Artif Intell Rev
DOI 10.1007/s10462-016-9516-4

Categorizing feature selection methods for multi-label classification

Rafael B. Pereira[1] · **Alexandre Plastino**[1] · **Bianca Zadrozny**[2] · **Luiz H. C. Merschmann**[3]

Abstract In many important application domains such as text categorization, biomolecular analysis, scene classification and medical diagnosis, examples are naturally associated with more than one class label, giving rise to multi-label classification problems. This fact has led, in recent years, to a substantial amount of research on feature selection methods that allow the identification of relevant and informative features for multi-label classification. However, the methods proposed for this task are scattered in the literature, with no common framework to describe them and to allow an objective comparison. Here, we revisit a categorization of existing multi-label classification methods and, as our main contribution, we provide a comprehensive survey and novel categorization of the feature selection techniques that have been created for the multi-label classification setting. We conclude this work with concrete suggestions for future research in multi-label feature selection which have been derived from our categorization and analysis.

Keywords Multi-label learning · Feature selection · Classification · Data mining

✉ Rafael B. Pereira
rbarros@ic.uff.br

Alexandre Plastino
plastino@ic.uff.br

Bianca Zadrozny
biancaz@br.ibm.com

Luiz H. C. Merschmann
luizhenrique@iceb.ufop.br

[1] Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil

[2] IBM Research - Brazil, Rio de Janeiro, RJ, Brazil

[3] Universidade Federal de Ouro Preto (UFOP), Ouro Preto, MG, Brazil

Published online: 23 September 2016

1 Introduction

A large body of research in supervised learning deals with the analysis of single-label data, where instances are associated with a single label from a set of class labels (Tsoumakas and Katakis 2007). More specifically, the single-label classification problem can be stated as the process of predicting the class label of new instances described by their feature values.

However, in many important data mining applications, such as text categorization, biomolecular analysis, scene classification and medical diagnosis, the instances are associated with more than one class label. This characterizes the multi-label classification problem, a recent and relevant topic of research, that has become a very common real-world task (Zhang and Zhou 2007).

In a broad way, two groups of classification strategies have been proposed to deal with multi-label data. In the first group, the multi-label data is converted into single-label data and then the classification problem is solved using single-label classifiers. The second group is related with proposals for adapting or extending single-label classifiers to cope with multi-label data. In the former group one can find popular methods like label powerset and binary relevance transformations, and in the latter group common adaptations are: multi-label k-nearest neighbors, multi-label Naive Bayes, multi-label AdaBoost, among others (Tsoumakas and Katakis 2007; de Carvalho and Freitas 2009; Gibaja and Ventura 2015).

The performance of a classification method is closely related to the inherent quality of the training data. Redundant and irrelevant features may not only decrease the classifier's accuracy but also make the process of building the model or running the classification algorithm slower. Feature selection is a data preprocessing step which aims at identifying relevant features for a target data mining task—specifically in this work, the classification task. Feature selection techniques are usually applied for removing from the training set features that do not contribute to, or even decrease, the classification performance (Guyon et al. 2008; Liu and Motoda 2008a). There is an extensive literature regarding feature selection for single-label classification, which has been summarized in surveys such as in Dash and Liu (1997), Guyon et al. (2008), Molina et al. (2002), Khalid et al. (2014) and Chandrashekar and Sahin (2014).

In the last few years, given the increasing popularity of multi-label classification, there has been significant research specifically in the area of feature selection for multi-label classification. Spolaôr et al. (2015) indicates the need to consider a taxonomy specific for multi-label feature selection. With the intent of organizing the current knowledge in this specific area and pointing to directions for future work, in this paper we review the current literature in feature selection for multi-label classification. Our main contribution consists of a taxonomy for multi-label feature selection along with a categorization and review of existing techniques in light of the proposed taxonomy.

The remainder of this paper is organized as follows. Section 2 reviews previous work and revisits a categorization of the multi-label classification problem. In Sect. 3, we describe the feature selection task in detail, and propose a categorization for the existing multi-label feature selection techniques. Finally, in Sect. 4, we make our concluding remarks and point to directions for future work.

2 Multi-label classification

In general, the classification task can be stated as the process of predicting one or more class labels for an instance described by a vector of feature values, given a training set where each

instance is described by a vector of features and by one or more class labels. Traditional classification is performed as a single-label task, where each data instance is associated with a single class label. Well-known single-label classification techniques include decision trees (Quinlan 1986, 1993), k-Nearest Neighbors (k-NN) (Cover and Hart 1967; Dasarathy 1991), Naive Bayes (Duda et al. 2001), neural networks (Ripley 1996), associative classifiers (Liu et al. 1998), Support Vector Machines (SVM) (Burges 1998) and others.

On the other hand, in the multi-label classification task, each data instance may be associated with multiple labels. Multi-label classification is suitable for many domains such as text categorization, scene and video classification, medical diagnosis, applications in microbiology (Read 2010), and it is also a challenging problem in bioinformatics (Li et al. 2010a). In all these cases, the task is to assign for each unseen instance a label set whose size is unknown a priori (Zhang and Zhou 2007).

The strategies proposed to deal with multi-label classification rely mainly on problem transformation, where the multi-label problem is transformed into a set of one or more single-label problems, and on algorithm adaptation, where the single-label learning algorithms are adapted to handle multi-label data directly (de Carvalho and Freitas 2009; Tsoumakas et al. 2010; Zhang and Zhou 2014). Both paradigms are presented in the next two sections.

2.1 Strategies based on data set transformation

The simplest way to apply a classification strategy to a multi-label data set is to transform it into single-label data set. Then a traditional classification technique—like k-NN or decision tree—can be employed to perform the classification task. This way, the transformation technique allows the usage of one or more single-label classification algorithms, which have been thoroughly studied and perfected over the last decades.

A simple transformation technique used to convert a multi-label data set into a single-label one consists in selecting for each instance just one label from its multi-label subset. This label can be the most frequent label in the data set (select-max), the least frequent label (select-min) or a random label (select-random) (Chen et al. 2007; Tsoumakas and Katakis 2007). Another option is to simply discard every multi-label example (select-ignore), although this is not useful if most of the data set consists of multi-label instances.

Another type of transformation consists in copying each multi-label instance n times, where n is the number of labels assigned to that instance. Each copied instance is then assigned one distinct single label from the original set. A variation of the copy transformation is the copy-weight, which associates a weight $1/n$ to each copied instance, according to the number n of labels of the original instance (Tsoumakas et al. 2010). This variation can only be employed if the classifier is able to handle weighted instances.

The drawback of using the simple transformation approach is that it treats each label of a data instance independently and thus it results in loss of information about label dependency, which can be essential to achieve good performance in some multi-label classification problems.

Label powerset (LP) is another kind of transformation which creates one label for each different subset of labels that exists in the multi-label training data set. Thus, the new set of labels corresponds to the powerset of the original set of labels. After this transformation process, a single-label classification algorithm can handle the transformed data set. This classifier can then be used to assign one of these new labels to new instances, which can then be mapped back to the corresponding subset of the original labels (Tsoumakas and Vlahavas 2007).

Label powerset is recommended only for data sets with a small number of labels, as the possible powerset combinations are 2^L, where L is the number of distinct labels in the data set. For data sets with a large number of labels, the resulting powerset data tends to become sparse and therefore making it harder for the classifier to work (Dembczyński et al. 2012).

