UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Pre-processing techniques in multi-label classification

Daan Van Rossen

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

dr. Gonzalo NAPOLES RUIZ

2019
2020

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Pre-processing techniques in multi-label classification

Daan Van Rossen

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

dr. Gonzalo NAPOLES RUIZ

Pre-processing techniques in multi-label classification*

Daan Van Rossen - Gonzalo Nápoles

Faculty of Business Economics, Hasselt University, Belgium

Abstract. Classification is an important task within machine learning, a generalization of traditional classification is multi-label classification. In multi-label classification, an instance is associated with a subset of labels, this allows modelling of a broader range of real-world problems. However, multi-label datasets introduce two major challenges, the curse of dimensionality and label imbalances. These challenges directly reduce the scalability and predictive performance of multi-label classifiers. In order to overcome these challenges, preprocessing techniques can be applied to reduce the dimensionality and balance of the dataset. A large variety of preprocessing techniques are presented within the literature. Different methods correspond to different conceptual ideas, assumptions or algorithmic properties. Therefore, in this thesis, a taxonomy of preprocessing techniques for multi-label classification is constructed by reviewing existing literature. Each categorization is described and their respective techniques are reviewed.

Keywords: multi-label classification · preprocessing · dimensionality reduction · imbalanced learning

1 Introduction

Technological advancements have increased our ability to process, storage and transmit data; consequently, the availability of data was never higher [94]. Nowadays, digital applications have the increasing need to discover knowledge from this collected data [50]. To satisfy this need, the field of machine learning has transitioned from an "academic discipline" to that of an "applied science" [63]. One of the most important tasks within machine learning is that of automatically classifying data [117]. Classification is the task of training a computational model using a set of labeled instances in order to correctly classify, never seen before, unlabeled instances [2]. However, traditional classification is not applicable to all real-world problems, a broad range of digital applications such as text categorization [101, 173], image annotation [64, 11], scene classification [12], video segmentation [42], protein function classification [36, 31] and music classification [90, 144], all have data instances that are naturally associated with more

* This master thesis was written during the COVID-19 crisis in 2020. This global health crisis might have had an impact on the (writing) process, the research activities and the research results that are at the basis of this thesis.

than one class label. In order to classify these real-world problems, multi-label classification was introduced. Multi-label classifiers have the ability to predict a subset of labels and thus are capable of modelling real-world problems with data instances that are naturally associated with more than one class label. The effectiveness and efficiency of multi-label classification is, like any classification algorithm, closely related to the inherent quality of the training data [125]. Unfortunately multi-label datasets whether they are real-world or synthetic, suffer from high-dimensionality in both feature and label spaces [18]. Additionally they have an intrinsically imbalanced nature [21]. These properties increase the difficulty of correctly classifying multi-label datasets. Therefore, the use of preprocessing techniques play a key role in the multi-label classification process. The goal behind preprocessing is to increase the data quality, specifically for multi-label classification. This consists of applying data reduction techniques to reduce the data dimensionality and balance the datasets.

In this paper, we hope to provide a starting point and reference for researchers interested in preprocessing techniques for multi-label classification. Therefore we present a taxonomy of preprocessing techniques taking into account the nature of multi-label datasets. More specifically we focus on data reduction techniques that alleviate the negative effects of high-dimensionality and label imbalance learning. The remainder of this paper is structured as follows. In Section 2 we briefly describe preprocessing and multi-label classification and their role within the Knowledge Discovery in Databases process. Followed by a theoretical description of the challenges multi-label learning in Section 3. Section 4 provides a taxonomy of methods for overcoming the challenges described in Section 3. Section 5 discusses the categorization of dimensionality reduction techniques and reviews existing techniques. In Section 6 we elaborate further on the categorization of resampling techniques and review existing techniques within the literature. Finally, in Section 7, we make our concluding remarks.

2 Preliminary

Preprocessing and multi-label classification are techniques situated within the Knowledge Discovery in Databases process. The KDD process is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [37]. The KDD process can be divided into stages, there are several methods to make this division, each with advantages and disadvantages [50]. In this paper we adopt the categorization of these stages according to [40] into six stages:

1. Problem Specification,
2. Problem Understanding,
3. Data Preprocessing,
4. Data Mining,
5. Evaluation,
6. Result Exploitation

The KDD process can involve significant iteration and can contain loops between two stages [50]. In the following subsections we briefly describe the data preprocessing stage and multi-label classification problem in the data mining stage.

2.1 Data Preprocessing

Real-world data tends to be dirty, incomplete, and inconsistent [56, 50], often it contains a lot of irrelevant, redundant, and noisy information [89]. It can be stated that real-world data is often of low quality [56, 89, 120] and as is well-known, the effectiveness and efficiency of the learning algorithm in the data mining stage is dependent on the quality of data. Low quality data leads to low-quality performance in the data mining stage and thus to a lower quality of knowledge in general [120]. Hence, data preprocessing is a fundamental stage in the KDD process, wherein preprocessing techniques are applied to the data to improve data quality. As a result, the accuracy and efficiency of the learning, conducted in the subsequent data mining stage, are improved [50, 57]. In [40], data preprocessing methods are divided into two categories: data preparation [118, 147] and data reduction [118, 147, 99]. The former, data preparation, compromises the set of techniques that initialize the raw data properly to serve as input for a certain learning algorithm in the data mining stage [40]. Without initializing the data, the learning algorithm might not work. Or might work incorrectly such that the results will not make sense and thus isn't considered as accurate knowledge. Data preparation include data transformation, integration, cleaning, and normalization methods. The latter, data reduction, comprises the set of techniques that aim to simplify and clean the raw data and thus obtain a reduced representation of the original data [40]. Although the learning algorithm in the data mining stage will work without applying data reduction, major issues arise when data reduction is skipped in the preprocessing stage. In short, applying data reduction methods enables learning algorithms to learn faster and perform more accurate. In this paper we focus on data reduction preprocessing methods since these enhance the performance of learning algorithms needed to solve multi-label classification tasks.

2.2 Multi-Label Classification

In general, classification within the data mining stage can be stated as the process of predicting one or more class labels for an unseen instance described by a vector of feature values by analyzing the training set [2]. In traditional classification these training sets are composed of a set of input features and a unique value in the output attribute, the class or label [9, 175]. However, in Multi-Label Classification (MLC) [57, 146], the processed or predicted instance is not associated with a single class label, instead each instance can belong to a subset of class labels at the same time. Multi-label classification is thus, a generalization of traditional classification, this allows it to be applicable to a wide range of real-life domains such as: bioinformatics [36, 31, 88, 174], emotion analysis [7],

text analysis [101, 173, 73, 108, 169, 182], image analysis [64, 11], video analysis [12, 42], music analysis [90, 144], among others. Most MLC methods follow one of two main approaches to deal with the multi-label classification problem [147]:

- The problem transformation approach [134, 147]: which transforms the multi-label problem into one or several single-label classification or regression problems. This transformation allows the use of existing single-label classification algorithms.
- The algorithm adaptation approach [134, 147]: which adapt existing algorithms to make them able to handle multi-label data directly.

A review of MLC learning algorithms is provided in [72, 175]. Although multi-label classification allows the modelling of a wide variety of real-world problems, it also increases the difficulty of correctly classifying the patterns in the data. The additional challenges that multi-label modelling bring are discussed in Section 3.

3 MLC Challenges

In general, all challenges faced by multi-label classification are challenges that pre-existed within traditional classification problems. However, they are further enhanced due to the fact that in multi-label classification, each instance can be associated with more than one class. In general two primary challenges are associated with multi-label classification: curse of dimensionality and label imbalance.

The curse of dimensionality [6, 67, 170, 171] refers to problems derived from the presence of many dimensions. As the number of dimensions increases, so does the volume of the solution space or search space. As a consequence, data points in this volume tend to be sparse as the dimensions grow, and distances between them tend to be less significant. Thus, to draw meaningful conclusions generally a larger collection of data points is needed. This implies more time to build the classifiers and usually a degraded predictive performance by most algorithms. Reference [57] states that multi-label classification has to account for three possible high dimensionality spaces: instance space, feature space and label space. While traditional classification can suffer from the curse of dimensionality in the instance and feature space, the curse of dimensionality affects multi-label classification even more due to the addition a third high dimensional label space. Multi-label datasets often have hundreds or even thousands of labels resulting in a high-dimensional label space that increases the solution space even further and enhances the curse of dimensionality.

The label imbalance problem emerges when there are many instances belonging to some classes (majority classes), but only a few representing others (minority classes). Learning from an imbalanced dataset [53] is a well-known and researched problem in a traditional non-multi-label classification context. In general, classifiers tend to underperform when learning from an imbalanced dataset, this can be contributed to their design. Aimed to reduce the global

error rate, classifiers favor the majority class, labeling new instances with this class at the expense of the minority class [63]. Once again, the multi-label classification context further enhances this problem, this can be explained due to the fact that multi-label datasets have an intrinsically imbalanced nature, as was experimentally validated by Charte et al. [21]. Meaning that although the total number of distinct labels is usually large (high-dimensional label space), the amount of instances associated to a label isn't (sparseness). Additionally the labels are correlated within the multi-label dataset.

4 Categorization for MLC preprocessing

Preprocessing techniques can overcome the aforementioned challenges. However, there is a large variety of preprocessing techniques, which correspond to different conceptual ideas, assumptions of their underlying model, or algorithmic properties. Therefore, in this section, we provide a taxonomy of multi-label classification preprocessing techniques. A taxonomy helps to understand the variety of techniques, their interrelation and grouping. In order to construct this taxonomy we reviewed 147 articles, resulting in 91 reviewed techniques. Figure 1 presents our proposed taxonomy, it consists of two data reduction preprocessing categories: dimensionality reduction and resampling.

Dimensionality reduction techniques can be further divided into two categories, single space reduction and dual space reduction [112]. Single space reduction is split into two categories: feature space reduction and label space reduction. The former processes the initial high-dimensional feature space into a reduced feature space. It can be achieved using feature selection [48, 74, 105] or feature extraction [47, 135] techniques. The latter processes the initial high-dimensional label space into a reduced label space. It can be achieved using label selection or label embedding techniques [162]. Dual space reduction consider both the curse of dimensionality in the feature space and the sparseness problem in the label space jointly by reducing dimensionality in both feature and label space. In Section 5, we review dimensionality reduction approaches which aim at reducing the number of features, labels, or both in order to overcome the curse of dimensionality. Reducing dimensionality results in a reduction of the associated computational burden and thus improving scaling properties of the classifiers along with their predictive performances [132].

Resampling techniques can be further divided into two categories, oversampling algorithms and undersampling algorithms [23]. Algorithms in the former group produces new samples with the minority class, while the latter removes instances linked to the majority class. In Section 6, we review resampling techniques for multi-label imbalanced learning which aim to overcome the class imbalance problem. Applying resampling techniques to an imbalanced dataset, results in an improved classifier performance [22].

