A Comparison of Multi-Label Feature Selection Methods Using the Random Forest Paradigm

Ouadie Gharroudi, Haytham Elghazel, and Alex Aussem

Université de Lyon, Université Lyon 1, LIRIS UMR CNRS 5205, F-69622, France
firstname.lastname@liris.cnrs.fr

Abstract. In this paper, we discuss three wrapper multi-label feature selection methods based on the Random Forest paradigm. These variants differ in the way they consider label dependence within the feature selection process. To assess their performance, we conduct an extensive experimental comparison of these strategies against recently proposed approaches using seven benchmark multi-label data sets from different domains. Random Forest handles accurately the feature selection in the multi-label context. Surprisingly, taking into account the dependence between labels in the context of ensemble multi-label feature selection was not found very effective.

Keywords: Feature Selection, Multi-label Learning, Ensemble Methods, Random Forest.

1 Introduction

The problem of single-label classification is concerned with learning from a set of examples $\mathcal{X}$, where each example is associated with a single label λ from a finite set of disjoint labels $\mathcal{L}$ of size L, with $L > 1$. If $L = 2$, then the learning task is called binary classification, while if $L > 2$, then it is called multi-class classification. On the other hand, the task of learning a mapping from an instance $x \in X$ to a set of labels $Y \in \mathcal{L}$ is referred to as a multi-label classification. Multi-label classification is a challenging problem that emerges in several modern applications such as text categorization, gene function classification, and semantic annotation of images [3,24]. The issue of learning from multi-label data has recently attracted significant attention from many researchers and a considerable number of approaches have been proposed [5,17,25]. Basically, they can be summarized into two categories: (a) algorithm adaptation methods and (b) problem transformation methods. Algorithm adaptation methods extend specific learning algorithms to handle multi-label data directly. Problem transformation methods, on the other hand, transform the multi-label learning problem into either several binary classification problems, such as the Binary Relevance (BR) approach, or one multi-class classification problem, such as the Label Powerset (LP) approach. The single-label classification problems are then solved with a commonly used single-label classification approach and the output is transformed back into a multi-label representation.

M. Sokolova and P. van Beek (Eds.): Canadian AI 2014, LNAI 8436, pp. 95–106, 2014.

The identification of relevant subsets of random variables among thousands of potentially irrelevant and redundant variables is a very important topic of pattern recognition research that has attracted much attention over the last few years. In traditional single-label learning, feature selection algorithms use information from labeled data to find the relevant subsets of variables, i.e., those that conjunctively prove useful to construct an efficient classifier from data. It enables the classification model to achieve good or even better solutions with a restricted subset of features [10]. As in the single-label case, multi-label feature selection had been widely studied and have encountered some success in many applications [9,16,20].

Although considerable attention has been given on the problem of using Random Forest (RF) to estimate the feature importance for traditional supervised [4], unsupervised [11,12,8] and semi-supervised [1] learning, little attention has been given to exploiting the power of this ensemble method with a view to identify and remove the irrelevant features in a multi-label setting. The way internal estimates are used to measure variable importance in RF paradigm [4] have been influential in our thinking. In this study, we propose and experimentally evaluate three wrapper multi-label feature selection methods, which use the RF paradigm. The main idea is to run three variants of RF for Multi-label learning (BRRF, RFLP and RFPCT) and then exploit the RF permutation importance measure [4] to evaluate the goodness of a feature. BRRF, for *Binary Relevance Random Forest* and RFLP, for *Random Forest Label power-Set*, consists of the two problem transformation approaches BR and LP, to previously transform the multi-label data into single-label data, which is then used to perform a Random Forest. However, RFPCT [15] (Random Forest of Predictive Clustering Trees) is another extension of RF that uses as base classifier PCT [2], a decision tree predicting multiple target attributes at once. We would like to mention that feature selection using RFPCT was initially proposed in [13], nonetheless, it was evaluated on a single biological data set and only compared to a trivial random feature ranking algorithm in [14]. To the best of our knowledge, this study is the first attempt to compare several RF-based feature selection methods in the context of multi-label classification.

Empirical results on seven multi-labeled datasets will be presented to answer the following questions: (1) Is there any benefit of exploiting label dependence structure in the context of multi-label feature selection as suggested by several authors [9,25]? (2) How can we extend the RF approach to address the multi-label feature selection problem? (3) Are these RF-based methods competitive with other state-of-the-art feature selection methods?