The original label powerset technique has been extended and improved in subsequent work. Two variations are the pruned problem transformation (PPT), proposed in Read (2008), which prunes away label sets that occur fewer times; and the random k-labelsets (RAKEL), proposed in Tsoumakas and Vlahavas (2007), which constructs an ensemble of LP classifiers trained using different and small random subsets of the set of labels (Tsoumakas et al. 2010).

Binary relevance (BR) is a transformation technique that produces a binary classifier for each different label of the original data set. The method is called "binary relevance", because each label is classified as relevant or non-relevant for an instance. The data transformation is applied to the multi-label data set, generating L single-label data sets, where L is the number of distinct labels in the original data set. Each single-label classifier yields a positive or negative result for each instance. The BR classification result is the union of the labels that are positively predicted by each classifier, as each one is capable of predicting one single label. As binary relevance learns a single binary model for each different label, it has linear complexity with respect to the number of labels (Tsoumakas et al. 2009).

Binary relevance does not take into account label correlations. Without this information, some relevant label dependences will not be considered (Tsoumakas et al. 2009). In order to minimize this drawback, several techniques have been proposed to extend and improve the binary relevance technique (Fürnkranz et al. 2008; Godbole and Sarawagi 2004; Hüllermeier et al. 2008; Mencía and Furnkranz 2008; Read et al. 2009).

2.2 Strategies based on algorithm adaptation

Most traditional classification algorithms employed in single-label problems have been adapted to the multi-label paradigm (Tsoumakas et al. 2010). The C4.5 decision-tree learning algorithm has been adapted to handle multi-label data in Clare and King (2001) by allowing multiple labels in the leaves of the tree. An SVM algorithm that minimizes the ranking loss metric has been proposed in Elisseeff and Weston (2001). A multi-label adaptation of the Naive Bayes algorithm was proposed in Zhang et al. 2009. Multi-class, Multi-label Associative Classification (MMAC) is an algorithm that follows the paradigm of associative classification which deals with the construction of multi-label classification rule sets using association rule mining (Tsoumakas and Katakis 2007).

Several k-NN adaptations were proposed Tsoumakas et al. (2010), and one of them is the Multi-label k-NN (ML-KNN) (Zhang and Zhou 2007). For each unseen instance, it identifies the k nearest neighbors in the training set. Then, based on statistical information gained from the subset of labels of these neighboring instances, the maximum a posteriori (MAP) principle is employed to determine the label set for the unseen instance.

Instance Based Learning by Logistic Regression (IBLR) (Cheng and Hüllermeier 2009) is an adaptation which combines the instance-based learning concept of the k-NN algorithm with logistic regression. It also considers the labels of neighboring instances as features, in order to aid the classification.

Another algorithm that can be considered as an adaptation is the BR + KNN using a single search instead of transforming the multi-label data set using the BR approach. It finds the k nearest neighbors and at the same time it makes independent predictions for each label (Sorower 2010). While BR followed by k-NN has a computational complexity of L times the cost of computing the k nearest instances, where L is the number of labels in the data

set, this adaptation runs much faster (linear complexity), and is more scalable than other classification algorithms based on transformation.

3 Categorization for multi-label feature selection

According to Guyon and Elisseeff (2006), feature selection techniques are employed to identify relevant and informative features, primarily to improve the classifier predictive accuracy. In general, besides this main goal, there are other important motivations: the reduction and simplification of the data set, the acceleration of the classification task and the simplification of the generated classification model.

Traditional feature selection techniques can generally be categorized into three approaches: embedded, wrapper or filter (Liu and Motoda 2008b). Embedded strategies are incorporated into the algorithm responsible for the induction of the classification model. Wrapper and filter strategies are performed in a preprocessing phase and they search for the most suitable feature set to be used by the classification algorithm or by the classification model inducer. In wrapper feature selection, the adopted classification algorithm itself is used to evaluate the quality of candidate feature subsets, while in filter feature selection, feature quality is evaluated independently from the classification algorithm using a measure which generally considers the features and class label distributions.

Feature selection techniques intended specifically for multi-label classification have been developed in recent years. Even though there are many proposals on this topic, it is still considered an active research area (Doquire and Verleysen 2011; Spolaôr et al. 2013).

The main contribution of this work is providing a comprehensive survey and a taxonomy of multi-label feature selection techniques. Figure 1 shows our proposed taxonomy for categorizing multi-label feature selection. It aims at categorizing the feature selection techniques according to characteristics inherent to the multi-label paradigm.

This taxonomy is composed of two main categories based on the multi-label classification paradigms already explained in our work: transformation-based methods and direct methods. The transformation-based and direct categories, as well as their subcategories, are described in the next sections.

3.1 Multi-label feature selection based on transformation

The simplest way to employ feature selection to a multi-label data set is to change it into a single-label data set and apply a traditional feature selection technique. There are plenty of algorithms to transform a multi-label data set into a single-label one. These methods to transform a multi-label data set into single-label data were described in the previous section in the multi-label classification context. In the next subsections, we review them in the feature selection context.

3.1.1 Strategies based on single data transformation

Single data transformation for multi-label feature selection consists in changing the multi-label data into one single-label data set and then applying a traditional feature selection technique. Single data transformation encompasses both simple and label powerset transformations.

The following common simple transformation techniques: select-max, select-min, select-random, select-ignore, copy and copy-weight; and the label powerset transformation, used

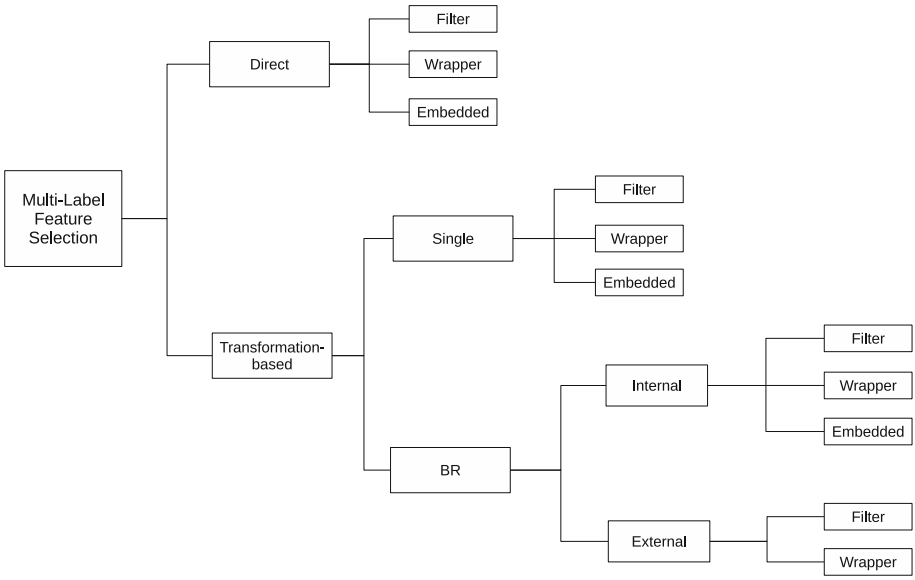

Fig. 1 Taxonomy proposed for multi-label feature selection

to convert a multi-label data set into a single-label one were described in Sect. 2.1. These transformations have also been employed to perform feature selection over multi-label data.