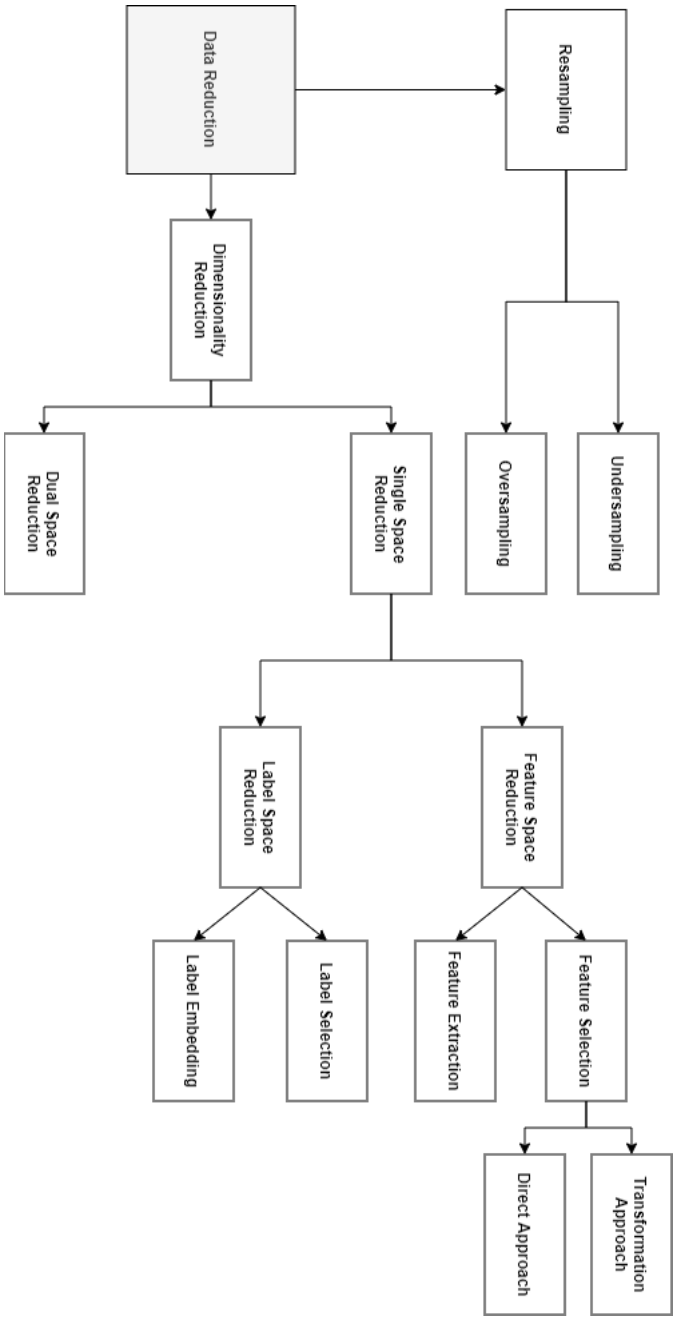

Fig. 1. Taxonomy of preprocessing techniques for MLC problems.

5 Dimensionality Reduction

Dimensionality reduction techniques are imperative for dealing with the curse of dimensionality and large data in general. Dimensionality reduction techniques aim to reduce the number of features, labels, or both by obtaining a reduced or compact representation of the original data and preserving the essence of the original data [50, 100]. The reduction of features and/or labels has the goal to alleviate the computation burden of the classifier, resulting in an improvement of scaling properties and predictive performances [132].

As stated in Section 4, our proposed taxonomy further divides dimensionality reduction techniques into two categories, single space reduction and dual space reduction [112]. In the following subsections we elaborate on each category and the respective reduced dimensionality spaces. Furthermore, we review preprocessing techniques for each category and further categorize where possible.

5.1 Single Space Reduction: Feature Space Reduction

Within traditional classification algorithms, the curse of dimensionality, is traditionally associated with the input feature space. As a consequence, literature covers many proposed techniques to deal with this problem [47]. Many of the traditional feature space reduction techniques can be adapted to multi-label data [57]. In general and as visualized in Figure 2, feature space reduction techniques

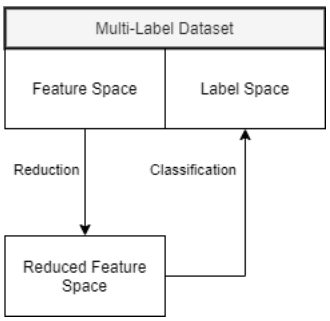

Fig. 2. Process of feature space reduction.

process the initial large feature space into a reduced feature space. This reduced feature space can be later used in the data mining stage to make a classification [132]. Note that although these methods can successfully reduce the dimensionality of the feature space, the resulting multi-label dataset will still suffer from high dimensionality in the label space. Hence, the classification task might only benefit from a sub-optimal improvement in terms of scaling properties and predictive performances.

Feature space reduction can either be achieved using feature selection [48, 74, 105] or feature extraction [47, 135] techniques. Both approaches can effectively

reduce data dimensionality by removing irrelevant and/or redundant features, speeding up learning algorithms without sacrificing performance [137]. The difference between the two categories lies in the reduction and representation of the original dataset. Feature selection approach evaluates the relevance of attributes already present in the original data, the features themselves remain unchanged. Feature Extraction approach [47] generates new features from the original features by transforming data into a low-dimensional space while preserving its structure.

Feature Selection

Feature selection is a preprocessing approach that aims to find a small subset of features that describes the dataset as well as, or even better than, the original set of features does. It achieves this by selecting features, which provides the most useful information to the learning algorithm. The result of feature selection is a subset of features from the original set of features [95]. Feature selection can be categorized according to different perspectives [72]: the label perspective, search strategy perspective, interaction with learning algorithm perspective, and data format perspective. For our taxonomy we closely followed the categorization proposed in [116]. Therefore first a division is made based on the data format perspective into: a direct approach and transformation approach. Thereafter we consider the interaction with the learning algorithm to make a selection of techniques applicable to multi-label classification. The interaction perspective divides feature selection techniques in three main approaches: the filter, wrapper and embedded approach [72]. Only the first two approaches can be performed in the data preprocessing step; therefore, we consider the embedded approach out of the scope of this paper [68].

- Filter feature selection methods use general properties of the dataset to remove irrelevant and/or redundant features from it, regardless of the learning algorithm [57]. Since they rely exclusively on the original dataset, the selection process remains unbiased. The advantage of being classifier independent is that the selection process only needs to be performed once for all learning algorithms [122]. However, being classifier independent is conjointly a major disadvantage of the filter method. That is, the selection process ignores the effects of the selected features on the performance of a learning algorithm and thus the selected features may be sub-optimal for a given classification algorithm [141, 138].
- Wrapper feature selection methods; on the other hand, are designed to optimize the subset of features internally using a given learning algorithm in the evaluation process [57, 136]. This makes the approach classifier dependent and thus biases the learning algorithm to influence the feature selection method. Consequently, the wrapper approach has the advantage of selecting features that have the highest impact on the performance of the learning algorithm. The downside of being classifier dependent is the associated high computational cost, since for each feature set the learning algorithm has to be called [141].

Usually filter methods are more efficient and less complex to implement in comparison to wrapper methods, where as wrapper methods provide better performance [68, 136].

The transformation approach is a two-step process that consists of transforming the multi-label dataset into single-label datasets before applying a traditional feature selection technique. Note that after the transformation, we can only use traditional feature selection methods of the filter category. The process is described in Figure 3, initially the multi-label data is converted into a non-multi-label dataset. This is achieved by using a data transformation approach, note that this data transformation has some major drawbacks. Firstly, the label correlations are often ignored and secondly, the computational costs increase exponentially with size of the label space [3]. After the data transformation, the data consist out of instances with each a set of independent features and one dependent label. Then, a traditional (non-multi-label) feature selection method can be used to evaluate each feature or subset of features. This is done using a collection of metrics, the result is a list of selected features that can be used to remove the non-selected features from the the original multi-label dataset.

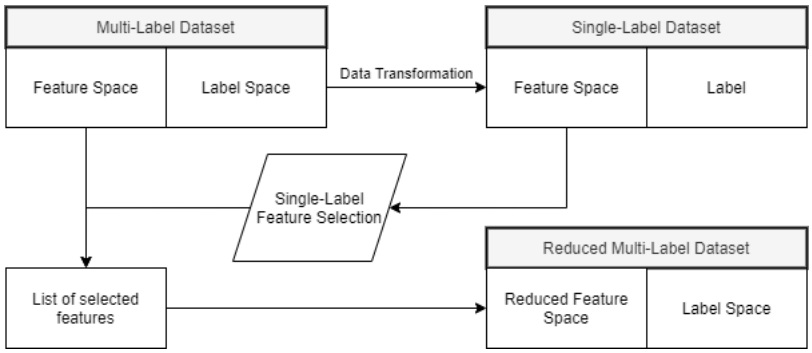

Fig. 3. Multi-label feature selection using the transformation approach.

The use of traditional (single-label) feature selection methods is a major advantage of the transformation approach since these methods have been thoroughly studied and perfected over the last decades. In the literature, we find various transformation strategies combined with traditional feature selection approaches. Below we briefly discuss the most important transformation techniques, afterwards we review the existing literature on transformation based multi-label feature selection.

– The Binary Relevance or BR transformation [146] transforms the multi-label dataset into a set of binary datasets or sub-problems. Afterwards the discriminative power of features to each individual label is evaluated. Note that the label is evaluated in isolation from the rest of the labels. Lastly

an aggregation strategy is used to obtain a global ranking for each feature. Features that are above a certain threshold are chosen.
- Label Power set or LP transformation [146] transform the multi-label dataset into a multi-class dataset or sub-problem. It achieves this by mapping each distinct label combination as class identifier in the corresponding multi-class problem. Subsequently feature selection is performed using traditional (non-multi-label) algorithms.
- Pruned Problem Transformation or PPT [121] is an adaptation of LP. After the multi-label dataset is converted into a multi-class dataset, pruning is used to discard instances with classes that occur too infrequently by considering a predefined threshold. This ensures that all classes are represented by at least a number of instances equal to a preset threshold.

In feature selection, feature-label dependency should be maximized so that the remaining features give the best description of the label space. At the same time, feature-feature dependency should be minimized so that redundant features are removed [72, 91]. It is generally agreed upon that BR-based transformation techniques outperform LP-based transformation techniques [115, 128, 145]. A disadvantage of BR-based transformations is that they do not consider potential label-label dependencies, which is a factor assumed to have an important role in improving performance of multi-label selection. However this assumption is controversial within the literature, several papers confirm the positive impact of taking label dependency in consideration [85, 86, 119]. Whereas several papers claim the opposite [43, 72]. On the other hand LP-based transformations do take into account label correlations, but struggle to deal with large label spaces. As the label space grows, so does the high computational cost. On top of that the LP based transformations are prone to cause imbalanced multi-class data when faced with large label-spaces [3]. To overcome these disadvantages of LP-based transformations, PPT-based transformations were proposed. Note that PPT-based transformation are in itself irreversible and may result in loss of class information [91]. Below we review the existing literature concerning multi-label feature selection using a transformation based approach. An overview of the reviewed methods can be found in Table 1.

The first example of using a data tranformation approach for multi-label feature selection was given in [181]. The author combines a data transformation along with twelve single-label feature selection measures to evaluate the usefulness of features. The feature selection measures used in the study are:

- Laplacian Score [54] is an univariate feature weighting algorithm for unsupervised learning that preserves the data manifold structure. In Laplacian Score, features are evaluated independently. Therefore, the optimization problem defined above can be solved by greedily picking the top features, which have the minimal Laplacian Score values. Since features are evaluated individually, Laplacian Score cannot handle feature redundancy.
- SPEC [180] is an univariate feature weighting algorithm for supervised and unsupervised learning that extends the Laplacian Score. In SPEC, the feature relevance is measured by three different criteria to assign similar values

to instances that are near each other. Note that SPEC also evaluates features individually; therefore, it cannot handle feature redundancy

- Fisher score [35] is an univariate feature weighting algorithm for supervised learning that is a special case of Laplacian Score. Fisher Score assign similar values to the samples from the same class and different values to samples from different classes. By greedily selecting the features with the largest Fisher Scores, the top k features can be obtained. Note that Fisher Score also evaluates features individually, therefore it cannot handle feature redundancy.
- ReliefF [123] is an univariate feature weighting algorithm for supervised learning. It uses a distance measurement as evaluation criterion, to select features that contribute to the seperation of samples.
- t-score [28] is an univariate feature weighting algorithm for supervised learning. It is used for binary problems that have unequal sample size and unequal variance.
- F-score [30] is an univariate feature weighting algorithm for supervised learning. It is used to test if a feature is able to separate samples well from different classes by considering between class variance and within class variance.
- Chi-square Score [96] is an univariate feature weighting algorithm for supervised learning. Chi-square is used in feature selection as a test of independence to assess whether the class label is independent of a particular feature.
- Gini Index [129] is an univariate feature weighting algorithm for supervised learning. It is used to quantify a feature's ability to distinguish between classes. The smaller the Gini Index, the more relevant the feature is. Since the Gini Index evaluates features individually, it cannot handle feature redundancy.
- Information Gain [164] is an univariate feature weighting algorithm for supervised learning. It is used to measure the dependence between the feature and the class label. The higher the Information Gain, the more relevant the feature is. Since features are evaluated individually, Information Gain cannot handle feature redundancy.
- Fast Correlation Based Filter (FCBF) [167] is a multivariate feature set algorithm for supervised learning. FCBF measures the feature-class and feature-feature correlation to find a subset of features that are highly correlated to the class, but not highly correlated to the other features. FCBF evaluates features jointly; therefore, FCBF is able to handle feature redundancy.
- Correlation-based Feature Selection (CFS) [49] is a multivariate feature set algorithm for supervised learning. CFS uses a correlation based heuristic to evaluate the worth of features. Meaning it calculates feature-class and feature-feature correlations using symmetrical uncertainty and then selects a subset of features using the Best First search with a stopping criterion. The advantage of CFS is it selects the maximum relevant feature and avoids the reintroduction of redundancy. But the drawback is that CFS cannot handle problems where the class is numeric.