The rest of the paper is organized as follow: Section 2 reviews recent studies on multi-label feature selection and ensemble methods. Section 3 introduces the three RF-based multi-label feature selection methods and describes how variable importance used in RF can be extended in multi-label context. Experiments using conventional benchmark data sets are presented in Section 4. We raise several issues for future work in Section 5 and conclude with a summary of our contributions.

2 Related Work

In this section, we briefly review the multi-label feature selection and multi-label ensemble learning approaches that appeared recently in the literature.

2.1 Multi-Label Feature Selection

In multi-label classification, most feature selection tasks have been addressed by extending the techniques available for single-label classification using the bridge provided by multi-label transformations. These methods propose a previous transformation of multi-label data to single-label data, *i.e.*, to binary data or multi-class data using respectively the BR or the LP approach. Thus, when the BR strategy is used, it is straightforward to employ a filter approach on each binary classification task, and then combining somehow the results (by an averaging for example). In this context, different feature importance measures have been used, such as Information Gain [20] and ReliefF [20]. Since each label is treated independently, these methods fail to consider the correlation among different labels. On the other hand, methods which perform feature selection using the same evaluation measures according to the LP approach take into account the label correlation [7,20]. Furthermore, the PMU approach in [16] is considered as the first filter approach that takes into account label interactions in evaluating the dependency of given features without resorting to problem transformation. The proposed method is presented as a multivariate mutual information-based feature selection method for multi-label learning that naturally derives from mutual information between a set of features and a set of labels.

In contrast to these previous filter approaches, Gu el al. propose an embedded-style feature selection method for multi-label learning called CMLFS [9]. CMLFS (for Correlated Multi-Label Feature Selection) is based on LaRank SVM, which is among state-of-the-art multi-label learning methods. In the proposed method, the goal is to find a subset of features, based on which the label correlation regularized loss of label ranking is minimized. Although this method considers correlation among labels, it optimizes a set of parameters during feature selection process to tune the kernel function of multi-label classifier making it impractical from the viewpoint of computational cost [16].

2.2 Multi-Label Ensemble Learning

The ensemble methods for multi-label learning are developed on top of the common problem transformation or algorithm adaptation methods. The most well known problem transformation ensembles are the RAkEL system by Tsoumakas et al. [21] and ensembles of classifier chains (ECC) [18]. RAkEL (for RAndom k-labELsets) is an ensemble of LP classifiers. It proposes breaking the initial set of labels into a number m of small-sized random subsets, called k-labelsets (k is the labelset size) and employing LP to train a corresponding classifier. A simple voting process determines the final classification set. In this manner, RAkEL take into account correlation between labels, and at the same time, avoid the weakness of LP methods, by reducing the the number of labels handled by the LP classifiers. On the other hand, ECC are ensemble methods that are based on classifier chain CC. The algorithm use p classifier chains $C_1, C_2, \ldots, C_p$; where in each C_i a random subset of instances is chosen as training data and the order of learners is performed using a random sequence of labels. For the multi-label classification of an unlabeled instance, the decisions of all CC classifiers are gathered and combined. A threshold is used to choose the final multi-label set. Both RAkEL and ECC are ensemble methods based on problem transformation algorithms. In algorithm adaptation category, RFPCT [15] is a standard Random Forest, which uses PCT [2] as base learner. PCT is an algorithm adaptation decision tree capable of predicting multiple target attributes at once.The induction process in PCT uses the sum of the Gini indices throughout all labels to identify the best separation at each node. In RFPCT, each tree makes multi-label predictions, and then predictions are combined using a majority or a probability voting scheme. The diversity among trees is promoted using two strategies; bootstrap sampling of training data and random selection of feature subsets.

3 Ensemble Feature Selection for Multi-Label Learning

RF has several desirable characteristics for feature selection: It is robust, exhibits high-quality predictive performance, does not overfit and handles simultaneously categorical and continuous features [4]. Furthermore, RF have proved to be efficient in traditional supervised [4], semi-supervised [1], and unsupervised [8] feature selection process. This section introduces and experimentally evaluates three wrapper multi-label feature selection methods, which use the RF paradigm. In this way, we discuss three variants of RF for Multi-label learning *Random forest of predictive clustering trees* (RFPCT), *Binary Relevance Random Forest* (BRRF), and *Random Forest Label Power-set* (RFLP); and then exploit the *RF permutation importance measure* [4] to evaluate the goodness of a feature. Before introducing the proposed methods, we recall how RF with permutation based out-of-bag (oob) measures feature importance.