Figure 2 presents a feature selection model to represent this category of transformations applied to the multi-label data. It initially converts the original multi-label data into a single-label data set using one of the transformations. Then a traditional single-label feature selection is employed to the data. The output of this process is a list of the selected features. Optionally, a subsequent process—indicated with dashed lines—can be employed to deliver the original multi-label data containing only the corresponding selected features. This way a multi-label classifier can be used to perform its predictions over the multi-label data.

Chen et al. (2007) used these data set transformations to allow the application of traditional feature selection techniques to the text categorization problem. According to the model in Fig. 2, the multi-label data were transformed into a single-label data set after executing the following simple transformations: copy, select-ignore, select-max and select-min. They also proposed a new transformation—from multi-label into single-label data—based on the entropy measure, which reweights each instance using this measure as a variation of the copy-weight transformation described before.

Note that after employing a feature selection technique, it is possible to deliver to the classification algorithms either the transformed single-label data set, or the original multi-label data set maintaining only the corresponding selected features. Nonetheless, as these feature selection techniques based on a simple data transformation disregard the correlation between labels or subsets of labels, they might fail to identify labels that have a strong relationship with each other, like dependency or co-occurrence.

The label powerset transformation is also directly applied to the task of multi-label feature selection based on transformation, as it is capable of delivering a single-label data set with each subset of labels converted into a new class label.

Trohidis et al. (2008) evaluated and compared several multi-label classification strategies for the task of automated decision of emotion in a music data set. For the empirical evaluation

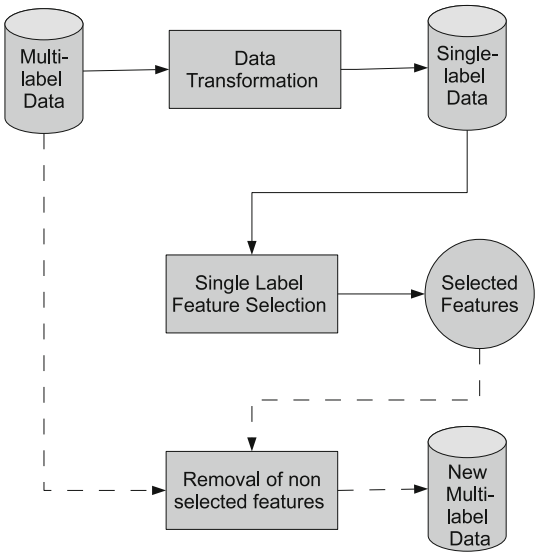

Fig. 2 Transformation-Based/Single multi-label feature selection

of feature selection, the use of a label powerset transformation was proposed to produce a single-label data set, and then a common feature selection measure was employed (χ^2 statistic) to select the best features. They verified that, for the evaluated data set, using the ML-KNN algorithm (Zhang and Zhou 2007) as the classification algorithm and the label powerset to apply the feature selection achieved a better Hamming Loss result than without feature selection.

The label powerset transformation was also used for feature selection in Spolaôr et al. (2013), in conjunction with the ReliefF and information gain measures. With this feature selection, it was possible to reduce the dimensionality of the data sets without compromising the classification performance.

The label powerset transformation tends to create too many classes, causing overfitting and imbalance problems (Lee and Kim 2013). Doquire and Verleysen (2011) used the pruned problem transformation (PPT) (Read 2008), an improvement over the label powerset, for multi-label feature selection on three real-world data sets from different domains: gene, semantic scene and emotion (in music) classification. Then a multi-label k-NN algorithm was employed over the original multi-label data containing only the selected features. When compared with the χ^2 statistic adopted in Trohidis et al. (2008), and also with a non feature selection scenario, the mutual information measure allowed the classification phase to achieve a better result in terms of the Hamming Loss and the accuracy of the classifier.

Reyes et al. (2015) presented three extensions of the ReliefF algorithm that work in the multi-label learning context. One of the extensions was the PPT-ReliefF, which uses the PPT problem transformation method to convert the multi-label problem into a single-label problem. The results were compared with the BR-ReliefF, LP-ReliefF and two adaptations (ReliefF-ML and RReliefF-ML) proposed in the same work, and indicated an improvement over preceding results.

Traditional single-label feature selection techniques can be categorized as wrapper, embedded or filter. An algorithm from any of these categories can be applied after a single data transformation. However, all publications reviewed in this subsection are categorized as

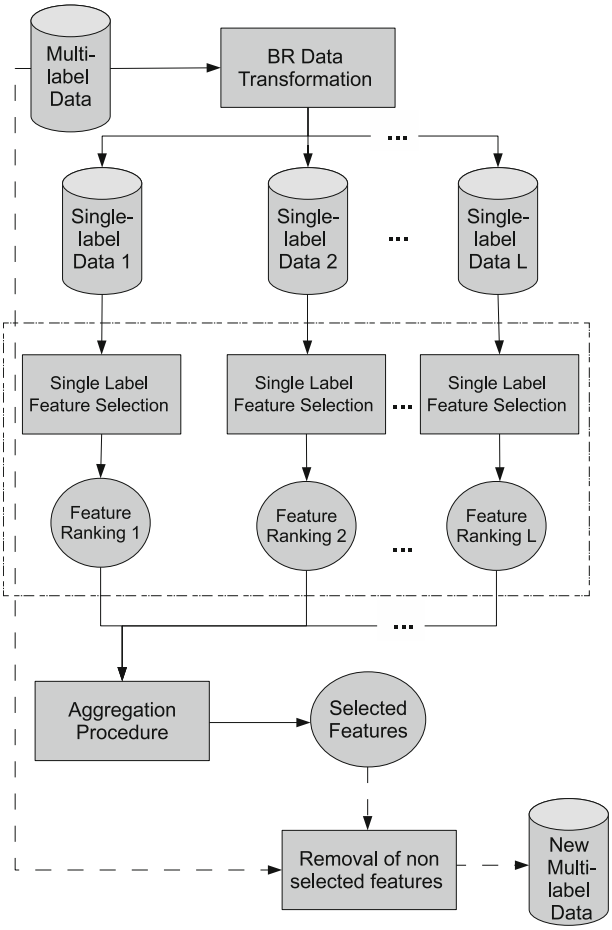

Fig. 3 Transformation-Based/Binary-Relevance multi-label feature selection

Transformation Based/Single/Filter. This means that there is a lack of work evaluating single-label embedded and wrapper feature selection techniques for multi-label classification.

3.1.2 Strategies based on binary relevance transformation

The process of transforming a multi-label data set into several single-label ones was explained in Sect. 2.1. The same technique can be employed for feature selection. For each different label in the original data set, a binary single-label data set is created, and then feature selection is executed.

Figure 3 represents a feature selection model based on the binary relevance (BR) transformation. Each label from the data set is considered individually in order to perform the feature selection. Then the single-label feature selection is applied once for each single-label data set.

There are two ways to handle the feature selection result on a BR approach. The first one is to combine the feature selection result of each binary model into a single output, which we call

the *External* approach. Then the reduced data set is given as input to a multi-label classifier. In this case there is the need to aggregate the feature selection results before classification.

Another way to handle the feature selection result of each binary model is to apply the classification method directly to each single-label data set obtained after the feature selection step. We call this the *Internal* approach. As the multi-label data set is transformed into a single-label data set, both the classifier and the feature selection techniques are able to handle the data. After the feature selection, each reduced single-label data set will serve as input for a single-label classifier. After the classification step, the results are combined analogously to a BR approach for multi-label classification.