– Minimum-Redundancy-Maximum-Relevance (MRMR) [114] is a multivariate feature set algorithm for supervised learning. MRMR is based on mutual information and measures correlation among features and correlation between features and group using this measure. It selects features that are mutually far away from each other, while they still have "high" correlation to the classification variable.

In [144] Trohidis et al. automatically detect emotions in music using a multi-label classification algorithm. The authors apply multi-label feature selection by first transforming the multi-label problem with the LP method into a multi-class problem. Subsequently the Chi-square Score [96] is used to rank the features.

Doquire et al. propose the PPT-MI method in [33] to improve the classification performance of image annotation and gene function classification. PPT-MI is a multi-label feature selection method using the Pruned Problem Transformation (PPT) [121] to transform the multi-label dataset into a multi-class dataset, followed by a sequential forward selection with the Mutual information (MI) [151] as search criterion. The paper extends the preliminary results that were presented in [32] and proposes a way to automatically select the pruning parameter for PPT. The PPT-MI algorithm is empirically validated and show better classification performance than PPT+CHI.

In [137] the authors propose the use of two data transformation approaches, BR and LP, to transform the multi-label data into single-label data. Subsequently, ReliefF [80, 165] and Information Gain [164] are used as feature evaluation measures for each label. The resulting methods are named BR-RF, LP-RF, BR-IG and LP-IG. These methods are then experimentally valuated. The authors conclude that the main advantage of ReliefF over other strictly univariate measures is that it takes into account the effect of interacting features. As a result, the ReliefF based algorithms, BR-RF and LP-RF are observed to outperform the LP-IG and BR-IG algorithms.

In [136] an algorithm, called Label Construction for Feature Selection (LCFS) is proposed. LCFS uses a two-step process firstly it constructs new labels based on pairwise relations between the original labels to augment the label set of the original dataset with second-order information. Then the augmented dataset is submitted to a feature selection algorithm based on BR and using the Information Gain [164] measure to select relevant features. The experimental evaluation of LCFS shows that LCFS outperforms classifiers without feature selection, as well as random feature selection. Additionally, LCFS is competitive with IG-BR at the cost of slightly increasing the computational burden due to the application of a binary operator.

Gharroudi et al. [43] propose two wrapper multi-label feature selection methods, called Binary Relevance Random Forest (BRRF) and Random Forest Label power-Set (RFLP). Both methods firstly transform the multi-label data using BR and LP, respectively. Then, a feature selection method based on the random forest paradigm [14] is applied. The experimental results of the study show that BRRF performs significantly better than PMU, RFLP and PPT-MI. Although

the proposed methods consider label dependence, the authors conclude that this consideration is not significantly effective in multi-label feature selection.

In the study of Reyes et al. [122] the PPT-ReliefF method is proposed. Similar to PPT-MI [32], PPT-ReliefF firstly uses the Pruned Problem Transformation (PPT) to transform the multi-label problem into a multi-class problem. Subsequently, the ReliefF algorithm is used to measure and rank the features according to their usefulness in distinguishing instances. In their experiments, the authors show that PPT-ReliefF outperforms the BR-RF [137] and LP-RF [137] algorithms. As well as improving the performance compared to classifiers without feature selection.

In [72] twelve transformation based multi-label feature selection methods (BR-RF, LP-RF, BR-IG, LP-IG, BR-FSCORE, LP-FSCORE, BR-FCBF, LP-FCBF, BR-CHI, LP-CHI, BR-CFS and LP-CFS) were experimentally validated. The authors conclude that it was not possible to find significant difference among the methods.

The direct approach consists of feature selection techniques that deal directly with multi-label data, meaning there is no need to transform the dataset. Direct multi-label feature selection methods are often an extension or adaptation of existing single-label feature selection methods. The process of direct multi-label feature selection can be divided into two steps. First, a metric e.g. correlation coefficient [48, 66], mutual information [38, 133] is used to measure the importance of candidate features and construct an objective function. Secondly, a search strategy is constructed to solve the given optimization function [91]. In the following paragraphs, we review the literature concerning direct feature selection.

As far as we know, Zhang et al. [172] were the first to propose an adaptation of the traditional naïve Bayes classifiers to multi-label datasets. The proposal, named Multi-Label classification with Naive Bayes (MLNB), is a wrapper method that incorporated a two-stage feature selection strategy. In the first stage Principeal Component Analysis (PCA) [113] is employed for feature construction, subsequently in the second stage the Genetic Algorithm (GA) [149] is used as a heuristic approach for feature selection. The fitness function of GA includes both hamming loss and ranking loss. MLNB suffers from three disadvantages, firstly, due to PCA being an unsupervised method, a degradation in classifier performance can be expected when used with multi-label datasets [104]. Secondly, the GA-based algorithms consume much time to reach the optima and may results in premature convergence [72]. Lastly the MLNB algorithm ignores the label interdependence.

In [84], the well-known technique FCBF [167] is extended to the multi-label setting. The authors propose an algorithm, called Multi-Label feature Ranker (MLfR) that employs a graphical model and applies a symmetrical uncertainty measure to represent the correlation and interdependence between all pairs of features and labels. Subsequently, the spanning tree of the complete undirected graph is computed. Then, a selection is performed by choosing the vertices corresponding to features whose distance from the whole set of labels is lower or equal to a given threshold. The proposed technique is evaluated and proved to achieve

significantly better results than applying an MLC learning algorithm directly without reducing the feature space. Moreover, the graph built by MLfR provides a valuable representation of the correlation and interdependence between labels and features.

In Kong and Yu's study [79] a new technique designed for graph classification was proposed, called gMLC. The technique is based on an efficient search for optimal subgraph features for graph objects with multiple labels, using a proposed gHSIC criterion that takes into account the dependence of the subgraph features with multiple labels of the graphs. Then, a branch-and-bound algorithm is proposed to efficiently search for a compact set of subgraph feature that is useful for the classification of graphs with multiple labels. The paper evaluates the proposed technique on Graph data sets and compares it to a BR-IG technique [137]. The results of this comparison favores the proposed gMLC technique.

Various adaptation of the ReliefF algorithm [80, 165] have been proposed for multi-label feature selection. This can be contributed due to the fact that Relief is one of the few algorithms capable of detecting feature dependencies [148]. Kong et al. study [78] was the first to presents MReliefF, a Relief algorithm for multi-label feature selection. Firstly, the MReliefF technique transforms the multi-label problem into a set of pairwise multi-label 2-class problems. Then, the cases where the "miss" and "hit" sets contain both classes are removed. A drawback of the MReliefF algorithms is that it only partially takes into account label correlations. In [109], ReliefF-ML is proposed. The ReliefF-ML extends ReliefF to assign weights to features according to their discriminative power, the feature weight reflects the ability of the feature to distinguish class labels. Finally, Reyes et al. [122] propose two extensions of the ReliefF algorithm, named ReliefF-ML and RReliefF-ML. The algorithms are extended to work in the multi-label learning context directly. The ReliefF-ML algorithm uses the same approach as in [109]. While the RReliefF-ML is based on RReliefF, a variant of the classical ReliefF algorithm specifically designed for regression problems. The authors compare the proposed methods transformation based counterparts: BR-RF [137], LP-RF [137] and PPT-ReliefF [122]. The results of the comparison favores the proposed methods in terms of classifier performance. In addition the proposed extensions by [122] are scalable on simple and complex multi-label datasets with different properties.

In [130] a hybrid optimization method called Hybrid Optimization based Multi-Label (HOML), is presented. 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. By combining these different optimization techniques to search for an optimal subset of features, HOML has the ability to avoid being trapped in a sub-optimal solution. HOML was compared against six other wrapper feature selection and extraction algorithms, the result shows HOML outperforming the other algorithms.

In the work of Jungjit et al. [70] ML-CFS, an adaptation of the Correlation-based Feature Selection (CFS) [49] technique is proposed. The algorithm is capable of evaluating a subset of features, instead of individual features. It uses

a simple hill-climbing algorithm to perform a heuristic search in the candidate feature space. The objective is to find feature subsets that maximize the features predictive accuracy and minimizes their dependency. The same author proposes two extensions of ML-CFS, called GA-ML-CFS [69] and LexGA-ML-CFS [68]. The former is a correlation-based feature selection method based on GA that has a single objective function. The latter is an improvement of GA-ML-CFS, which uses a fitness function with two objectives maximization of the classification accuracy and minimization of the number of selected features. Note that these two objectives are conflicting and may sacrifice classification accuracy. The experimental results of the studies [70, 69, 68] show that ML-CFS, GA-ML-CFS and LexGA-ML-CFS improved predictive accuracy, by comparison against a baseline. However, overall there was no statistically significant difference between the results of the extensions GA-ML-CFS and LexGA-ML-CFS and the ML-CFS method.

In [85] the PMU methods is proposed, the Mutual Information measure is adapted to select features with the most discriminating power. A score function is devised by decomposing Mutual Information between the feature and the label sets into a series of multivariate mutual information. Subsequently, an incremental selection strategy that maximizes the multivariate Mutual Information between selected features and the labels is performed by using a forward search strategy. PMU is the first multi-label filter feature selection method that considers label interactions in measuring the dependency of given features [72].

In Garroudi et al. [43] Random Forest Predictive Clustering Tree (RFPCT) is proposed. RFPCT is an extension of Random Forest (RF) [14] that uses a randomized variant of the non Pruned Predictive Clustering Tree (PCT) [10], as a base classifier. The diversity among trees is promoted using two strategies, bootstrap sampling of training data and random selection of feature subsets. RFPCT is capable of predicting multiple target attributes at once by measuring feature relevance on each PCT tree, and then averaging them over all the trees in the forest. RFCPT is able to exploit the underlying label dependencies; however, the experimental results indicate that ignoring correlation among labels within the feature selection process doesn't affect the quality of the multi-label classification, as was previously confirmed by [137].

The Max-Dependency and Min-Redundancy (MDMR) is proposed in [91]. It is a filter algorithm that uses a criterion for incremental multi-label feature selection based on the mRMR method. Features are incrementally selected until a predefined number of features is reached. The selection is based on a score calculated using mutual information, the dependency between features and labels and redundancy between non-selected features and the subset of the selected features are taken into account. MDMR does not consider the inter dependencies between labels. The study verifies the efficacy of MDMR by comparing its performance with MLNB, PMU and MDDM algorithms. MDMR achieves similar classification performance compared to PMU, and is superior to MLNB and MDDM.

Lee and Kim [87] propose a filter method, called multi-label feature selection based on scalable relevance evaluation (SCLS). The algorithm introduces a novel way of measuring feature redundancy, by using an incremental selection strategy that sequentially includes the top-ranked feature in the already selected feature subset. The ranking is performed using mutual information to determine the relevancy between features and label. A major drawback of SCLS is that it ignores the label correlations while evaluating the feature relevance. The newly proposed method is experimentally validated and the authors conclude that SCLS outperforms other mutual information-based multi-label feature selection methods by providing significantly better discriminating power.

A new filter method based on mutual information is proposed by Zhang et al. in [176]. The method, named Feature Selection based on Label Redundancy (LRFS), divides labels into independent and dependent groups. Then, a newly proposed feature relevance term, called Label Reduncancy (LR), is used to measure mutual information between a candidate feature and each already-selected feature. Then, based on this LR measure a ranking of features is made. A major advantage of LRFS is that it takes into account the effect of label redundancy on the evaluation of the feature relevance. Previous multi-label feature selection methods based on information theory ignored the effect of label redundancy. Various experimental results demonstrate that LRFS can effectively select the compact feature subset from the original data set and LRFS obtains better classification result than D2F, PMU, PPT-MI, PPT-CHI, SCLS, IGMF and MIFS alogrithms.

In [71] two novel multi-label filter feature selection methods based on the Pareto optimal set are presented, ParetoFS and ParetoCluster. Both methods are not inspired by any similar method in single-label feature selection and are exclusively designed for multi-label data. Both methods select features that are members of Pareto optimal set and investigate each label individually. The difference between the two methods is that the first one is a subset feature selection method, i.e. the number of features to be selected is defined by the algorithm. The second proposed algorithm is a variation of the first method, which ranks the features and selects a specified number of features that is defined by the user. The study performs two series of experiments to evaluate the proposed methods, the experimental results show that both of methods have better classification performance compared to PMU, ELA-CHI, PPT-CHI, PPT-MI, PPT-ReliefF. Additionally, the proposed methods are amongst the least time consuming algorithms.

In [51] Hashemi et al. design a fast multi-label filter feature selection algorithm using the PageRank algorithm, called MGFS. Similar to [71] and unlike previous methods, MGFS is no adaptation of a single-label feature selection method, instead it is specifically designed for multi-label data. MGFS method uses a multi-label graph-based theory, and the Google PageRank algorithm is employed to select the best feature subset. The MGFS algorithm is experimentally validated and is proven to have a low computational complexity. Furthermore,

it has better classification accuracy and less error compared to PMU, LRFS, MDFS, PPT-MI and ParetoCluster algorithms.

Mishra et al. [104] propose a wrapper method, called Feature Selection for Multi-Label classification using Clustering in feature-space (FS-MLC). FS-MLC first creates clusters for features and considers these created clusters as independent instances. Then, for each cluster a feature is selected as a representative for all the features in that cluster. Using sample-based precision and recall measures the representative features are ranked. FS-MLC has some positive characteristics, firstly, it is a wrapper method that does not require to create a large number of feature subsets linearly proportional to the number of labels in the dataset. Secondly, FS-MLC is a parameter tuning free approach meaning that the number of features are selected automatically. FS-MLC is experimentally compared to MDMR, SCLS and DFSC algorithms and has proven to be superior.

It is difficult to determine the best feature selection method because only partial comparisons are reported in the literature and to the best of our knowledge there exist no experimental studies that compare all the methods presented above and summarized in Table 2. In choosing a competitive feature selection method, a choice between the transformation and direct approach needs to be made. Based on the experimental results reported in the literature, methods based on the direct approach seem to perform better for MLC algorithms in terms of accuracy [122, 72, 176, 71, 51]. However, for certain datasets there are only minor performance differences with the PPT-MI method. Moreover, the PPT-MI method has a significantly lower computational burden than direct algorithms [71, 51]. This can partly be explained due to the fact that all transformation based feature selection methods are filter algorithms. Since the structure of filter algorithms are very simple, it provides a more efficient calculation of features relevance. On the other hand, the structure of the wrapper approach is more complicated and thus tends to be better at maximizing predictive accuracy. However, this improvement in predictive accuracy comes at a higher computational cost. Therefore, the literature favors filter methods for large scale datasets. The most competitive filter feature selection method seems to be the MGFS algorithm [51]. When faced with a smaller scale dataset, a wrapper approach can be more effective in terms of classifier accuracy, the FS-MLC algorithm seems to be the most competitive one. Note that the choice of feature selection method is highly dependent on the nature of the datasets, given this fact and the lack of a comprehensive experimental study, the generalization of the respective competitive methods should be considered cautiously.

Feature Extraction

Feature extraction [47, 135] is a task different from feature selection [94]. In the sense that it is a preprocessing approach that projects the original feature space into another low-dimensional space made up of new features that are a linear combination of the original features. Similar to feature selection, feature extraction removes redundant and irrelevant features while preserving maximum information of the original data set.

Feature extraction does not support a data transformation approach because there is no way to aggregate artificial features of the different single-label subproblems to get the final multi-label solution. Therefore, a large body of literature has extended single-label feature extraction algorithms to directly handle multi-label data sets. In our taxonomy we make a division based on the exploitation of label information into unsupervised and supervised feature extraction methods. Hereby we follow the general categorization proposed in [132].

Unsupervised feature extraction methods aim to reduce dimensionality of the feature space without taking into account the label information of each instance. They rely on the analysis of input feature space; therefore, traditional (single-label) methods can be applied right out of the box to multi-label data. Although unsupervised feature extraction can be directly applied, they fail to make use of the label information. Clearly these techniques are not ideally suited for reducing the feature space of a multi-label dataset. As a result, they are not the focus of this paper. For the curious reader, a brief overview of unsupervised feature extraction techniques is provided in Table 3.

Supervised feature extraction methods exploit the label information by using it to guide the reduction of the feature space. This strengthens the link between the extracted reduced feature space and the label space. Since the label space is taken into account, supervised feature extraction methods have to be adapted in some way before being applicable to multi-label data [57]. Below we review feature extraction methods adapted for use in the multi-label setting, an overview of the reviewed methods can be found in Table 4.

Partial Least Squares (PLS) is a family of methods that can be used for dimensionality reduction [124]. PLS extracts latent features vectors from the dataset by maximizing the covariance and correlation between the reduced feature space and the label space. PLS has several drawbacks, firstly, PLS cannot capture high order correlation information among different labels. Secondly, it is unable to find a space with a larger dimensionality than the number of labels; thus, its generalization performance on new dimensions of outputs is restricted [158].

Yu et al. [166] propose the Multi-label informed Latent Semantic Indexing (MLSI). MLSI is an extension of the traditional LSI algorithm [29]. Although LSI is an unsupervised algorithm, MLSI is not, since it makes use of the labeling information to build the lower dimensional feature projection. It achieves this by computing a mapping function for both the feature space and the corresponding labels. The final solution is obtained by solving an eigen problem where eigenvectors with largest eigenvalues are directly integrated into the mapping function. In doing so, MLSI creates a new feature space wherein the feature variances and binary label variances are maximized. The new feature space preserves the information of the original feature space and captures the dependencies among the labels [91]. In the study MLSI is empirically validated and has shown to work well on a number of tasks. Similar to PLS, a drawback of MLSI is that it fails to capture high order correlation information among different labels. In addition,

MLSI has a high computational burden due to the time consuming large-scale inverse matrix computation [161]. A nonlinear kernel variant is proposed in [178], but is not experimentally validated for multi-label applications.

In [65] Tang et al. propose the Multi-Label Least Square (MLLS) algorithm, a general framework for extraction of shared structures in multi-label datasets. The MLLS methods assumes that a subspace is shared among multiple labels, it uses this assumption to extract the common features among multiple labels. MLLS computes a linear transformation (least squares loss) to discover this shared subspace. Then, an optimal solution can be computed via solving a generalized eigenvalue problem. An advantage of the MLLS formulation is that the performance is not sensitive to the high dimensionality of the shared subspace, it is, however, sensitive to low dimensionality of the shared subspace. MLLS is experimentally validated in the study, MLLS achieves a performance improvement in comparison to the SVD algorithm. However, the authors note that MLLS is dominated by SVD in terms of computational burden.

Zhang and Zhou [178] propose a dimensionality reduction algorithm, called Multi-label Dimensionality reduction via Dependence Maximization (MDDM). The algorithm projects the feature space into a lower dimensional space such that the dependencies between the features and the corresponding labels is maximized. The dependency is measured using the Hilbert-Schmidt Independence Criterion (HSIC) [46] as metric. In the process, the dimensionality reduction problem is formulated by an eigen-decomposition problem, whose solutions constitutes the reduced feature space. The study delivers two algorithms, MDDMp that uses the HSIC with orthonormal projection directions as metric and MDDMf that uses HSIC orthonormal projected features respectively. Both MDDM method are experimentally compared against other feature extraction methods, the results show that MDDM is slightly superior to PCA and LPP and significantly superior to MLSI.

In [152], Wang et al. propose a generalization for the LDA [34] method, called Multi-label Linear Discriminant Analysis (MLDA). The MLDA method uses a weighted form to estimate the label correlation instances and labels, to obtain an optimal low-dimensional sub-space by maximizing the between-class scatter measure and minimizing the within-label scatter at the same time. A drawback of the MLDA method is that the low-dimensional sub-space can not exceed the number of classes minus 1 [152]. The paper validates the proposed MLDA method and finds that MLDA significantly reduces the feature dimensionality as well as improve classifier performance. To overcome MLDA's limit on the dimensionality of the reduced space, Oikonomou and Tefas propose in [107] the Direct Multi-label Linear Discriminant Analysis (DMLDA) method.

The MDDM method [178] has several drawbacks, MDDM uses dense matrices eigen-decomposition that are computationally expensive for use with high-dimensional data. In addition, MDDM cannot be guaranteed to capture the correlation between multiple labels. Therefore, Shu et al. [131] proposes a new hybrid method of MLLS and MDDMf [132], called Shared Subspace multi-label Dimensionality reduction via Dependence Maximization (SSMDDM). The au-

thors reformulates the MDDM method as a least squares problem and thus avoid the use of direct eigen-decomposition on the large scale matrix. This allows the SSMDDM methods to be computationally faster and effectively uncover multiple label interactions. Various experimental results demonstrate the effectiveness of SSMDDM in both predictive and computational performance, SSMDDM significantly outperforms the original MDDM. In [161], Xu et al. proposes the Maximizing feature Variance and feature label Dependence (MVMD) method, which is a hybrid method of PCA and MDDMp. Similar to [131], the authors reformulate the MDDM method as a least squares problem. Subsequently, the objective function of PCA is linearly combined with the objective function of MDDMp, to both maximize feature variance and maximize feature-label dependence simultaneously. The proposed MVMD method is empirically validated on eight datasets against seven existing feature extraction methods: PCA, MDDMp, MDDMf, CCA, MLSI, MLDA and DMLDA. It is proven that MVMD improves the multi-label performance and outperforms the seven existing feature extraction methods. In [132], MDDM, MVDM and SSMDDM are directly compared against each other, it is concluded that the SSMDMM algorithm outperforms the MVDM and MDDM algorithms in terms of F1-score and Hamming loss. However, the authors state that the results are dependent on the choice of datasets.

It is a challenging and active research task to search for a novel well-performed criterion for feature extraction [162]. The results reported in the literature greatly depend on both the used quality measures and used datasets [132]. Three algorithms MVMD [161], SSMDDM [131] and MDDM [178] based on the Hilbert-Schmidt Independence Criterion (HSIC) [46] appear to be the most competitive supervised feature extraction methods. Of which SSMDDM seems to be the best choice due to being computationally faster and having the ability to effectively uncover multiple label interactions. Furthermore, a choice between feature selection and feature extraction needs to be made. From a dimensionality reduction point of view, both approaches are effective in alleviating the curse of dimensionality in the feature space and improving scalability and prediction performance of the classifier. Within the literature, few studies experimentally compare both approaches [130, 91], and further research is needed in order to draw a conclusion. However, feature extraction creates a set of new features; therefore, losing the physical meanings of these features and limiting the further analysis. This poses an obstacle to scientific understanding of scientific problems [130]. Feature selection maintains physical meanings of the original features and gives models better readability and interpretability. The choice between feature selection and feature extraction is dependent on the need to preserve the structure (feature extraction) and the need to know the physical interpretation of the features (feature selection).

5.2 Single Space Reduction: Label Space Reduction

Within Feature Space Reduction, Feature Selection and Extraction Methods were already well researched methods for traditional classification. However, the

need to reduce label space dimension is specific to multi-label classification and did not exist in binary and multi-class datasets [57]. Several MLC algorithms (such as BR methods) can't handle datasets with a high-dimensional label space [97]. Hsu et al. [61] pose the sparseness problem in the label space, stating that although the label space may be very high dimensional, the actual labels are often sparse. In this subsection we aim to categorize Label Space Reduction techniques that focus on overcoming this sparseness problem. Similar to Feature Space Reduction, the goal of label Space Reduction is to reduce the training cost of classification algorithms by removing irrelevant, redundant or noisy information of labels while maintaining classification performance [27]. The reduction

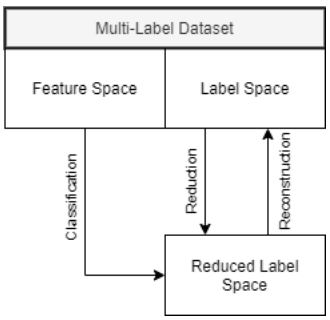

Fig. 4. Process of label space reduction.

process is visualized in Figure 4. In the preprocessing stage, label space reduction methods first reduces the original high-dimensional label space to a lower dimension label space. In the data mining stage, the multi-label classifier algorithm is trained using the feature space and reduced label space. In the Label Space Reduction process a additional postprocessing stage is added to reconstruct the reduced label space to the original high-dimensionality label space [112, 162]. The ability to decompose and reconstruct the label space is a key aspect of label reduction methods. Due to the fact that all labels need to be provided to the multi-label classifier, labels cannot simply be removed from an multi-label datasets [57]. Dually to feature space reduction, existing label space reduction techniques are divided into two categories: label selection and label embedding [162]. The former group selects an informative label subset set directly from the original label space. A major advantage of this method is the reservation of original meaning of the labels, this allows for an improved interpretability of the predicted results [142]. The latter group uses a linear or non-linear transformation to transform the original label space into a low-dimensional embedded space [154, 83]. A representation of the hidden structure of the label space is kept in the embedded space. Label embedding methods successfully save computational power and storage while improving classifier performance [93]; however, a drawback is the loss of original label meaning [142]. There is relatively few liter-

ature on Label Space Reduction methods and the proposed techniques all follow a unique approach [57]. In the following paragraph we discuss these proposed techniques, an overview of the techniques and their classification can be found in Table 5.

To the best of our knowledge, Hsu et al. [61] proposed the first label (embedding) space reduction method, called Compressed Sensing (CS). The proposed method makes the premise that the original label space is sparse. Firstly, a linear compression function is used to encode the label space into a small number of linear random projections. Then, in the decoding step, an optimization problem needs to be solved to reconstruct the labels with respect to the sparsity error. Note that the encoding component of CS is linear, but the decoding component is not. As a result, the decoding step can be time-consuming. A variant of the CS method is proposed in [183]. The method, called Compressed Labeling (CL), firstly extracts the label dependence information from the original multi-label dataset and stores them in distilled labelsets. In the postprocessing stage these distilled labelsets are used to reconstruct the original label space. Subsequently, the label space is transformed into a lower dimensional space using the signs of the linear Gaussian random projections to preserve the binary nature of the original one. In comparison with CS, the (linear) decoding stage is faster and more efficient. Although the encoding and decoding stages of the CL algorithm are fast, the initial storing of label dependence information in distilled labelsets is not [8]. Both CS and CL significantly reduce the label space using label compression; thus, substantially alleviating the problem of label dependence and high label dimensionality. Although this greatly benefits the learning task, we have to note that both methods transform the multi-label classification problem into another type of problem. The CS method can only be wrapped around regression classifiers, whereas the CL method can only be combined with binary classifiers. This transformation of the MLC problem limits the application range of the CS and CL method [18].

Landmark Selection Method for Multiple Output Prediction (MOPLMS) [4] is a label selection label reduction method based on a groupsparse problem. The algorithm assumes that all the output labels can be recovered from a small subset. The study experimentally validates that MOPLMS is able to considerably reduce the sample complexity when the output dimensionality is high. However, MOPLMS suffers from a major drawback, it relies on an expensive and complex optimization problem to select labels. As a consequence, MOPLMS is often unfeasible to run on larger datasets. Furthermore, a small drawback of MOPLMS is that the size of the label subset cannot be explicitly controlled [8].

Tai et al. [140] found the decoding step of CS to be inefficient and proposed Principal Label Space Transformation (PLST). PLST can be viewed as the counterpart of PCA in the label space and is feature-unaware, meaning it only considers the label information during reduction. Firstly, the original label space is transformed into a smaller linear label space using singular value decomposition (SVD) [45, 150]. Subsequently, using the properties of an orthogonal matrix derived from SVD, this matrix can be used to transform the reduced label

space back to the original space. By a number of experiments PLST was shown to successfully improve MLC performance in both accuracy and computational time, especially for datasets with a large number of labels. Furthermore, PLST proved to faster and more accurate than the CS method [8, 110, 140]. A variant of PLST, called Conditional Principal Label Space Transformation (CPLST), is proposed in [27, 140]. The CPLST algorithm combines the concepts of PLST and CCA [60] and attempts to improve PLST through the addition of feature information. The goal is to simultaneously minimize both the encoding error and training error in the reduced dimensional space. The authors experimentally compare PLST and CPLST and conclude that CPLST is at least as good as, and usually better than, PLST. However, Appice et al. [1] empirically showe that CPLST is computationally more expensive than PLST, while CPLST has no significant performance difference with PLST [27, 8, 1]. This suggests that taking into account feature information only provides little information for label transformation or selection [8].

Bi and Kwok [8] propose an efficient CSSP [13] variant called Multi-Label Column Subset Selection Problem (ML-CSSP). ML-CSSP firstly uses a partial singular value decomposition (SVD) on the label space to derive label weights. Then the most relevant labels are selected using a randomized sampling algorithm. Consequently, the ML-CSSP alogrithm reduces the label space dimension by selecting labels where the sampling probability or weight of each label reflects its performance over all the labels. Unlike regular CSSP, the ML-CSSP algorithm uses an adaptive technique to determine the number of sampling trials. The authors experimentally validate ML-CSSP and prove that their proposed algorithm outperforms MOPLMS, PLST, CPLST and CL algorithms in terms of classification accuracy. Additionally they state that, compared to label embedding methods, ML-CSSP has a lower training error, as the selected labels are easier to learn.

In reference [18], Charte et al. propose Label Inference for Multilabel Classification (LI-MLC). LI-MLC is a pre-/postprocessing label selection algorithm designed to reduce label dimensionality. In the preprocessing stage, an association rule mining algorithm is used to obtain label dependency information. The result is a multi-label dataset with a reduced label space; thus, any existing multi-label classifier can be used in the data mining stage to generate predictions. In the postprocessing stage, the same set of rules allows to infer the missing labels from the presence of others in the predicted label space. The use of any existing multi-label classifier is a major advantage compared to the above proposed algorithms CS and CL. It allows LI-MLC to preserve the original multi-label nature of the problem and to be applicable to broader range of applications [18]. The authors experimentally validate the LI-MLC approach and conclude that the LI-MLC algorithm allows classifiers to be trained in less time, resulting in simpler models and improved classification results in many cases. Note that to the best of our knowledge, no experimental comparison between LI-MLC and other label reduction methods can be found in the literature.

The above described label embedding methods (CS, CL, PLST and CPLST) and ML-CSSP perform label space reduction in a function-based manner and require an encoding function (e.g linear encoding) to be used in the preprocessing stage [92]. However, since an optimal transformation to a lower-dimensional embedded space can be complicated and even indescribable, assuming an function-based or explicit encoding function may not well model it [93]. Lin et al. [92] state that an end-to-end or implicit approach, namely directly learning from the label space without using an encoding functions can be feasible and even preferable. Therefore, Wicker et al. [157] proposed the first implicit label embedding method, called Multi-Label Classification Using Boolean Matrix Decomposition (MLC-BMaD). The proposed method uses end-to-end label space encoding via Boolean matrix factorization [102] to factorize the label matrix into a factor matrix of latent labels and a factor matrix of the interdependencies among these labels. A classifier can then learn from the latent labels instead of the actual labels. The predicted label space is then reconstructed by Boolean matrix multiplication using the second factor matrix. The MLC-BMaD was experimentally validated by Wicker et al. and showed better classifier performance, compared to the PLST [111, 132]. However a disadvantage, the MLC-BMaD method is a feature-unaware method, meaning the correlations between the latent space and the feature space are not considered. According to Zhou et al. [183] this results in a less predictable latent space and thus a possibly degraded classifier performance.

Therefore, Lin et al. [93] propose Feature-aware Implicit label space Encoding (FaIE), via combining PLST and CCA linearly with orthonormal projection labels. The proposed method directly learns a feature aware code matrix and a linear decoding matrix via jointly maximizing recoverability of the original label space and the predictability of the latent space. A major advantage of FaIE is that it is an end-to-end label reduction approach that takes into account feature correlations. In the study, FaIE is experimentally validated, the study concludes that although FaIE has a higher computational cost, it achieves superior classification performance compared to the CS, PLST, CPLST, ML-CSSP and MLC-BMaD algorithms. In [92] FaIE is extended into several variants that are more efficient and effective.

Mineiro and Karampatziakis [103] developed a function-based, feature aware label space (embedding) reduction method, named Response EMBedding via RANDomized Techniques (Rembrandt). The proposed method uses REML [160] to decompose the label matrix into a low-rank structure and sparse component. The low-rank structure represents label correlations, while the sparse component represents the outliers. The advantage of the Rembrandt algorithms is that it works for both multi-class and multi-label problems, additionally Rembrandt exponentially speeds up the running time compared to traditional algorithms making it applicable to large datasets. However, to the best of our knowledge, no experimental comparison between Rembrandt and other label reduction methods can be found in the literature.

In [168] the authors propose a new function-base, feature aware label space reduction method, called Dependence Maximization based Label space Reduction (DMLR). The method uses a linear integration of PLST with MDDM to maximize the dependence between feature vectors and code vectors via the Hilbert–Schmidt Independence Criterion [46] while minimizing the encoding loss of labels. DMLR differs from the above mentioned label space reduction methods in the assumption that the objective function should consist of two components: encoding loss and dependence loss. Whereas the latter only considers either encoding or dependence loss in their objective function. The study experimentally validates that DMLR achieves superior accuracy results compared to PLST and CPLST algorithms at the cost consuming slightly more training time.

Huang et al. [62] propose Cost-sensitive Label Embedding via Multidimensional Scaling (CLEMS) algorithm. The proposed end-to-end, feature unaware label space reduction method, firstly embeds the label and cost information into an embedded space using a classic multidimensional scaling approach for manifold learning. From the embedded space, CLEMS can make cost-sensitive predictions efficiently and effectively by decoding to the nearest neighbor within a proper candidate set. The authors experimentally validate the CLEMS method, the results show CLEMS outperforming the non-cost-sensitive algorithms PLST, CPLST, and FaIE. In [139] it is stated that CLEMS can struggle on larger data sets due to its considerable complexity.

In [155] the authors propose a novel feature aware label embedding method for multi-label classification via maximizing global recoverability and dependency, and minimizing Local Variances (ML-mLV). The function-based algorithm implements, besides two existing global label embedding criteria (global recoverability and dependency-based one), a new local label recoverability criterion based on local consistency [153]. The algorithm has the objective to maximize the global label recoverability and dependency, and minimize the local label recoverability. The study experimentally compares the proposed ML-mLV with PLST, CPLST, MLC-BMaD and FaIE algorithms. The ML-mLV method performs the best, compared with the other approaches according to two performance metrics for large-scale label sets.

In general we can conclude that all label space reduction methods can help to reduce the total training costs while maintaining classifier performance [168]. It is difficult to determine the best label space reduction methods because only partial comparisons are generally reported in the literature and to the best of our knowledge there exist no experimental studies which compare all the methods presented above and summarized in Table 5. The choice of label space reduction method is highly dependant on the nature and size of the dataset. Based on the partial comparisons in the literature, we believe the ML-mLV [155] is the most competitive label embedding method, capable of dealing with large-scale datasets. The cost sensitive label embedding algorithm CLEMS, outperforms traditional non-cost sensitive algorithm when dealing with non-large-scale datasets. However, a pairwise comparison between ML-mLV and CLEMS is needed before a conclusion can be drawn. In some cases, a reservation of the original label

meaning is needed to keep the interpretability of the predicted results. In this setting a label subset selection method should be used, literature shows that ML-CSSP is the most competitive label subset selection method. We note that although the LI-MLC algorithm shows great promise, further research is needed to asses the limitations and performance of the method against existing state-of-the-art methods.

5.3 Dual Space Reduction

The reviewed single space reduction methods in the previous subsections have proven to be effective in improving classifier performance. Nevertheless, they are faced with a trade-off, the use of feature reduction methods to overcome the high dimensionality problem in the feature space fails to consider the sparseness problem [61] in the label space. Likewise, label space reduction methods overcome the sparseness problem in the label space, but fail to consider the high dimensionality of the feature space. Consequently, single space reduction methods might not be best-suited for a multi-label dataset with both a high dimensional feature and label space. Therefore, dual space reduction methods [112] consider both the curse of dimensionality in the feature space and the sparseness problem in the label space. By reducing the feature and label space jointly, better performance is achieved compared to single space reduction methods [110–112]. Dual space

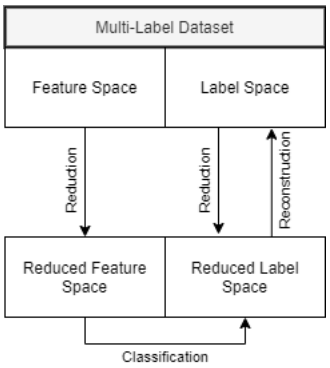

Fig. 5. Process of dual space reduction.

reduction methods perform a combination of feature and label space reduction methods; therefore, the process, visualized in Figure 5, is similar to the feature and label reduction process. In the preprocessing stage, both the feature and label space are reduced to a lower dimensional space. Using the reduced feature space and reduced label space, the multi-label classifier algorithm is trained. Finally, the reduced label space is reconstructed in an additional postprocessing stage. Below we review the dual space reduction techniques presented in the existing literature. Pacharawongsakda and Theeramunkong [110] were the first to

attempt to reduce both the feature and label space simultaneously. Their study introduces two dual space reduction approaches, the first method, named Dependent Dual Space Reduction (DDSR), calculates the cross-covariance matrix of feature and label spaces to represent dependency between features and labels. Then, both matrices are projected into a single reduced space using SVD [45, 150]. The reduced space is subsequently used in the data mining stage to perform predictions. Note that although the DDSR method considers correlation information between features and labels, it neglects the dependency amongst features and the dependency amongst labels. The second method, named Independent Dual Space Reduction (IDSR), calculates the cross-variance matrix of features and the cross-variance matrix of labels separately. Then, both matrices are projected to separate lower-dimensional spaces using independent SVD's [45, 150]. By representing the features and labels dependency separately, we can assume there is no linear dependency between the two matrices. As a result, we can take into account the dependency among features and the dependency among labels. In addition, the study examines classification performance when both spaces are simultaneously reduced. The proposed methods, DDSR and IDSR, transform the feature and label spaces into lower-dimensional spaces with less computational time than other well-known methods. Moreover, they successfully improve classification performance in terms of accuracy compared to other methods. However, the proposed methods suffer from a number of limitations. Firstly the SVD method is used to project spaces into a lower dimension, SVD is known to be computationally expensive when the dimensions are large. Secondly a threshold selection process is required, an improper selected threshold may introduce a bias. Thirdly the use of projections may trigger information loss of some degree.

The same authors propose an alternative dual space reduction framework in [111] namely, Two-Stage Dual Space Reduction (2SDSR) framework. The proposed framework overcomes a limitation of DDSR, which is that the number of the reduced dimensions cannot exceed the dimensionality of the label matrix. Therefore, DDSR cannot find the optimal reduced feature dimensions [110, 112]. The framework consists of two stages. In the first stage the dimensional label space is projected to a lower-dimensional label space using an existing feature-aware label space reduction method. As a second stage the high-dimensional feature space is transformed to a reduced feature space using an existing dependent feature space reduction method. In the study four dimensionality reduction techniques MDDM, CCA, SVD and BMD are applied to the 2SDSR framework and experimentally validated. The 2SDSR variant where MDDM and BMD were exploited achieved the best performance.

The studies [110–112] experimentally show that simultaneously reducing both the feature and label space achieves better classification performance compared to traditional single space reduction methods. Additionally, there is no dual space reduction method that performs well for all evaluation metrics [112].

6 Resampling

To solve the problem of imbalance learning for multi-label classification, a data resampling approach can be applied. A resampling approach is a preprocessing approach that aims to rebalance the class distribution by deletion [81] or recreation [25] of instances. Being a preprocessing approach, it has the advantages of being independent of the classification process and thus having a broader application range. The use of resampling techniques has been extensively researched in a non-multi-label context [52] and has shown their general effectiveness [41]. However, traditional (non-multi-label) resampling methods are designed to work with one output class/label only and assume that there are only one minority and one majority class/label in the dataset. While in a multi-label dataset each instance is associated with more than one label and additionally one instance may have both a minority and majority label concurrently [16]. Thus, a new multi-label resampling approach that takes into account these characteristics of MLC problems has to be designed.

Within multi-label resampling approaches a division can be made based on the rebalance method. In the first approach, referred to as undersampling, some of the samples are removed from the majority class. In the second approach, referred to as oversampling, samples are added to the minority class. Undersampling algorithms usually perform worse than oversampling ones [41], since they cause a loss of information by removing instances. When undersampling is applied to multi-label datasets, an even greater information loss occurs since the removed instance is representing several labels instead of one. [19] In the following paragraph we review under- and oversampling methods that have been proposed for multi-label datasets. An overview of the reviewed methods can be found in Table 6.

A first approach for multi-label resampling was presented by Charte et al. [15]. The study presents a group of measures aimed to evaluate the imbalance level in multi-label datasets. These measures, IRlbl and MeanIR can be used to assess the imbalance level, and indicate if the multi-label dataset needs to be resampled. Additionally, the paper presents two random resampling techniques, called Label Powerset Random Oversampling (LP-ROS) and Label Powerset Random Undersampling (LP-RUS). Both LP-RUS and LP-ROS algorithms are based on the LP transformation and considers each label combination as class identifiers. In [21] the same author proposes two additional algorithms: Multi-Label Random Oversampling (ML-ROS) and Multi-Label Random Undersampling (ML-RUS). Both algorithms, individually evaluate the imbalance level of each label using the IRlbl measure[15]. The paper experimentally validates the algorithms and concludes that ML-ROS obtains the best overall results and overall significantly improved classification results when applied to imbalanced multi-label datasets.

In the work of Charte et al. [20] a method of Multi-Label edited Nearest Neighbor (MLeNN) is proposed. The method is a heuristic undersampling algorithm that is build upon the ENN rule [159]. It compares all instances with classes containing a majority labels against the class of its nearest neighbours.

Instances whose class differs from the class of two or more nearest neighbours are removed. The authors introduce two new key ideas to adapt ENN for multi-label classification. Note that unlike the methods proposed in [21], the samples to be removed are heuristically selected instead of random. The study experimentally validates the proposed algorithm and concludes that MLeNN is able to improve classification results when applied to imbalanced multi-label datasets. Moreover, MLeNN achieves better performance than LP-RUS [20].

In [44] the authors analyze different strategies aimed to apply the original Synthetic Minority Over-sampling Technique (SMOTE) algorithm [26] to multi-label datasets. The result is an heuristic oversampling algorithm, SmoteUG. Note that SmoteUG only takes into account one minority label and thus ignores the intrinsic nature of multi-label datasets [22]. In [22] an extension of the SMOTE algorithm is presented for multi-label datasets, named Multi-Label Synthetic Minority Over-sampling Technique (MLSMOTE). This extension overcomes the limitation of SmoteUG and is able to take into account multiple minority labels. The authors experimentally show that MLSMOTE achieves a statistically significant improvement against the results obtained without preprocessing. MLSMOTE is able to reduce the imbalance level in multi-label datasets and outperforms the oversampling algorithms: ML-ROS, LP-ROS and SmoteUG in terms of prediction results.

A major disadvantage of the above mentioned algorithms is that they use a cloning approach, meaning the algorithm always works over a full labelset. As a result, applying these methods to multi-label datasets with a high SCUMBLE [16] or a high level of concurrence between imbalanced labels can be counterproductive. To overcome this limitation, Charte et al. propose in [17] a new method, called REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels (REMEDIAL). The REMEDIAL method is designed to deal with multi-label datasets having a high SCUMBLE level. The algorithm works as an oversampling method and as an editing technique. It decomposes instances containing both a minority and majority label into two easier instances, one of which merely contains majority labels and another only with minority labels. The study experimentally shows that REMEDIAL is recommended for resampling of multi-label datasets with a high SCUMBLE level and when BR or LP based classifiers are going to be used. Under these conditions REMEDIAL is able to improve classification results. The same authors propose an improvement for the REMEDIAL technique, named REMEDIAL Hybridization, with Resampling (REMEDIALHwR) in [24].

Finally, in [117], an extension of the standard Tomek Link resampling algorithm [143], called Multi-Label Tomek Link (MLTL), is proposed. The proposed algorithm can either be used as heuristic undersampling algorithm or as a postprocessing cleaning step. MLTL uses the multi-label imbalances measure (MeanIR) to find the samples from the majority classes in combination with Hamming Distance to determine how different two labelsets are.

The success and effectiveness of resampling techniques is highly influenced by how imbalanced the data is (the concurrence of minority and majority labels

in the same instances) [16]. Within the literature it is generally agreed upon that undersampling algorithms whether random or heuristic perform worse than oversampling ones since they cause a loss of essential information by removing instances[16, 21, 41]. Therefore, undersampling techniques should not be applied to multi-label datasets, which are not truly imbalanced [20]. Literature shows that in general ML-ROS is the most effective resampling method; however, for the most imbalanced datasets MLTL achieves better results [117].

7 Concluding Remarks

This work aims to categorize preprocessing techniques in multi-label classification. The preprocessing techniques in this work are grouped, based on the multi-label classification challenges they attempt to solve: dimensionality reduction and resampling. The former tries to overcome the curse of dimensionality, by representing the original data on a low dimensional representation. The latter tries to overcome the challenge of learning from imbalanced data. We proposed an original taxonomy of dimensionality reduction and resampling techniques that could be performed independently from the choice of multi-label classifier. Each categorisation is described and their respective techniques are reviewed. We note that it is difficult to determine the most efficient and/or competitive technique for each category since only partial comparisons are generally reported in the articles. To the best of our knowledge there exist no experimental studies, which compare all the approaches within a category. Moreover, the reported experimental results are highly dependent on the nature of the used multi-label dataset and the chosen measures. As far as we know, up until now the dimensionality reduction and resampling techniques were scattered among literature. By collecting and categorizing the existing literature, it is our hope that this thesis can be used as a starting point and reference for future researchers.

Table 1. Table of Multi-Label Feature Selection methods using the Transformation Approach

Abbreviation	**Name**	**Type**	**References**
LP-CHI	Label Powerset - Chi-square	Filter	[144, 72]
BR-CHI	Binary Relevance - Chi-Square	Filter	[72]
PPT-MI	Pruned Problem Transformation - Mutual Information	Filter	[32, 33]
BR-RF	Binary Relevance - Relieff	Filter	[137, 72]
BR-IG	Binary Relevance - Information Gain	Filter	[137, 57, 72]
LP-RF	Label Powerset - Relieff	Filter	[137, 72]
LP-IG	Label Powerset - Information Gain	Filter	[137, 72]
LCFS	Label Construction for multi-label Feature Selection	Filter	[136]
BRRF	Binary Relevance - Random Forest	Wrapper	[43]
RFLP	Random Forest - Label Powerset	Wrapper	[43]
PPT-ReliefF	Pruned Problem Transformation - ReliefF	Filter	[122]
BR-CLS	Binary Relevance - Constrained Laplacian Scores	Filter	[3]
LP-CLS	Label Powerset - Constrained Laplacian Scores	Filter	[3]
BR-FSCORE	Binary Relevance - Fscore	Filter	[72]
LP-FSCORE	Label Powerset - Fscore	Filter	[72]
BR-FCBF	Binary Relevance - Fast Correlation Based Filter	Filter	[72]
LP-FCBF	Label Powerset - Fast Correlation Based Filter	Filter	[72]
BR-CFS	Binary Relevance - Correlation-based Feature Selection	Filter	[72]
LP-CFS	Label Powerset - Correlation-based Feature Selection	Filter	[72]

Table 2. Table of Multi-Label Feature Selection methods using the Direct Approach

Abbreviation	**Name**	**Type**	**References**
MLNB	Multi-Label classification with Naive Bayes	Wrapper	[172]
MLFR	Multi-Label Feature Ranker	Filter	[84]
gMLC	ML-FS method for graph classification	Filter	[79]
MRelief	Relief algorithm for multi-label feature selection	Filter	[78]
ReliefF-ML	extension of ReliefF algorithm to multi-label learning	Filter	[109, 122]
RReliefF-ML	extension of RReliefF algorithm to multi-label learning	Filter	[122]
HOML	Hybrid Optimization based Multi-Label	Wrapper	[130]
ML-CFS	adaptation of the Correlation-based Feature Selection	Filter	[70]
GA-ML-CFS	Genetic Algorithm for Multi-label Correlation-based Feature Selection	Filter	[68]
LexGa-ML-CFS	Lexicographic multi-objective Genetic Algorithm for Multi-Label Correlation-based Feature Selection	Filter	[69]
PMU	ML-FS method based on multivariate mutual information	Filter	[85]
RFPCT	Random Forest Predictive Clustering Tree	Wrapper	[43]
MDMR	Max-Dependency and Min-Redundancy criterion	Filter	[91]
SCLS	multi-label feature selection based on scalable relevance evaluation	Filter	[87]
LRFS	Feature Selection based on Label Redundancy	Filter	[176]
ParetoCluster	feature selection method based on the Pareto optimal set	Filter	[71]
MGFS	multi-label graph-based feature selection algorithm via PageRank centrality	Filter	[51]
FS-MLC	Feature Selection for Multi-Label classification using Clustering in feature-space	Wrapper	[104]

Table 3. Table of Unsupervised Feature Extraction methods

Abbreviation	**Name**	**Type**	**Reference**
PCA	Principal Component Analysis	Linear	[113]
SVD	Singular Value Decomposition	Linear	[45, 150]
PP	Projection Pursuit	Linear	[39]
MDS	Multidimensional scaling	Non-Linear	[82]
SOM	Self-Organizing Map	Non-Linear	[75, 76]
KPCA	Kernel Principal Component Analysis	Non-Linear	[127]
LSI	Latent Semantic Indexing		[29]
LLE	Locally Linear Embedding	Non-Linear	[126]
t-SNE	T-Stochastic neighbour embedding	Non-Linear	[59]
LE	Laplacian Eigenmap	Non-Linear	[5]
LTSA	Local Tangent Space Alignment	Non-Linear	[179]
LPP	Locality Preserving Projection	Linear	[55]
CNMF	Constrained Non-negative Matrix Factorization		[98]
RPCA	Random Principal Component Analysis	Linear	[156]
AE	Auto-Encoder	Non-Linear	[58]
MSSBoost	Model-Shared Subspace Boosting	Linear	[163]
OLPP	Orthonormal Locality Preserving Projection	Linear	[77]
ONPP	Orthonormal Neighborhood Preserving Projection	Linear	[77]
MLLP-DR	MultiLocal Linear Pattern Dimensionality Reduction	Linear	[177]
FMEU	Flexible Manifold Embedding for Unsupervised case	Non-Linear	[106]

Table 4. Table of Supervised Feature Extraction methods

Abbreviation	**Name**	**Type**	**Reference**
PLS	Partial Least Square	Linear	[124]
MLSI	Multi-label informed Latent Semantic Indexing	Linear	[166]
MLLS	Multi-Label Least Square	Linear	[65]
MDDM	Multi-Label Dimensionality reduction via Dependance Maximization	Linear	[178]
MLDA	Multi-label Linear Discriminant Analysis	Linear	[152]
DMLDA	Direct Multi-label Linear Discriminant Analysis	Linear	[107]
SSMDDM	Shared Subspace multi-label Dimensionality reduction via Dependence Maximization	Linear	[131]
MVMD	Maximizing feature Variance and feature label Dependence	Linear	[161]

Table 5. Table of Label Space Reduction methods

	Abbreviation	Name	Dependancy	Type	Reference
Label Selection	MOPLMS	Landmark Selection Method for Multiple Output Prediction	Independent	Explicit	[4]
	ML-CSSP	Multi-Label Column Subset Selection Problem	Independent	Explicit	[8]
	LI-MLC	Label Inference for Multilabel Classification	Independent		[18]
Label Embedding	CS	Compressed Sensing	Independent	Explicit	[61]
	CL	Compressed Labeling	Independent	Explicit	[183]
	PLST	Principal Label Space Transformation	Independent	Explicit	[140]
	CPLST	Conditional Principal Label Space Transformation	Dependent	Explicit	[140, 27]
	MLC-BMaD	Multi-Label Classification Using Boolean Matrix Decomposition	Independent	Implicit	[157]
	FaIE	Feature-aware Implicit label space Encoding	Dependent	Implicit	[93]
	Rembrandt	Response EMBedding via RANDomized Techniques	Dependent	Explicit	[103]
	DMLR	Dependence Maximization based Label space Reduction	Dependent	Explicit	[168]
	CLEMS	Cost-sensitive Label Embedding via Multidimensional Scaling	Independent	Implicit	[62]
	ML-MIV	Multi-Label classification via maximizing global recoverability and dependency , and minimizing Local Variances	Dependent	Explicit	[155]

Table 6. Table of Resampling methods

Abbreviation	**Name**	**Type**	**Sampling**	**Reference**
LP-RUS	Label Powerset Random Undersampling	Random	Undersampling	[15]
ML-RUS	Multi-Label Random Undersampling	Random	Undersampling	[21]
LP-ROS	Label Powerset Random Oversampling	Random	Oversampling	[15]
ML-ROS	Multi-Label Random Oversampling	Random	Oversampling	[21]
ML-ENN	MultiLabel edited Nearest Neighbor	Heuristic	Undersampling	[20]
SmoteUG	SMOTE with UG strategy	Heuristic	Oversampling	[44]
ML-SMOTE	Multi-Label Synthetic Minority Over-sampling Technique	Heuristic	Oversampling	[22]
REMEDIAL	REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels	Heuristic	Oversampling	[17]
REMEDIALHwR	REMEDIAL Hybridization with Resampling	Heuristic	Oversampling	[24]
MLTL	Multi-Label Tomek Link	Heuristic	Undersampling	[117]

References

1. Machine Learning and Knowledge Discovery in Databases. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-23528-8
2. Aggarwal, C.C.: Data classification. In: Data Mining. pp. 285–344. Springer (2015)
3. Alalga, A.: Semi-supervised multi-label feature selection. Unpublished (2017). https://doi.org/10.13140/RG.2.2.34571.11040
4. Balasubramanian, K., Lebanon, G.: The landmark selection method for multiple output prediction. In: Proceedings of the 29th International Coference on International Conference on Machine Learning. p. 283–290. ICML'12, Omnipress, Madison, WI, USA (2012)
5. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6), 1373–1396 (Jun 2003). https://doi.org/10.1162/089976603321780317
6. Bellman, R.: Dynamic programming and lagrange multipliers. Proceedings of the National Academy of Sciences of the United States of America **42**(10), 767 (1956)
7. Bhowmick, P.K., Basu, A., Mitra, P., Prasad, A.: Multi-label text classification approach for sentence level news emotion analysis. In: International Conference on Pattern Recognition and Machine Intelligence. pp. 261–266. Springer (2009)
8. Bi, W., Kwok, J.T.: Efficient multi-label classification with many labels. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. p. III–405–III–413. ICML'13, JMLR.org (2013)
9. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
10. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. arXiv preprint cs/0011032 (2000)
11. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9), 1757–1771 (Sep 2004). https://doi.org/10.1016/j.patcog.2004.03.009, http://dx.doi.org/10.1016/j.patcog.2004.03.009
12. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern recognition **37**(9), 1757–1771 (2004)
13. Boutsidis, C., Mahoney, M.W., Drineas, P.: An improved approximation algorithm for the column subset selection problem. In: Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms. pp. 968–977. SIAM (2009)
14. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
15. Charte, F., Rivera, A., del Jesus, M.J., Herrera, F.: A First Approach to Deal with Imbalance in Multi-label Datasets, p. 150–160. Springer Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-40846-5_16, http://dx.doi.org/10.1007/978-3-642-40846-5_16
16. Charte, F., Rivera, A., del Jesus, M.J., Herrera, F.: Concurrence among Imbalanced Labels and Its Influence on Multilabel Resampling Algorithms, p. 110–121. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-07617-1_10, http://dx.doi.org/10.1007/978-3-319-07617-1_10
17. Charte, F., Rivera, A., del Jesus, M.J., Herrera, F.: Resampling Multilabel Datasets by Decoupling Highly Imbalanced Labels, p. 489–501. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-19644-2_41, http://dx.doi.org/10.1007/978-3-319-19644-2_41
18. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Li-mlc: A label inference methodology for addressing high dimensionality in the label space for multilabel classification. IEEE Transactions on Neural Networks and Learning Systems **25**(10), 1842–1854 (Oct 2014). https://doi.org/10.1109/tnnls.2013.2296501

19. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: MLeNN: A First Approach to Heuristic Multilabel Undersampling, p. 1–9. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-10840-7_1, http://dx.doi.org/10.1007/978-3-319-10840-7_1
20. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: MLeNN: A First Approach to Heuristic Multilabel Undersampling, p. 1–9. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-10840-7_1, http://dx.doi.org/10.1007/978-3-319-10840-7_1
21. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163**, 3–16 (Sep 2015). https://doi.org/10.1016/j.neucom.2014.08.091, http://dx.doi.org/10.1016/j.neucom.2014.08.091
22. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Systems **89**, 385–397 (Nov 2015). https://doi.org/10.1016/j.knosys.2015.07.019, http://dx.doi.org/10.1016/j.knosys.2015.07.019
23. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Dealing with difficult minority labels in imbalanced mutilabel data sets. Neurocomputing **326–327**, 39–53 (Jan 2019). https://doi.org/10.1016/j.neucom.2016.08.158, http://dx.doi.org/10.1016/j.neucom.2016.08.158
24. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Remedial-hwr: Tackling multilabel imbalance through label decoupling and data resampling hybridization. Neurocomputing **326–327**, 110–122 (Jan 2019). https://doi.org/10.1016/j.neucom.2017.01.118, http://dx.doi.org/10.1016/j.neucom.2017.01.118
25. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (Jun 2002). https://doi.org/10.1613/jair.953, http://dx.doi.org/10.1613/jair.953
26. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
27. Chen, Y.N., Lin, H.T.: Feature-aware label space dimension reduction for multi-label classification. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. p. 1529–1537. NIPS'12, Curran Associates Inc., Red Hook, NY, USA (2012)
28. Davis, J.C., Sampson, R.J.: Statistics and data analysis in geology, vol. 646. Wiley New York (1986)
29. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE **41**(6), 391–407 (1990)
30. Ding, S.: Feature selection based f-score and aco algorithm in support vector machine. In: 2009 Second International Symposium on Knowledge Acquisition and Modeling. vol. 1, pp. 19–23. IEEE (2009)
31. Diplaris, S., Tsoumakas, G., Mitkas, P.A., Vlahavas, I.: Protein classification with multiple algorithms. In: Panhellenic Conference on Informatics. pp. 448–456. Springer (2005)
32. Doquire, G., Verleysen, M.: Feature Selection for Multi-label Classification Problems, p. 9–16. Springer Berlin Heidelberg (2011). https://doi.org/10.1007/978-3-642-21501-8_2

33. Doquire, G., Verleysen, M.: Mutual information-based feature selection for multilabel classification. Neurocomputing **122**, 148–155 (Dec 2013). https://doi.org/10.1016/j.neucom.2013.06.035
34. Duda, R., Hart, P., G.Stork, D.: Pattern Classification, vol. xx (01 2001)
35. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. John Wiley & Sons (2012)
36. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in neural information processing systems. pp. 681–687 (2002)
37. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine **17**(3), 37–37 (1996)
38. Forman, G.: An extensive empirical study of feature selection metrics for text classification [j]. Journal of Machine Learning Research - JMLR **3** (03 2003)
39. Friedman, J., Tukey, J.: A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers **C-23**(9), 881–890 (Sep 1974). https://doi.org/10.1109/t-c.1974.224051
40. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-10247-4, http://dx.doi.org/10.1007/978-3-319-10247-4
41. García, V., Sánchez, J., Mollineda, R.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems **25**(1), 13–21 (Feb 2012). https://doi.org/10.1016/j.knosys.2011.06.013, http://dx.doi.org/10.1016/j.knosys.2011.06.013
42. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5958–5966 (2018)
43. Gharroudi, O., Elghazel, H., Aussem, A.: A Comparison of Multi-Label Feature Selection Methods Using the Random Forest Paradigm, p. 95–106. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-06483-3_9
44. Giraldo-Forero, A.F., Jaramillo-Garzón, J.A., Ruiz-Muñoz, J.F., Castellanos-Domínguez, C.G.: Managing Imbalanced Data Sets in Multi-label Problems: A Case Study with the SMOTE Algorithm, p. 334–342. Springer Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-41822-8_42, http://dx.doi.org/10.1007/978-3-642-41822-8_42
45. Golub, G.H., Reinsch, C.: Singular Value Decomposition and Least Squares Solutions, p. 134–151. Springer Berlin Heidelberg (1971). https://doi.org/10.1007/978-3-642-86940-2_10
46. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring Statistical Dependence with Hilbert-Schmidt Norms, p. 63–77. Springer Berlin Heidelberg (2005). https://doi.org/10.1007/11564089_7
47. Guyon, Nikravesh, G., Zadeh: Feature Extraction. Springer Berlin Heidelberg (2006). https://doi.org/10.1007/978-3-540-35488-8
48. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(null), 1157–1182 (Mar 2003)
49. Hall, M.A.: Correlation-based feature selection for machine learning (1999)
50. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
51. Hashemi, A., Dowlatshahi, M.B., Nezamabadi-pour, H.: Mgfs: A multi-label graph-based feature selection algorithm via pagerank centrality. Expert Systems with Applications **142**, 113024 (Mar 2020). https://doi.org/10.1016/j.eswa.2019.113024, http://dx.doi.org/10.1016/j.eswa.2019.113024

52. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009)
53. He, H., Ma, Y.: Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons (2013)
54. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in neural information processing systems. pp. 507–514 (2006)
55. He, X., Niyogi, P.: Locality preserving projections. In: In Advances in Neural Information Processing Systems 16. MIT Press (2003)
56. Hernández, M.A., Stolfo, S.J.: Data Mining and Knowledge Discovery **2**(1), 9–37 (1998). https://doi.org/10.1023/a:1009761603038, http://dx.doi.org/10.1023/A:1009761603038
57. Herrera, F., Charte, F., Rivera, A.J., del Jesus, M.J.: Multilabel Classification. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-41111-8
58. Hinton, G.E.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (Jul 2006). https://doi.org/10.1126/science.1127647
59. Hinton, G.E., Roweis, S.T.: Stochastic neighbor embedding. In: Advances in neural information processing systems. pp. 857–864 (2003)
60. Hoteling, H.: Relations between two sets of variates. Biometrika **28**(3–4), 321–377 (Dec 1936). https://doi.org/10.1093/biomet/28.3-4.321
61. Hsu, D.J., Kakade, S.M., Langford, J., Zhang, T.: Multi-label prediction via compressed sensing. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 772–780. Curran Associates, Inc. (2009)
62. Huang, K.H., Lin, H.T.: Cost-sensitive label embedding for multi-label classification. Machine Learning **106**(9–10), 1725–1746 (Aug 2017). https://doi.org/10.1007/s10994-017-5659-z
63. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis **6**(5), 429–449 (2002)
64. Ji, S., Sun, L., Jin, R., Kumar, S., Ye, J.: Automated annotation of drosophila gene expression patterns using a controlled vocabulary. Bioinformatics **24**(17), 1881–1888 (Jul 2008). https://doi.org/10.1093/bioinformatics/btn347, http://dx.doi.org/10.1093/bioinformatics/btn347
65. Ji, S., Tang, L., Yu, S., Ye, J.: Extracting shared subspace for multi-label classification. pp. 381–389 (08 2008). https://doi.org/10.1145/1401890.1401939
66. Jian, L., Li, J., Shu, K., Liu, H.: Multi-label informed feature selection. In: IJCAI. pp. 1627–1633 (2016)
67. Jiang, L., Zhang, L., Li, C., Wu, J.: A correlation-based feature weighting filter for naive bayes. IEEE transactions on knowledge and data engineering **31**(2), 201–213 (2018)
68. Jungjit, S., Freitas, A.: A lexicographic multi-objective genetic algorithm for multi-label correlation based feature selection. In: Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference - GECCO Companion '15. ACM Press (2015). https://doi.org/10.1145/2739482.2768448
69. Jungjit, S., Freitas, A.: A new genetic algorithm for multi-label correlation-based feature selection (04 2015)
70. Jungjit, S., Michaelis, M., Freitas, A.A., Cinatl, J.: Two extensions to multi-label correlation-based feature selection: A case study in bioinformatics. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE (Oct 2013). https://doi.org/10.1109/smc.2013.262, http://dx.doi.org/10.1109/SMC.2013.262

71. Kashef, S., Nezamabadi-pour, H.: A label-specific multi-label feature selection algorithm based on the pareto dominance concept. Pattern Recognition **88**, 654–667 (Apr 2019). https://doi.org/10.1016/j.patcog.2018.12.020, http://dx.doi.org/10.1016/j.patcog.2018.12.020
72. Kashef, S., Nezamabadi-pour, H., Nikpour, B.: Multilabel feature selection: A comprehensive review and guiding experiments. WIREs Data Mining and Knowledge Discovery **8**(2) (Jan 2018). https://doi.org/10.1002/widm.1240
73. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: Proceedings of the ECML/PKDD. vol. 18, p. 5 (2008)
74. Kohavi, R., John, G.H., et al.: Wrappers for feature subset selection. Artificial intelligence **97**(1-2), 273–324 (1997)
75. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990). https://doi.org/10.1109/5.58325
76. Kohonen, T.: The self-organizing map. Neurocomputing **21**(1–3), 1–6 (Nov 1998). https://doi.org/10.1016/s0925-2312(98)00030-7
77. Kokiopoulou, E., Saad, Y.: Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2143–2156 (Dec 2007). https://doi.org/10.1109/tpami.2007.1131
78. Kong, D., Ding, C., Huang, H., Zhao, H.: Multi-label relieff and f-statistic feature selections for image annotation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE (Jun 2012). https://doi.org/10.1109/cvpr.2012.6247947
79. Kong, X., Yu, P.S.: gmlc: a multi-label feature selection framework for graph classification. Knowledge and Information Systems **31**(2), 281–305 (May 2011). https://doi.org/10.1007/s10115-011-0407-3
80. Kononenko, I.: Estimating attributes: analysis and extensions of relief. In: European conference on machine learning. pp. 171–182. Springer (1994)
81. Kotsiantis, S., Pintelas, P.: Mixture of expert agents for handling imbalanced data sets. Annals of Mathematics, Computing & Teleinformatics **1**, 46–55 (01 2004)
82. Kruskal, J.B.: Multidimensional scaling. No. 11, Sage (1978)
83. Kumar, V., Pujari, A.K., Padmanabhan, V., Kagita, V.R.: Group preserving label embedding for multi-label classification. Pattern Recognition **90**, 23–34 (Jun 2019). https://doi.org/10.1016/j.patcog.2019.01.009, http://dx.doi.org/10.1016/j.patcog.2019.01.009
84. Lastra, G., Luaces, O., Quevedo, J.R., Bahamonde, A.: Graphical feature selection for multilabel classification tasks. In: IDA (2011)
85. Lee, J., Kim, D.W.: Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters **34**(3), 349–357 (Feb 2013). https://doi.org/10.1016/j.patrec.2012.10.005
86. Lee, J., Kim, D.W.: Fast multi-label feature selection based on information-theoretic feature ranking. Pattern Recognition **48**(9), 2761–2771 (Sep 2015). https://doi.org/10.1016/j.patcog.2015.04.009
87. Lee, J., Kim, D.W.: Scls: Multi-label feature selection based on scalable criterion for large label set. Pattern Recognition **66**, 342–352 (Jun 2017). https://doi.org/10.1016/j.patcog.2017.01.014, http://dx.doi.org/10.1016/j.patcog.2017.01.014
88. Li, G.Z., You, M., Ge, L., Yang, J.Y., Yang, M.Q.: Feature selection for semi-supervised multi-label learning with application to gene function analysis. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. pp. 354–357 (2010)

89. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection. ACM Computing Surveys **50**(6), 1–45 (Jan 2018). https://doi.org/10.1145/3136625, http://dx.doi.org/10.1145/3136625
90. Li, T., Ogihara, M.: Toward intelligent music information retrieval. IEEE Transactions on Multimedia **8**(3), 564–574 (2006)
91. Lin, Y., Hu, Q., Liu, J., Duan, J.: Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing **168**, 92–103 (Nov 2015). https://doi.org/10.1016/j.neucom.2015.06.010
92. Lin, Z., Ding, G., Han, J., Shao, L.: End-to-end feature-aware label space encoding for multilabel classification with many classes. IEEE Transactions on Neural Networks and Learning Systems **29**(6), 2472–2487 (Jun 2018). https://doi.org/10.1109/tnnls.2017.2691545, http://dx.doi.org/10.1109/TNNLS.2017.2691545
93. Lin, Z., Ding, G., Hu, M., Wang, J.: Multi-label classification via feature-aware implicit label space encoding. vol. 2 (06 2014)
94. Liu, H., Motoda, H.: Less Is More, p. 3–12. Springer US (1998). https://doi.org/10.1007/978-1-4615-5725-8_1
95. Liu, H., Motoda, H.: Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series). Chapman & Hall/CRC (2007)
96. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. pp. 388–391. IEEE (1995)
97. Liu, T.Y., Yang, Y., Wan, H., Zeng, H.J., Chen, Z., Ma, W.Y.: Support vector machines classification with a very large-scale taxonomy. Acm Sigkdd Explorations Newsletter **7**(1), 36–43 (2005)
98. Liu, Y., Jin, R., Yang, L.: Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1. p. 421–426. AAAI'06, AAAI Press (2006)
99. Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., Herrera, F.: Big Data Preprocessing. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-39105-8, http://dx.doi.org/10.1007/978-3-030-39105-8
100. Mandziuk, J., Zychowski, A.: Dimensionality reduction in multilabel classification with neural networks. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE (Jul 2019). https://doi.org/10.1109/ijcnn.2019.8852156, http://dx.doi.org/10.1109/IJCNN.2019.8852156
101. McCallum, A.K.: Multi-label text classification with a mixture model trained by em. In: AAAI 99 Workshop on Text Learning (1999)
102. Miettinen, P.: The boolean column and column-row matrix decompositions. Data Mining and Knowledge Discovery **17**(1), 39–56 (Jul 2008). https://doi.org/10.1007/s10618-008-0107-0
103. Mineiro, P., Karampatziakis, N.: Fast Label Embeddings via Randomized Linear Algebra, p. 37–51. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-23528-8_3
104. Mishra, N.K., Singh, P.K.: Fs-mlc: Feature selection for multi-label classification using clustering in feature space. Information Processing & Management **57**(4), 102240 (Jul 2020). https://doi.org/10.1016/j.ipm.2020.102240, http://dx.doi.org/10.1016/j.ipm.2020.102240

105. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(3), 301–312 (2002)
106. Nie, F., Xu, D., Tsang, I.W.H., Zhang, C.: Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. IEEE Transactions on Image Processing **19**(7), 1921–1932 (Jul 2010). https://doi.org/10.1109/tip.2010.2044958
107. Oikonomou, M., Tefas, A.: Direct multi-label linear discriminant analysis. vol. 383, pp. 414–423 (09 2013). https://doi.org/10.1007/978-3-642-41013-0_43
108. Olsson, J.O.S., