefore, (9) can be used to define only a suboptimal classification reliability measure to be used for any $\mathbf{x}$. In this paper we define $R(\mathbf{x})$ exactly as the right-hand side of (9):

$$R(\mathbf{x}) = \frac{TP(\mathbf{x}) + A}{FP(\mathbf{x}) + \beta^2 FN(\mathbf{x}) + B} \ . \tag{10}$$

[1] Due to lack of space, the proof of these properties is reported here, and can be found at `http://prag.diee.unica.it/pra/bib/pillai_iciap2011_rj`.

As explained above, any pair of values $A > 0$ and $B > 0$ can be used, if the conditions mentioned above hold. To take into account the cases when they do not hold, we can set A and B such that $R(\mathbf{x})$ approximates the value of expression (8). Namely, we can set $A = \hat{TP}(S)$, $B = \hat{FP}(S) + \beta^2\hat{FN}(S)$, where $\hat{TP}(S)$, $\hat{FP}(S)$ and $\hat{FN}(S)$ are estimated from validation data. The values $TP(\mathbf{x})$, $FP(\mathbf{x})$ and $FN(\mathbf{x})$ can be estimated from validation data as well. For instance, $TP(\mathbf{x})$ (true positives) can be estimated as the number of correct category assignments on the subset of the K validation samples nearest to $\mathbf{x}$, for some K. An alternative method can be used for classifiers which provide a score $s_k(\mathbf{x}) \in \mathbb{R}$ for each class ω_k, as many multi-label classifiers do: one can consider for each class ω_k the K validation samples whose scores are closest to $s_k(\mathbf{x})$.

Consider now the macro-averaged F measure of Eq. (5). It is not difficult to see that using the criterion (7) to define a classification reliability measure, the contribution of a sample $\mathbf{x}$ is not independent on the other samples, similarly to the case of the micro-averaged F measure. However, note that F_β^{M} is defined as the mean of $F_{\beta,k}$ of the N classes. It turns out that under some conditions on $TP_k(\mathbf{x})$, $FP_k(\mathbf{x})$ and $FN_k(\mathbf{x})$, the analogous of Eq. (9) holds for the individual $F_{\beta,k}$, for any $A_k, B_k > 0$. As a suboptimal classification reliability measure we chose therefore to use:

$$R(\mathbf{x}) = \frac{1}{N} \sum_{k=1\ldots N} \frac{TP_k(\mathbf{x}) + A_k}{FP_k(\mathbf{x}) + \beta^2 FN_k(\mathbf{x}) + B_k} \ , \tag{11}$$

where the values A_k, B_k, $FP_k(\mathbf{x})$, $FN_k(\mathbf{x})$ and $TP_k(\mathbf{x})$ can be estimated form validation data as explained above, independently for each category ω_k.

5 Experimental Evaluation

The aim of our experiments was to evaluate the trade-off between automatic classification accuracy and the fraction of rejected samples that can be attained by a multi-label classifier using the approach proposed in this paper.

The experiments were carried out on three widely used benchmark data sets, related to two text categorisation and one image annotation task: the "ModApte" version of "Reuters 21578",[2] the Heart Disease subset of the Ohsumed data set [8], and the Scene data set[3]. Their main characteristics are reported in Table 1.

The *bag-of-words* feature model with the term frequency–inverse document frequency (tf-idf) features [12] was used for Reuters and Ohsumed. A feature selection pre-processing step was carried out for Reuters and Ohsumed, through a four-fold cross-validation on training samples, by applying stemming, stop-word removal and the information gain criterion. This lead to the selection of $15,000$ features for both data sets.

To implement a N-class multi-label classifier we used the well known *binary relevance* approach. It consists in independently constructing N two-class

[2] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`
[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html`

Data set	Reuters	Ohsumed	Scene
N. of training samples	7769	12775	1211
N. of testing samples	3019	3750	1196
Feature set size	18157	17341	295
N. of classes	90	99	6
Distinct sets of classes	365	1392	14
N. of labels per sample (avg./max.)	1.23 / 15	1.492 / 11	1.06 / 3
N. of samples per class (min./max.)	1.3E-4 / 0.37	2.4E-4 / 0.25	0.136 / 0.229

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

classifiers using the one-vs-all strategy [12, 13]. We used as the base two-class classifier a support vector machine (SVM) implemented by the `libsvm` software [3]. A SVM linear kernel was used for Reuters and Ohsumed, as it is considered the state of the art classifier for text categorisation tasks. A radial-basis function (RBF) kernel was used for Scene instead. 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 estimated by 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 N threshold values can be chosen as the ones which optimise the considered performance measure. In these experiments we used the F measure with $\beta = 1$. To maximise the macro-averaged F measure of Eq. (5), it is known that the threshold can be set by independently maximising the individual F measure of each class, Eq. (1) [15]. No optimal algorithm exists for maximising the micro-averaged F measure instead. We used the suboptimal iterative maximisation algorithm recently proposed in [5]. In both cases the thresholds were estimated through a five-fold cross-validation on training data.

In the experiments several values of the rejection rate $r_{\max}$ were considered, ranging in $[0, 0.3]$ with a step of 0.05. For each $r_{\max}$ value, we implemented a decision rule with the reject option by using the reliability measures $R(\mathbf{x})$ of Eq. (10) and Eq. (11), respectively when the micro- and macro-averaged F measure was used. For any input sample $\mathbf{x}$, the values of TP, FP and FN in $R(\mathbf{x})$ were estimated using the scores $s_k(\mathbf{x})$, $k = 1 \dots N$ of the SVMs, as described in Sect. 4. To this aim, we estimated the score distribution for each class on training samples, using 20 bins histograms, where the bins correspond to disjoint intervals of the score range. The rejection threshold T was set to the value that lead to the desired rejection rate $r_{\max}$ on training samples. Note that for $r_{\max} = 0$ we obtain a standard multi-label classifier without a reject option.

For each data set ten runs of the experiments were carried out, using 80% of the patterns of the original training set. To this aim, ten different training sets were obtained by randomly partitioning the original one into ten disjoint subsets of identical size, and using at each run only eight partitions as the training set. The original test set was used at each run.

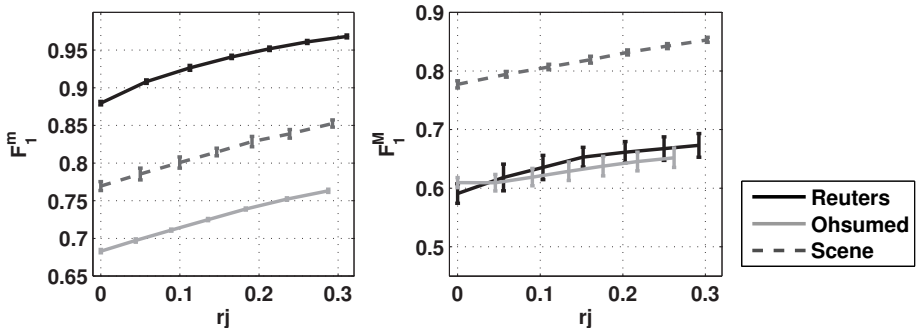

Fig. 1. Test set averaged F_1^m (left) and F_1^M (right) versus the rejection rate on the three data sets. The standard deviation is denoted by vertical bars.

In Fig. 1 we report the average micro- and macro-averaged F measure over the ten runs, as a function of $r_{\max}$. The standard deviation is also reported as vertical bars. Note that the decision thresholds of the N two-class SVM classifiers were computed by optimising the same performance measure (either the micro- or macro-averaged F measure) used to evaluate the classifier performance.

The results in Fig. 1 show that the classification accuracy attained on non-rejected samples always increases as the rejection rate increases. In particular, rejecting up to 30% of the samples, the accuracy improvements are quite remarkable for the micro-averaged F measure, and also for the macro-averaged one in the Scene data set, taking also into account the small standard deviation. Another relevant result is that the rejection rate observed in the test set was always very close to the desired rejection rate ($r_{\max} = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$), which was set on training samples through the choice of the threshold T. This can be seen in Fig. 1, where the rejection rates correspond to the position of the standard deviation bars.

6 Conclusions

We proposed two approaches to implement a reject option in multi-label classifiers, aimed at reducing the manual annotation time in tasks like text categorisation and image annotation (either by using the tagging or browsing approach), attaining at the same time a higher classification accuracy on automatically classified samples than the one which can be obtained without a reject option. We also derived a classification reliability measure to decide whether a sample has to be rejected or not, for the case when the tagging approach is used, with the aim of maximising both the macro- and micro-averaged F measure on non-rejected samples. Reported experimental results related to text categorisation and image annotation tasks provided evidence that the proposed approach can allow to significantly improve the accuracy of an automatic classifier, even when only 30% of samples are rejected and must be manually labelled.

We mention two issues to be further investigated. One is the definition of more accurate reliability measures, especially for the macro-averaged F measure. The other one stems from the novel manual annotation approach proposed in [14], which combines tagging and browsing, and is more efficient than both of them for some applications. Accordingly, a hybrid rejection approach obtained as the combination of the two ones identified in Sect. 2 can be devised for this hybrid tagging-browsing approach.

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. Aronson, A., Rogers, W., Lang, F., Névéol, A.: 2008 report to the board of scientific counselors (2008), `http://ii.nlm.nih.gov/IIPublications.shtml`
2. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (March 2004)
3. 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`
4. Chow, C.K.: On optimum recognition error and reject tradeoff. IEEE Transactions in Information Theory 16(1), 41–16 (1970)
5. Fan, R.E., Lin, C.J.: A study on threshold selection for multi-label. Tech. rep., National Taiwan University (2007)
6. Fumera, G., Pillai, I., Roli, F.: Classification with reject option in text categorisation systems. In: Int. Conf. Image Analysis and Proc. (2003)
7. Fumera, G., Pillai, I., Roli, F.: A Two-Stage Classifier with Reject Option for Text Categorisation. In: Structural, Syntactic, and Statistical Patt. Rec. (2004)
8. Lewis, D.D., Schapire, R.E., Callan, J.P., Papka, R.: Training algorithms for linear text classifiers. In: SIGIR. pp. 298–306 (1996)
9. Nowak, S., Huiskes, M.: New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. Working Notes of CLEF 2010 (2010)
10. Pudil, P., Novovicova, J., Blaha, S., Kittler, J.V.: Multistage pattern recognition with reject option. In: ICPR. pp. II:92–95 (1992)
11. Ruiz, M., Aronson, A.: User-centered evaluation of the medical text indexing (mti) system (2007), `http://ii.nlm.nih.gov/IIPublications.shtml`
12. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (Mar 2002)
13. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data Mining and Knowledge Discovery Handbook pp. 667–685 (2010)
14. Yan, R., Natsev, A., Campbell, M.: An efficient manual image annotation approach based on tagging and browsing. In: Workshop on multimedia inf. retr. on The many faces of multimedia semantics. pp. 13–20. (2007)
15. Yang, Y.: A study of thresholding strategies for text categorization. In: Int. Conf. on Research and development in information retrieval. New York, USA (2001)