The variable importance measure in RF is based on the decrease of predictive performance when values of a descriptive variable in a node of a tree are permuted randomly. Basically, a bootstrap is used as training set to create trees in the forest. In each bootstrapped data set, almost 33% are left oob, *i.e.*, they are not used for the construction of the i^{th} corresponding model h^i ($i \in \{1, \ldots, T\}$). We refer to them as Oob_i. Thus, these patterns can be used to estimate non biased feature relevancies. In every tree grown in the forest, the values of the f^{th} feature in the Oob_i data, is randomly permuted to form Oob_i^f, and the tree h^i is used to predict the labels of the new oob patterns. The predictive performance of each tree h^i is evaluated on the untouched oob data and the permuted versions of the oob data. The importance of the f^{th} variable is then calculated as the relative increase of the error that is obtained when its values are randomly permuted (*c.f.* Equation 1). The average of this number over all trees in the forest is the raw importance score for variable f. We note that the greater the value of the importance measure, the more relevant is the feature,

$$I(f) = \frac{1}{T} \sum_{i=1}^{T} \frac{e(h^i(Oob_i^f)) - e(h^i(Oob_i))}{e(h^i(Oob_i))} \tag{1}$$

where T is the size of the forest and e is the error measure function.

Given a label space $\mathcal{L} = (\lambda_1, \lambda_2, ..., \lambda_L)$ and a data set $\mathcal{D}$ that consists of N instances each taking the form (x_i, y_i) where $x_i = (x_{i1}, ..., x_{iM})$ is a vector of M descriptive attributes and $y_i \in \mathcal{L}$ is the subset of labels associated to x_i (represented by a binary feature vector $(y_{i1}, ..., y_{iL}) \in \{0, 1\}^L$), we present, in the sequel, the three used variants of RF for multi-label learning and describe how variable importance used in RF can be extended in this context.

Binary Relevance Random Forest (BRRF) - This method transforms the multi-label dataset $\mathcal{D}$ into many single-label datasets, one for each individual label in $\lambda_i \in \mathcal{L}$. After this transformation, a RF is created for each label λ_i. The relevance of each feature according to each individual label is measured using the above Equation 1 for which e is the traditional single-label classification error. Finally, the average of the score of all features across all labels is considered. BRRF, focuses on each label individually and does not take into account label dependence.

Random Forest Label Power-set (RFLP) - In this method the multi-label feature selection problem is handled using Power-set strategy. This approach reduces the multi-label dataset $\mathcal{D}$ to a multi-class dataset by treating each distinct label set as an unique multi-class label. To avoid creating too many calsses with few instances, that may issue an overfitting and an imbalance multi-class problems the Pruned Problem Trans formation as in [7] was used; patterns with too rarely occurring labels are simply removed from the training set by considering label sets with a predefined minimum occurrence τ. A RF could be now performed and the above described feature selection procedure will be naturally applied using in Equation 1 the traditional single-label classification error e. In this way, this approach directly takes into account label correlation.

Random Forest Predictive Clustering Tree (RFPCT) - In contrast to both previous approaches (BRRF and RFLP) for which the RF grows many classification trees using a CART as a base classifier, RFPCT [15] is an extension of RF that use a randomized variant of the non Pruned Predictive Clustering Tree (PCT) [2], as a base classifier. In this approach, the multi-label data $\mathcal{D}$ is handled directly and is then able to provide an intuitive way for taking into account relationships between labels. Nevertheless, it is noteworthy that BRRF and RFPCT perform comparably for classification (see [15] for more details).

The feature selection problem with RFPCT follows the same procedure described above. Feature relevance is measured on each PCT tree, and then averaged over all the trees in the forest. However, since PCT is an adaptation method devoted to learning simultaneously all the labels, the RF-based feature evaluation procedure requires an appropriate multi-label error measure e in Equation 1 instead of the ordinary classification error used for BRRF and RFLP. As suggested in [13,14], the multi-label error measure e used in each tree is obtained by averaging the individual classification errors across the L labels.

4 Performances Analysis

In this section, we investigate the effectiveness of the aforementioned RF-based feature importance measures for multi-label feature selection and compare their performances against two recent multi-label feature selection methods on seven benchmark data sets.

4.1 Data Sets and Evaluation Protocol

Seven benchmark multi-label data sets, mostly obtained from the *Mulan's repository* [22], were used to assess the performance of feature selection algorithms. We selected these data sets as they have already been used in various empirical studies and cover different application domains: Biology, semantic scene analysis, music emotions and text categorization. Table 1 shows, for each data set, the number of examples (N), the number of features (M), where b indicates that the feature values are binary and n indicates that the feature values are numeric; the number of labels (L), the Label Cardinality (LC), which is the average number of single-labels associated with each example; the Label Density (LD), which is the normalized cardinality; and the number of Distinct Combinations (DC) of labels.

Table 1. Description of the multi-label data sets used in the experiments

Data	**Domain**	**N**	**M**	**L**	**LC**	**LD**	**DC**
Emotions	Music	593	72 n	6	1.869	0.311	27
Enron	Text	1702	1001 b	53	3.378	0.064	753
Genbase	Biology	662	1186 b	27	1.252	0.046	32
Medical	Text	978	1449 b	45	1.245	0.028	94
Scene	Image	2407	294 n	6	1.074	0.179	15
Slashdot-f	Text	3782	1079 b	22	1.180	0.041	156
Yeast	Biology	2417	103 n	14	4.237	0.303	198

We compared the three RF-based multi-label feature selection methods to two recently proposed ones: PPT-MI [7] and PMU [16]. PPT-MI is a multi-label feature selection method using the Pruned Problem Transformation (PPT) to improve the LP approach followed by a sequential forward selection with the

Mutual information (MI) as search criterion. PMU is a filter approach that takes into account label interactions in evaluating the dependency of given features without resorting to problem transformation. It is presented as a multivariate mutual information-based feature selection method for multi-label learning that naturally derives from mutual information between selected features and a set of labels. The classification performance of the five feature selection methods was measured using RAkEL multi-label classification algorithm [21]. We evaluated the performance of the methods using a 3-fold cross validation. In order to get reliable statistics over the performance metrics, the experiments were repeated 5 times. So the results obtained are averaged over 15 iterations which allows us to apply statistical tests in order to discern significant differences between the compared methods. Note that the LIBSVM (with linear kernel) is employed as the binary learner for classifier induction to instantiate RAkEL. The number of models m in RAkEL is set to $min(2 \times L, 100)$ for all datasets [17], the size of the label-sets k to half the number of labels $(L/2)$ [17] and the threshold value to 0.5. For PMU and PPT-MI, the numeric data sets are discretized using the Equal-width interval scheme, as suggested by the authors in [16]. Furthermore, the three variants of RF of multi-label learning (BRRF, RFLP and RFPCT) are tuned similarly. The number of variables to split on at each node and the committee size are set to $\sqrt{M}$, and 100, respectively.

In the multi-label classification problem, performance can be assessed by several evaluation measures. Here, we employed the subset accuracy measure (also called multi-label classification accuracy) defined as follow:

$$Subset\_Accuracy(h) = \frac{1}{|Te|} \sum_{i=1}^{|Te|} I(h(x_i) = \mathcal{Y}_i) \tag{2}$$

where $\mathcal{Y}_i$ is the set of true labels and $h(x_i)$ is the set of predicted labels. $|Te|$ is the cardinality of test data set and $I(true) = 1$ and $I(false) = 0$

This metrics takes values in the interval $[0; 1]$. The greater the value, the better the algorithm performance. Note that the subset accuracy implicitly takes into account the label correlations. It is therefore a very strict evaluation measure as it requires an exact match of the predicted and the true set of labels.

4.2 Comparison Results

This section presents the results obtained from our empirical study and concludes on the applicability and performance of RF for multi-label feature selection. Figure 1 plots the classification performance in terms of subset accuracy averaged over the 5x3 runs of the above five compared approaches against the 50 most important features. Due to space limitation, we only show experimental results for the four largest data sets. As may be observed, BRRF outperforms the other methods by generally achieving the highest subset accuracy values. On the other hand, RFLP perform the worst.

For the sake of completeness, we also averaged the subset accuracy averaged over the 15 runs for different numbers of selected features for an extensive

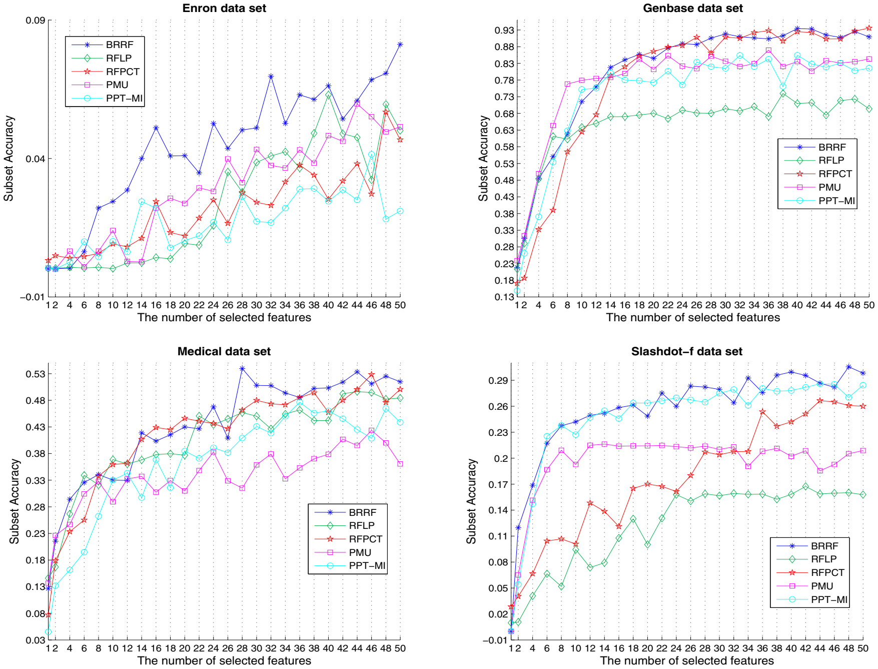

Fig. 1. Subset accuracy averaged over the 5x3 runs vs. different numbers of selected features on the four largest data sets

statistical analysis. The averaged metrics of the five feature selection methods over the top 50 features are depicted in Table 2. In order to better assess the results obtained for each feature selection algorithm on each metric, we adopt in this study the methodology proposed by [6] for the comparison of several algorithms over multiple datasets. In this methodology, the non-parametric Friedman test is firstly used to evaluate the rejection of the hypothesis that all the approaches perform equally well for a given risk level. It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2 etc. In case of ties it assigns average ranks. Then, the Friedman test compares the average ranks of the algorithms and calculates the Friedman statistic. If a statistically significant difference in the performance is detected, we proceed with a *post hoc* test. The Nemenyi test is used to compare all the classifiers to each other. In this procedure, the performance of two classifiers is significantly different if their average ranks differ more than some critical distance (CD). The critical distance depends on the number of algorithms, the number of data sets and the critical value (for a given significance level p) that is based on the Studentized range statistic (see [6] for further details). In this study, based on the values in table 2, the Friedman test reveals statistically significant differences ($p < 0.1$) for each metric. Furthermore, we present the

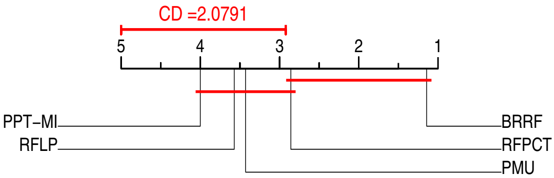

Fig. 2. Average rank diagram comparing the feature selection algorithms in terms of subset accuracy

result from the Nemenyi posthoc test with average rank diagram as suggested by Demsar [6]. This is given on Figure 2. The ranks are depicted on the axis, in such a manner that the best ranking algorithms are at the rightmost side of the diagram. The algorithms that do not differ significantly (at $p = 0.1$) are connected with a line. The critical difference CD is shown above the graph (here CD=2.0791).

Table 2. Subset accuracy averaged over the 5x3 runs and the 50 most important features for all algorithms and all data sets. Bottom row of the table presents the average rank of Subset accuracy mean used in the computation of the Friedman test

Data	**BRRF**	**RFLP**	**RFPCT**	**PMU**	**PPT-MI**
Emotions	0.266±0.05	0.253±0.04	0.260±0.05	0.245±0.04	0.248±0.05
Enron	0.046±0.01	0.024±0.00	0.022±0.01	0.029±0.01	0.019±0.01
Genbase	0.808±0.02	0.650±0.03	0.781±0.02	0.775±0.02	0.744±0.01
Medical	0.431±0.03	0.402±0.02	0.417±0.05	0.329±0.04	0.366±0.02
Scene	0.547±0.01	0.496±0.02	0.535±0.01	0.614±0.01	0.457±0.01
Slashdot	0.257±0.00	0.125±0.05	0.180±0.03	0.196±0.01	0.249±0.01
Yeast	0.157±0.01	0.155±0.01	0.155±0.01	0.139±0.01	0.152±0.01
Av Rank	1.1429	3.5714	2.8571	3.4286	4.0000

From Figure 2, we observe that BRRF performs significantly better than PMU, RFLP and PPT-MI, which seem to have equivalent performances. Although the average rank exhibit clear differences, the test doesn't allow us to conclude whether RFPCT is equivalent to BRRF or to the worst three methods. The Freidman test we use is known to be overly conservative. So to further exploit these rank comparisons, we compared, on each data set and for each pair of methods, the subset accuracy values obtained over the 15 iterations by using the paired t-test (with $\alpha = 0.1$). The results of these pairwise comparisons are depicted in Table 3 in terms of "Win-Tie-Loss" statuses of all pairs of methods; the three values in each cell (i, j) respectively indicate how times many the approach i is significantly better/not significantly different/significantly worse than the approach j. Following [6], if the two algorithms are, as assumed under the null-hypothesis, equivalent, each should win on approximately $n/2$ out of n data sets. The number of wins is distributed according to the binomial distribution and the critical number of wins at $\alpha = 0.1$ is equal to 6 in our case. Since tied matches support the null-hypothesis we should not discount them but split them evenly between the two classifiers when counting the number of wins; if there is an odd number of them, we again ignore one.

In Table 3, each pairwise comparison entry (i, j) for which the approach i is significantly better than j is boldfaced. The analysis of this table reveals that the approach that is never beaten by any other approach is BRRF.

Overall, these experiments confirm the ability of RF, that showed promising results for multi-label classification in [17], to rank the relevant features accurately in a multi-label context. More specifically, they suggest a relative superiority of the feature selection method built using the BRRF approach, compared with the ones that use RFPCT and RFLP. Indeed, it is more effective to use a RF that treats each label independently (*i.e.,* BRRF) rather than exploiting the underlying dependencies between labels (*i.e.,* RFPCT and RFLP) for evaluating feature importance in a multi-label setting. However, it was expected that methods that take the interaction among labels into consideration (*i.e.,* RFPCT and RFLP) would show better results than the ones using the BR approach (*i.e.,* BRRF). Nevertheless, this observation corroborate the previous finding in [20], namely that ignoring correlation among labels within the feature selection process doesn't affect the quality of the multi-label classification.

The superiority of BRRF compared to the remaining RF-based approaches (RFLP and RFPCT) in the feature selection process could be further motivated by the following reasons:

- The RFLP approach is based on the LP algorithm which suffers from class size issues, *i.e.,* the large number of label sets appearing in the training set (class values for the single-label classifier of LP), makes the learning task quite hard as many of these label sets are usually associated with very few training examples [21] giving rise to a poor feature importance estimation in a wrapper way. Although the Pruned Problem Transformation were used to avoid this problem of creating too many rarely classes in RFLP, the latter remains inefficient and does not give competitive results.
- With RFPCT, the classification error does not vary significantly when the values of a specific feature are randomly permuted. Indeed, we noticed that the label errors often compensate each other. This is why the classification error vary moderately after shuffling a variable. This issue worsen as the number of labels is increased. To confirm this observation from an experimental point of view, we analyzed the average gap between classification error before and after the variable shuffling in Equation 1. We observed error variations of the magnitude of 10^{-7} on the data sets with a large number of labels (*e.g.* Enron, Medical).

Table 3. Pairwise t-test comparisons of FS methods in terms of Subset accuracy. Bold cells (i, j) highlights that the approach i is significantly better than j according to the sign test at $\alpha = 0.1$.

	BRRF	**RFLP**	**RFPCT**	**PMU**	**PPT-MI**
BRRF	–	**7/0/0**	**5/2/0**	**6/0/1**	**7/0/0**
RFLP	0/0/7	–	0/4/3	2/1/4	4/1/2
RFPCT	0/2/5	3/4/0	–	2/2/3	5/1/1
PMU	1/0/6	4/1/2	3/2/2	–	3/1/3
PPT-MI	0/0/7	2/1/4	1/1/5	3/1/3	–

5 Conclusion

This work proposed and experimentally evaluated three wrapper multi-label feature selection methods, which use the RF paradigm: BRRF, RFLP and RF-PCT. These extensions differ in the way they consider label dependence within the feature selection process. The performance of the methods were compared against recently proposed approaches using seven benchmark multi-label data sets emerging from different domains. The result of this evaluation is three-fold: 1) Random Forest handles accurately the feature selection in a multi-label context and; 2) Surprisingly, BRRF appears more suitable for multi-label feature selection, as taking into account relationships between labels was not shown remarkably effective for multi-label feature selection using the RF paradigm.

Future work will be conducted to assess the stability of the feature selection methods [19] when noise is added to the data. We will also investigate the effectiveness of using label-specific feature selection [23] in the multi-label learning process. This is currently being undertaken and will be reported in due course.

References

1. Barkia, H., Elghazel, H., Aussem, A.: Semi-supervised feature importance evaluation with ensemble learning. In: ICDM 2010, pp. 31–40 (2011)
2. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: ICML, pp. 55–63 (1998)
3. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
4. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
5. Dembczynski, K., Waegeman, W., Cheng, W., Hüllermeier, E.: On label dependence and loss minimization in multi-label classification. Machine Learning 88(1-2), 5–45 (2012)
6. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
7. Doquire, G., Verleysen, M.: Feature selection for multi-label classification problems. In: Cabestany, J., Rojas, I., Joya, G. (eds.) IWANN 2011, Part I. LNCS, vol. 6691, pp. 9–16. Springer, Heidelberg (2011)

8. Elghazel, H., Aussem, A.: Unsupervised feature selection with ensemble learning. Machine Learning, 1–24 (2013)
9. Gu, Q., Li, Z., Han, J.: Correlated multi-label feature selection. In: CIKM, pp. 1087–1096 (2011)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
11. Hong, Y., Kwong, S., Chang, Y., Ren, Q.: Consensus unsupervised feature ranking from multiple views. Pattern Recognition Letters 29(5), 595–602 (2008)
12. Hong, Y., Kwong, S., Chang, Y., Ren, Q.: Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. Pattern Recognition 41(9), 2742–2756 (2008)
13. Kocev, D., Slavkov, I., Dzeroski, S.: More is better: Ranking with multiple targets for biomarker discovery. In: 2nd International Workshop on Machine Learning in Systems Biology, p. 133 (2008)
14. Kocev, D., Slavkov, I., Dzeroski, S.: Feature ranking for multi-label classification using predictive clustering trees. In: International Workshop on Solving Complex Machine Learning Problems with Ensemble Methods, in Conjunction with ECML/PKDD, pp. 56–68 (2013)
15. Kocev, D., Vens, C., Struyf, J., Dzeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recognition 46(3), 817–833 (2013)
16. Lee, J.-S., Kim, D.-W.: Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters 34(3), 349–357 (2013)
17. Madjarov, G., Kocev, D., Gjorgjevikj, D., Dzeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recognition 45(9), 3084–3104 (2012)
18. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning 85(3), 333–359 (2011)
19. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
20. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electr. Notes Theor. Comput. Sci. 292, 135–151 (2013)
21. Tsoumakas, G., Katakis, I., Vlahavas, I.P.: Random k-labelsets for multilabel classification. IEEE Trans. Knowl. Data Eng. 23(7), 1079–1089 (2011)
22. Tsoumakas, G., Xioufis, E.S., Vilcek, J., Vlahavas, I.P.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research 12, 2411–2414 (2011)
23. Zhang, M.-L.: Lift: Multi-label learning with label-specific features. In: IJCAI, pp. 1609–1614 (2011)
24. Zhang, M.-L., Zhou, Z.-H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering 18(10), 1338–1351 (2006)
25. Zhang, M.-L., Zhou, Z.-H.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering 99(PrePrints):1 (2013)