The process of combining the lists of features is also known as aggregation. This is the approach shown in Fig. 3. After the aggregation process, indicated by a dashed line, there is an optional step of removing the features from the original multi-label data set to produce a corresponding data set with the chosen features only.

A typical way to output a list of selected features is ensuring a score threshold or a fixed number of features across the rankings (e.g., the top 500 features). Other ways to combine the multiple feature rankings produced by the binary models is to consider the overall maximum score or the average score of each feature across the binary models (Trohidis et al. 2008). The feature selection used in this External strategy can be a filter or a wrapper technique.

Forman (2004) proposed a round robin aggregation method to a BR External strategy, which considers the best features of each binary model in sequence, and a variation named rand-robin, that selects the best features randomly with probabilities according to the frequency of each label in the original data set.

Yang and Pedersen (1997) evaluated common feature selection measures (document frequency, information gain, mutual information, χ^2 statistic and term strength) in a text categorization multi-label problem. Each label was evaluated separately, which is equivalent to an external binary relevance transformation. After applying the feature selection to this data set, the k-NN classification technique was employed. Up to 98 % of the features were removed without losing categorization accuracy, when using the information gain and χ^2 techniques; the same when 90 % of features were removed with the document frequency metric; and 50–60 % with term strength. Mutual information achieved an inferior performance compared to the other methods.

Some text classification work (Olsson and Oard 2006; Zheng et al. 2004) employed the binary relevance technique to apply single-label feature selection measures, like information gain and χ^2 statistic.

Rogati and Yang (2002) applied several filter feature selection techniques in text categorization data sets. Again, each label was considered individually, which is equivalent to a BR transformation. Then the following feature selection measures were applied to the data sets: document frequency, information gain, a binary version of information gain and the χ^2 statistics. From the resulting feature ranking of each measure, both the average and the maximum value were considered as an aggregated score. The empirical results showed by Rogati and Yang (2002) suggested that combining the use of multiple feature selection was advantageous for eliminating rare words in a consistent way across different classifiers. In the experimental evaluation by Doquire and Verleysen (2011), the maximum and average aggregation strategies were also used for the BR.

The BR transformation was also used for feature selection by Spolaôr et al. (2013), in conjunction with the ReliefF and information gain measures. This feature selection strategy was compared to the LP transformation using the same measures, reaching the conclusion that both methods achieved a similar performance in the experiments with data sets from various domains commonly used in multi-label work.

Spolaôr and Tsoumakas (2013) used BR to apply feature selection in conjunction with several aggregation techniques to data sets from the text categorization domain. The best results were achieved by using the maximum score across all labels with the χ^2 measure.

Tsoumakas and Vlahavas (2007) proposed the RAKEL method and evaluated it on three data sets from different domains (semantic scene, gene and textual classification). In the data transformation step, RAKEL constructs an ensemble of label powersets. In order to reduce the computational cost of training, a BR feature selection was applied to the textual data set. The χ^2 statistic was used separately for each label in order to obtain different rankings of all features, and in the aggregation step the top 500 features were selected (i.e., the features with the highest score over all labels). This same label-based approach was applied by Read (2010) for a text-categorization data set (Reuters) in conjunction with the information gain measure.

Dendamrongvit et al. (2011) used BR with the Internal strategy, as it was noticed that for text categorization, each label is characterized by a different set of features. Thus, an appropriate feature selection technique was applied separately to each label, and the output was sent to a k-NN classifier in the first experiment and an SVM for the second, both assessing text data sets. The results confirmed that the chosen feature selection increased the performance of the text categorization.

Wandekokem et al. (2010) builds a binary SVM for each label, selecting features for each predictor by employing a wrapper approach with the Area Under the ROC Curve (AUC) to estimate the performance of a candidate feature set. Two search strategies were used for the wrapper: Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). The ensemble of SVMs classifiers was used to assess data from real-world industrial machines, concluding that feature selection increased the individual accuracy of the SVM classifiers.

As it occurs with single transformation-based feature selection, there is a lack of work evaluating embedded feature selection techniques after a BR transformation. Most techniques described in this subsection are categorized as Transformation-Based/BR/Filter. Among the filters, all are subcategorized as External, except for Dendamrongvit et al. (2011), which is categorized as Internal. There is also one BR transformation followed by a wrapper technique.

As it occurs with classification, the use of the binary relevance transformation for feature selection can result in a loss of information because it ignores label dependences. This is also an important issue in multi-label feature selection according to Spolaôr and Monard (2014).

3.1.3 Summary of publications on transformation-based feature selection

Table 1 shows publications related to multi-label feature selection that rely on data transformation. The "Data Transformation" column specifies which transformation technique described in our taxonomy was used. In the case of the binary relevance transformation, we also specify how the multiple lists of features were combined (indicated by the '+' sign): either using the average or maximum score, in the case of the *Internal* approach, or selecting a specific number of top features, in the case of the *External* approach.

The "Feature Selection" column indicates which feature selection technique was used – all of them single-label techniques, relying on the data transformation executed before. The "Classification Algorithm" column shows which classification strategy was employed, in some cases preceded by some data transformation technique, indicated by the '+' sign (e.g., RAKEL + SVM). Finally, the "Data Sets" column lists the data sets used and to which domains they belong, in parentheses.

We observe that most of the publications employed a data transformation technique from just one paradigm (simple transformations, label powerset-based or binary relevance-based).

Table 1 Summary of publications on multi-label feature selection based on transformation

References	Data transformation	Feature selection	Classification algorithm	Data sets
Chen et al. (2007)	Single/Filter/copy	Information gain	SVM	Reuters-21578 (text)
	Single/Filter/select-ignore	χ^2 statistic		Reuters RCV1-v2 (text)
	Single/Filter/select-min	OCFS		
	Single/Filter/select-max			
	Single/Filter/entropy-based			
Yang and Pedersen (1997)	BR/External/Filter	Document frequency	k-NN	OSHUMED (text)
	+top features	Information gain	Regression (LLSF)	Reuters-22173 (text)
		Mutual information		
		χ^2 statistic term strength		
		Term strength		
Trohidis et al. (2008)	Single(LP)/Filter	χ^2 statistic	ML-KNN	Emotions (music)
	BR/External/Filter			
	+avg and max			
Zheng et al. (2004)	BR/External/Filter	χ^2 statistic	Naive Bayes	Reuters-21578 (text)
		Information gain	Logistic regression	
		Correlation coefficient odds ratio		
		Odds ratio		
Doquire and Verleysen (2011)	Single(PPT)/Filter	Mutual information	ML-KNN	Yeast (gene)
	Single(LP)/Filter	χ^2 statistic	SVM	Scene (image)
				Emotions (music)
Olsson and Oard (2006)	BR/External/Filter	Document frequency	k-NN	RCV1-v2 (text)
	+max, avg and min	χ^2 statistic		
		Information gain		

Table 1 continued

References	Data transformation	Feature selection	Classification algorithm	Data sets
Rogati and Yang (2002)	BR/External/Filter	Document frequency	k-NN	Reuters-21578 (text)
	+avg and max	Information gain	Naive Bayes	Reuters RCV1 (text)
		Binary information gain	SVM	
		χ^2 statistic	Rocchio	
Tsoumakas and Vlahavas (2007)	BR/External/Filter	χ^2 statistic	RAKEL + SVM	tmc2007 (text)
	+top 500 features		BR + SVM	
Dendamrongvit et al. (2011)	BR/Internal/Filter	Information gain	BR + kNN	Reuters RCV1 (text)
			BR + SVM	EUROVOC (text)
Read (2010)	BR/External/Filter	Information gain	BR + Naive Bayes	Reuters RCV1 (text)
	+top 500 features		LP + Naive Bayes	
Spolaôr et al. (2013)	Single(LP)/Filter	Information gain	BRKNN	Various domains
	BR/External/Filter	ReliefF		
Spolaôr and Tsoumakas (2013)	BR/External/Filter	χ^2 statistic	BR + SVM	Various domains
	+avg,max,round/rand-robin	bi-normal separation		
Sechidis et al. (2014)	Single(LP)/Filter	MI maximization	ML-KNN	Scene (image)
	BR/External/Filter	joint MI max.		Yeast (gene)
	+max			
Reyes et al. (2015)	Single(PPT)/Filter	ReliefF	ML-KNN	Various domains
			BRKNN	
Wandekokem et al. (2010)	BR/External/Wrapper	Area under the curve	Ensembles of SVM	Industrial machine data

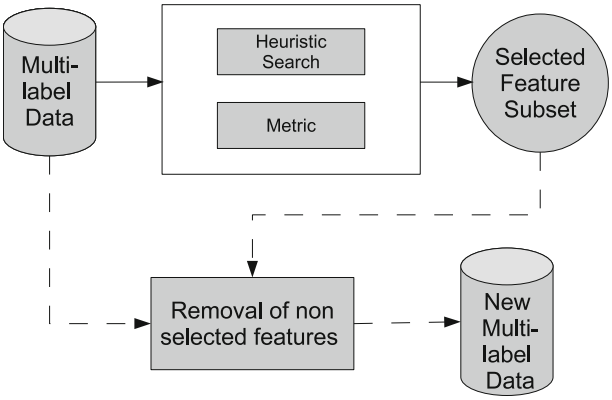

Fig. 4 Direct/Filter multi-label feature selection

The BR approach is usually External. The feature selection strategies used are simple filters that evaluate one feature at a time. The aggregation of partial results (subset of features) given by a BR model using a single-label feature selection algorithm is an unexplored topic. Furthermore, most of feature selection strategies aim to evaluate one or two classifiers, using data sets from just one or a few multi-label domains.

3.2 Direct multi-label feature selection

Several feature selection techniques were proposed to deal directly with the multi-label data. They consist mostly of algorithm adaptations of well-known feature selection techniques. Unlike the previous categories, in this case there is no transformation of the multi-label data.

We will categorize these multi-label feature selection techniques in three sub-categories: *Filter*, *Wrapper* and *Embedded*, in the same way they are categorized for their single-label counterparts (Liu and Motoda 2008b), and according to our proposed taxonomy.

3.2.1 Strategies based on the filter category

Filter strategies generally use an evaluation function which depends only on the properties of the data set, so they are independent of any particular learning algorithm.

Figure 4 illustrates this approach, which typically employs a heuristic search strategy and a metric able to evaluate subsets of features. The heuristic search can be also a ranker which evaluates each feature individually by a specific metric. Afterwards, the ranking is processed to output the selected features, either by establishing a metric value threshold or selecting the top n features from the ranking. Then this result can be combined with the original multi-label data to produce a new data set with only the selected features.

Lastra et al. (2011) extended the well-known technique Fast Correlation-Based Filter (FCBF), introduced by Yu and Liu (2004), to handle multi-label data. The technique consists in transforming the data set into a directed graph and applying the symmetrical uncertainty measure to evaluate the features of the data set. This feature selection was applied in conjunction with the IBLR (Cheng and Hüllermeier 2009) and Ensemble Classifier Chains (Read et al. 2011) classification algorithms, and data sets from multiple domains were evaluated.

Some feature extraction techniques were adapted from single-label counterparts, like Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI). They produce

a ranking of features after applying a technique to reduce the number of features, either by removing irrelevant features, or by creating a projection of the feature space. For instance, Yu et al. (2005) proposed the Multi-label Latent Semantic Indexing (MLSI). It is a feature extraction technique based on dimensionality reduction, as an extension of the LSI technique to make use of multi-label information. Feature extraction is a task different from feature selection (Liu and Motoda 2008a), so it is not the focus on this work.

Pereira et al. (2015) proposed a multi-label filter adaptation based on the information gain measure. The technique was evaluated on various multi-label data sets and coupled with ML-KNN, BR-KNN, Classifier Chains and other classification algorithms. It achieved an overall better result than the LP and copy transformations, and competitive results against the BR transformation. For an analysis of scalability, these algorithms were assessed with data sets from the Yahoo directory (more than 30,000 features), and the multi-label information gain adaptation outperformed the other transformation-based techniques. Li et al. (2014) proposed an adaptation based on the information gain measure, and the experimental results also confirmed it as an effective approach compared with other feature selection techniques.

Similarly, Zhang and Zhou (2010) proposed the MDDM method—Multi-label Dimensionality reduction via Dependence Maximization. It consists of a dimensionality reduction method, like PCA, aimed to the multi-label domain. It creates a ranking of features by maximizing the dependence between the features and the associated class labels using a well-known dependence measure. It was compared with other similar methods like PCA and MLSI coupled with the Multi-Label k-NN classifier and eleven Yahoo web-pages data sets. These experiments validate the performance of MDDM.

Filter strategies can also consider subsets of features instead of single features. After a number of iterations, the feature subset with the best metric value is selected. Like in other multi-label feature selection techniques, a subsequent process can be employed to deliver the original multi-label data with the corresponding selected features.

Kong and Yu (2012) proposed a new multi-label feature selection technique designed for graph classification, called gMLC. It is based on an efficient search for optimal subgraph features for graph objects with multiple labels, and evaluates each subset with a particular criteria. Graph data sets were evaluated with this method, which was compared with a BR transformation coupled with the information gain measure, and also with a technique that selects the top k-frequent subgraph features. The results favored the proposed gMLC technique.

Common single-label feature selection techniques were adapted to the multi-label paradigm recently. The ReliefF measure was adapted by Pupo et al. (2013) and Spolaôr and Monard (2014). The mutual information measure was adapted by Lee and Kim (2013). The correlation-based feature selection technique, capable of handling subset of features, instead of individual features, was adapted to the multi-label setting by Jungjit et al. (2013).

As described in Sect. 3.1.1, Reyes et al. (2015) presented three extensions (ReliefF-ML, RReliefF-ML and PPT-ReliefF) of the ReliefF algorithm that work in the multi-label learning context. The ReliefF-ML and RReliefF-ML adapted the classic ReliefF algorithm in order to handle the multi-label data directly. The results were compared with the BR-ReliefF, LP-ReliefF and PPT-ReliefF, indicating that the adaptations improved preceding results.

3.2.2 Strategies based on the wrapper category

The wrapper approach for feature selection (Kohavi and John 1997) consists in a method that searches for a relevant subset of features employing a classification technique to evaluate it.

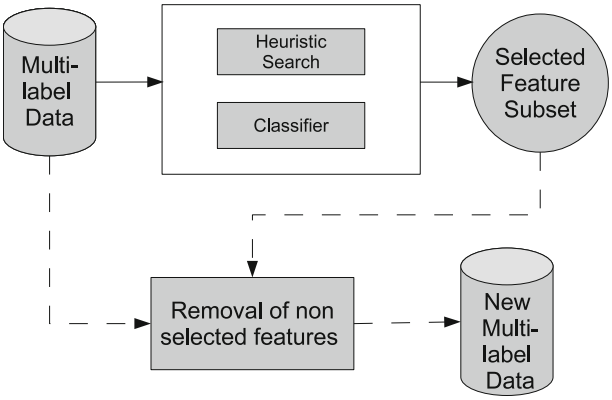

Fig. 5 Direct/Wrapper multi-label feature selection

In other words, given a multi-label learning algorithm, the method searches for the subset of features that optimizes a multi-label metric on the training data set (Tsoumakas et al. 2010).

Figure 5 shows a suitable model for the wrapper paradigm, that is capable of handling multi-label data directly. It works as follows: the data set is submitted to a heuristic search algorithm, and for each selected subset of features, the classification algorithm is used to evaluate it. The best subset of features according to the classification performance is then selected.

Note that in a wrapper approach the adopted classifier can belong to any one of the multi-label categories described in the multi-label classification sections.

Zhang et al. (2009) used a wrapper technique over the data set to identify the best feature set. The wrapper feature selection implements a genetic algorithm as the search component. To evaluate this method, the Multi-label Naive Bayes classifier—proposed in the same work—is employed to select the best features. The classification achieved a better result when coupled with the feature selection. Besides, Also, it achieved a better performance when compared with the following classification algorithms: ADTBoost.MH, Rank-SVM, BR + Naive Bayes and Constrained Non-negative Matrix Factorization (CNMF) (Liu et al. 2006).

Shao et al. (2013) proposed the Hybrid Optimization based Multi-Label (HOML) feature selection. It consists of a hybrid wrapper feature selection technique, combining simulated annealing, genetic algorithm and hill-climbing to optimize the search for an optimal subset of features. HOML was compared with other wrapper algorithms that employ the following heuristic search algorithms: simulated annealing, forward selection, backward selection and genetic algorithms; all of them coupled with the following base classification algorithms: ML-KNN (Zhang and Zhou 2007), BP-MLL (Zhang and Zhou 2006), Rank-SVM (Elisseeff and Weston 2001) and Multi-label Naive Bayes (Zhang et al. 2009). Experimental results on two multi-label data sets favored the HOML technique.

Li et al. (2010b) is another example which also used an ensemble of SVMs after a wrapper feature selection, but for removing irrelevant labels in an image annotation scenario. Then an SVM classifier was trained for each region and the labels combined.

Jungjit and Freitas (2015) used a simple univariate filter method that computes the average correlation between features and labels, and selected the 100–400 top features in the rank. Then a genetic algorithm with a lexicographic component was proposed to select a subset of features, and evaluated against a hill climbing algorithm and a BR technique. The

proposed genetic algorithm achieved a generally better predictive accuracy than the other feature selection techniques, when coupled with the Multi-Label k-NN classifier.

3.2.3 Strategies based on the embedded category

There is also the case of embedded feature selection algorithms, where the classification process itself performs the feature selection as part of learning. Techniques like decision trees (Quinlan 1986, 1993) are examples of classification algorithms that employ an embedded feature selection strategy. In order to build a decision tree model, the learning algorithm selects the features to build the internal nodes, and the leaves of the tree represent the class labels.

Other examples of classifier learning algorithms with embedded feature selection are neural networks, random forests and feature selection using the weight vector of SVM classifiers (Guyon et al. 2008). There is not a general model for the embedded strategy, as the selection of the subset of features is done into the classifier construction (Saeys et al. 2007), so it is highly dependent on the classifier.

Clare and King (2001) proposed a multi-label decision tree as an extension of the C4.5 algorithm, by allowing multiple labels in the leaves of the tree. De Comité et al. (2003) combined a multi-label boosting algorithm with decision trees to produce a novel method—ADTBoost.MH—capable of handling multi-label data.

Li et al. (2010a) proposed the PRECOMN technique, based on a previous technique named Multi-label Embedded Feature Selection (MEFS). It consists of an embedded technique coupled with the ML-KNN algorithm. It combines the sequential backward search algorithm with an evaluation measure, named prediction risk criterion, to evaluate the subset of features. The technique was evaluated on one data set, and results showed that it achieved a better performance than the ML-KNN classification without feature selection, and another classification technique called COMN, that was also proposed in the work.

The Correlated LaRank SVM method proposed by Gu et al. (2011) is a dimensionality reduction technique incorporated into the SVM classifier with a system of ranking labels, an extension of LaRank SVM (Label Ranking SVM). The feature selection is incorporated into the classification algorithm; hence it is categorized as an embedded technique.

3.2.4 Summary of publications on direct multi-label feature selection

Table 2 describes the publications that employ multi-label feature selection techniques that are capable of handling the task without transformation. They correspond to the Direct category in our proposed taxonomy. The "Feature Selection Category" column specifies which of the three categories—*Filter*, *Wrapper* or *Embedded*—the work is focused on. The "Feature Selection Technique" column indicates which multi-label techniques were used—all of them capable of handling the data directly. If the feature selection method is embedded, it is specified in which classifier the feature selection is inserted. "Classification Algorithm" column shows which classification strategy was employed and the "Data Sets" column lists the data sets used and to which domains they belong, in parentheses.

3.3 Experimental evaluation and further research

Among simple transformation methods, entropy-based transformation achieved better results for text data sets and coupled with SVM classifier in Chen et al. (2007). LP transformation for feature selection is popular and achieves competitive results in various domains. However, it

Table 2 Summary of publications on direct multi-label feature selection

References	Feature selection category	Feature selection technique	Classification algorithm	Data sets
Clare and King (2001)	Embedded	C4.5H		Phenotypic data (gene)
De Comité et al. (2003)	Embedded	ADTBoost.MH		Reuters-21450 (textual)
				Newsgroups (textual)
Gu et al. (2011)	Embedded	Correlated LaRank SVM		Scene (image)
				Yeast (gene)
				Yahoo webpages (web)
Li et al. (2010a)	Embedded	PRECOMN (based on ML-KNN)		Yeast (gene)
Kocev et al. (2013)	Embedded	Predictive Clustering Trees Feature Ranking		Bibtex (text)
				Emotions (music)
				Enron (text)
				Medical (text)
Lastra et al. (2011)	Filter	FCBF extension	IBRL-ML	Various domains
			ECC	
Kong and Yu (2012)	Filter	gMLC	BoosTexter SVM	Graph
Zhang et al. (2009)	Wrapper	Genetic Algorithm (wrapper)	Multi-label Naive Bayes	Scene (image)
				Yeast (gene)
				Synthetic data sets
Shao et al. (2013)	Wrapper	Hybrid optimization ML FS (HOML)	ML-KNN	Yeast (gene)
			BP-MLL	TCM CHD (medical)
			Rank-SVM	
			Multi-label Naive Bayes	
Pupo et al. (2013)	Filter	ReliefF-ML	ML Lazy Ranking	Various domains
Spolaôr and Monard (2014)	Filter	ReliefF-ML	BR-KNN	Synthetic data sets

Table 2 continued

References	Feature selection category	Feature selection technique	Classification algorithm	Data sets
Lee and Kim (2013)	Filter	Mutual Information ML	Multi-Label Naive Bayes	Enron (text) Scene (image) Yeast (gene)
Jungjit et al. (2013)	Filter	ML-Correlation-based FS	ML-KNN ML-RBF (neural network)	Bioinformatics gene
Pereira et al. (2015)	Filter	ML Information Gain	Various classifiers	Various domains
Li et al. (2014)	Filter	ML Information Gain	ML-KNN Rand-SVM	Various domains
Reyes et al. (2015)	Filter	ReliefF-ML RReliefF-ML	ML-KNN BRKNN	Various domains
Li et al. (2010b)	Wrapper	Accuracy	Ensembles of SVM	Image annotation
Jungjit and Freitas (2015)	Filter and Wrapper	ML-CFS + Genetic Algorithm ML-CFS + Hill Climbing	ML-KNN	Various domains

is generally outperformed by the BR transformation for most measures (Pereira et al. 2015; Trohidis et al. 2008; Spolaôr et al. 2013; Sechidis et al. 2014), with the exception of the Subset Accuracy measure that is more sensitive on label dependency.

Our categorization showed that there is a lack of work based on transformation methods which employ a wrapper or embedded strategy. Most of them, both single and BR transformations employ a filter strategy. The analysis of how well these other strategies scale and perform on transformed multi-label data sets is therefore an open topic for future research.

Direct methods achieve better results than transformation methods in terms of performance in the case of ReliefF and mutual information methods (Spolaôr and Monard 2014; Reyes et al. 2015 and in terms of computational scalability (Pereira et al. 2015). For a more thorough analysis, filter techniques like in Lastra et al. (2011) should be evaluated with other multi-label classifiers. Wrapper and embedded techniques should be assessed on more domains. Spolaôr et al. (2015) explained the abundance of filters due to the relative lower computational cost in comparison with other alternatives.

Imbalanced labels in multi-label text-categorization domains were analysed by Dendamrongvit et al. (2011). Although it is a valuable analysis in the topic of imbalanced labels, a more extensive analysis could be provided for other domains and feature selection algorithms.

The general scalability of algorithms was assessed by Pereira et al. (2015) and Reyes et al. (2015). In the former, a direct adaptation of information gain was 100 times faster than BR and LP approaches, when using large data sets; and in the latter, BR and LP presented a poor performance on data sets with higher number of labels, favoring also the adaptation methods. But the issue of how well current algorithms scale with respect to the number of labels is still underexplored.

Jungjit and Freitas (2015) adopted a hybrid approach, using an univariate filter method as a first feature selection, and then employing more time-consuming wrapper algorithms, as hill climbing and genetic algorithm. This hybrid approach could be explored by researchers in further experiments.

Other unexplored subjects in the multi-label feature selection domain are: how some methods benefit from the ability of considering label correlations; a theoretical and empirical comparison of representative methods from each category (transformation and adaptation), in order to better visualize the pros and cons of the each one; evaluating and comparing the performance of the filter, wrapper and embedded subcategories on each main category; evaluating and comparing methods which apply binary relevance feature selection externally with the ones which apply it internally; variations in the classification result according to the use of different error measures (Hamming Loss, Subset Accuracy, Ranking Loss, etc.); and answering if a specific category is able to generally achieve better experimental results when combined with a specific classification method or a specific multi-label domain.

4 Conclusions

In this work, we have surveyed previous work on feature selection techniques that deal with multi-label data and we have also proposed an original taxonomy to categorize this kind of techniques. Up to this time, they were scattered in the literature with no common framework to describe them. With this new categorization, we expect to make it more straightforward to describe, classify, evaluate, compare and combine multi-label feature selection algorithms.

Also, we have presented directions for future research, based on the analysis of the current literature. There are various unexplored subjects in the multi-label feature selection domain

that require a closer attention from researchers. A list of these future work topics was produced based on our survey analysis and categorization of the existing methods.

References

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121–168

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chen W, Yan J, Zhang B, Chen Z, Yang Q (2007) Document transformation for multi-label feature selection in text categorization. In: Proceedings of the 7th IEEE international conference on data mining. pp 451–456

Cheng W, Hüllermeier E (2009) Combining instance-based learning and logistic regression for multilabel classification. Mach Learn 76(2–3):211–225

Clare A, King RD (2001) Knowledge discovery in multi-label phenotype data. In: Proceedings of the 5th European conference on principles of data mining and knowledge discovery. pp 42–53

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13:21–27

Dasarathy BV (1991) Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156

de Carvalho ACPLF, Freitas AA (2009) A tutorial on multi-label classification techniques. In: Abraham A, Hassanien A-E, Snášel V (eds) Foundations of Computational Intelligence Volume 5. Springer, Berlin, pp 177–195

De Comité F, Gilleron R, Tommasi M (2003) Learning multi-label alternating decision trees from texts and data. In: Proceedings of the 3rd international conference on machine learning and data mining in pattern recognition. Springer, pp 35–49

Dembczyński K, Waegeman W, Cheng W, Hüllermeier E (2012) On label dependence and loss minimization in multi-label classification. Mach Learn 88(1–2):5–45

Dendamrongvit S, Vateekul P, Kubat M (2011) Irrelevant attributes and imbalanced classes in multi-label text-categorization domains. Intell Data Anal 15(6):843–859

Doquire G, Verleysen M (2011) Feature selection for multi-label classification problems. In: Proceedings of the 11th conference on artificial neural networks on advances in computational intelligence. Springer, pp 9–16

Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York

Elisseeff A, Weston J (2001) A kernel method for multi-labelled classification. Adv Neural Inf Process Syst 14:681–687

Forman G (2004) A pitfall and solution in multi-class feature selection for text classification. In: Proceedings of the 21st international conference on machine learning. ACM, pp 1–38

Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73(2):133–153

Gibaja E, Ventura S (2015) A tutorial on multilabel learning. ACM Comput Surv (CSUR) 47(3):52

Godbole S, Sarawagi S (2004) Discriminative methods for multi-labeled classification. In: Proceedings of the 8th Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 22–30

Gu Q, Li Z, Han J (2011) Correlated multi-label feature selection. In: Proceedings of the 20th ACM international conference on information and knowledge management. pp 1087–1096

Guyon I, Elisseeff A (2006) An introduction to feature extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA (eds) Feature extraction, foundations and applications. Springer, Berlin, pp 1–24

Guyon I, Gunn S, Nikravesh M, Zadeh LA (2008) Feature extraction: foundations and applications, vol 207. Springer, Berlin

Hüllermeier E, Fürnkranz J, Cheng W, Brinker K (2008) Label ranking by learning pairwise preferences. Artif Intell 172(16–17):1897–1916

Jungjit S, Freitas A (2015) A lexicographic multi-objective genetic algorithm for multi-label correlation based feature selection. In: Proceedings of the companion publication of the 2015 annual conference on genetic and evolutionary computation. ACM, pp 989–996

Jungjit S, Michaelis M, Freitas AA, Cinatl J (2013) Two extensions to multi-label correlation-based feature selection: a case study in bioinformatics. In: Proceedings of the IEEE international conference on systems, man, and cybernetics. IEEE, pp 1519–1524

Khalid S, Khalil T, Nasreen S (2014) A survey of feature selection and feature extraction techniques in machine learning. In: Proceedings of the science and information conference (SAI). IEEE, pp 372–378

Kocev D, Slavkov I, Dzeroski S (2013) 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

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324

Kong X, Yu PS (2012) gmlc: a multi-label feature selection framework for graph classification. Knowl Inf Syst 31(2):281–305

Lastra G, Luaces O, Quevedo JR, Bahamonde A (2011) Graphical feature selection for multilabel classification tasks. In: Proceedings of the 10th international conference on advances in intelligent data analysis. pp 246–257

Lee J, Kim DW (2013) Feature selection for multi-label classification using multivariate mutual information. Pattern Recognit Lett 34(3):349–357

Li GZ, You M, Ge L, Yang JY, Yang MQ (2010) Feature selection for semi-supervised multi-label learning with application to gene function analysis. In: Proceedings of the 1st ACM international conference on bioinformatics and computational biology. pp 354–357

Li L, Liu H, Ma Z, Mo Y, Duan Z, Zhou J, Zhao J (2014) Multi-label feature selection via information gain. In: Advanced data mining and applications, lecture notes in computer science. Springer International Publishing, pp 345–355

Li R, Zhang Y, Lu Z, Lu J, Tian Y (2010) Technique of image retrieval based on multi-label image annotation. In: Proceedings of the 2nd international conference on multimedia and information technology (MMIT), vol 2. IEEE, pp 10–13

Liu B, Hsu W, Ma Y (1998) Integrating classification and association rule mining. In: Proceedings of the 4th international conference on knowledge discovery and data mining. pp 80–86

Liu H, Motoda H (eds) (2008) Less is more. In: Computational methods of feature selection. Chapman & Hall/CRC, Boca Raton, pp 3–17

Liu Y, Jin R, Yang L (2006) Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proceedings of the 21st national conference on artificial intelligence. pp 421–426

Mencía EL, Furnkranz J (2008) Pairwise learning of multilabel classifications with perceptrons. In: Proceeding of the 2008 IEEE international joint conference on neural networks. pp 2899–2906

Molina LC, Belanche L, Nebot A (2002) Feature selection algorithms: a survey and experimental evaluation. In: Proceedings of the 2002 IEEE international conference on data mining. pp 306–313

Olsson J, Oard DW (2006) Combining feature selectors for text classification. In: Proceedings of the 15th ACM international conference on information and knowledge management. ACM, pp 798–799

Pereira RB, Plastino A, Zadrozny B, Merschmann LH (2015) Information gain feature selection for multi-label classification. J Inf Data Manag 6(1):48

Pupo OGR, Morell C, Soto SV (2013) ReliefF-ML: an extension of ReliefF algorithm to multi-label learning. In: Ruiz-Shulcloper J, Sanniti di Baja G (eds) Progress in pattern recognition, image analysis, computer vision, and applications. Springer, Berlin, pp 528–535

Quinlan JR (1986) Induction of decision trees. Mach Lear 1:81–106

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, Massachusetts

Read J (2008) A pruned problem transformation method for multi-label classification. In: Proceedings of the New Zealand computer science research student conference. pp 143–150

Read J (2010) Scalable multilabel classification. Ph.D. dissertation, Hamilton

Read J, Pfahringer B, Holmes G, Frank E (2009) Classifier chains for multi-label classification. In: Proceedings of the 20th European conference on machine learning and knowledge discovery in databases. pp 254–269

Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. Mach Learn 85(3):333–359

Reyes O, Morell C, Ventura S (2015) Scalable extensions of the relieff algorithm for weighting and selecting features on the multi-label learning context. Neurocomputing 161:168–182

Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge

Rogati M, Yang Y (2002) High-performing feature selection for text classification. In: Proceedings of the 11th international conference on information and knowledge management. ACM, pp 659–661

Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Sechidis K, Nikolaou N, Brown G (2014) Information theoretic feature selection in multi-label data through composite likelihood. In: Fränti P, Brown G, Loog M, Escolano F, Pelillo M (eds) Structural, syntactic, and statistical pattern recognition. Springer, Berlin, pp 143–152

Shao H, Li G, Liu G, Wang Y (2013) Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine. Sci China Inf Sci 56(5):1–13

Sorower MS (2010) A literature survey on algorithms for multi-label learning. Technical Report, Oregon State University, Corvallis

Spolaôr N, Monard MC (2014) Evaluating relieff-based multi-label feature selection algorithm. In: Proceedings of the 14th edition of the Ibero-American conference on artificial intelligence. Springer, pp 194–205
Spolaôr N, Tsoumakas G (2013) Evaluating feature selection methods for multi-label text classification. In: Proceedings of the first workshop on bio-medical semantic indexing and question answering
Spolaôr N, Cherman EA, Monard MC, Lee HD (2013) A comparison of multi-label feature selection methods using the problem transformation approach. Electron Notes Theor Comput Sci 292:135–151
Spolaôr N, Monard MC, Tsoumakas G, Lee HD (2015) A systematic review of multi-label feature selection and a new method based on label construction. Neurocomput Prog Intell Syst Des 180:3–15
Trohidis K, Tsoumakas G, Kalliris G, Vlahavas IP (2008) Multi-label classification of music into emotions. In: Bello JP, Chew E, Turnbull D (eds) Proceedings of the 9th international conference on music information retrieval. pp 325–330
Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Warehous Min 3(3):1–13
Tsoumakas G, Vlahavas I (2007) Random k-labelsets: an ensemble method for multilabel classification. In: Proceedings of the 18th European conference on machine learning. pp 406–417
Tsoumakas G, Dimou A, Spyromitros E, Mezaris V, Kompatsiaris I, Vlahavas I (2009) Correlation based pruning of stacked binary relevance models for multi-label learning. In: Proceedings of the 1st international workshop on learning from multi-label data. pp 101–116
Tsoumakas G, Katakis I, Vlahavas I (2010) Mining multi-label data. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, Berlin, pp 667–685
Wandekokem ED, Varejão FM, Rauber TW (2010) An overproduce-and-choose strategy to create classifier ensembles with tuned svm parameters applied to real-world fault diagnosis. In: Progress in pattern recognition, image analysis, computer vision, and applications, Lecture notes in computer science, vol 6419. Springer, pp 500–508
Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proceedings of the 14th international conference on machine learning. pp 412–420
Yu K, Yu S, Tresp V (2005) Multi-label informed latent semantic indexing. In: Proceedings of the 28th ACM SIGIR conference on research and development in information retrieval. pp 258–265
Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
Zhang ML, Zhou ZH (2006) Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans Knowl Data Eng 18:1338–1351
Zhang ML, Zhou ZH (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recognit 40(7):2038–2048
Zhang ML, Zhou ZH (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26:1819–1837
Zhang ML, Peña JM, Robles V (2009) Feature selection for multi-label naive bayes classification. Inf Sci 179(19):3218–3229
Zhang Y, Zhou ZH (2010) Multilabel dimensionality reduction via dependence maximization. ACM Trans Knowl Discov Data 4(3):1411–1421
Zheng Z, Wu X, Srihari R (2004) Feature selection for text categorization on imbalanced data. ACM SIGKDD Explor Newslett 6(1):80–89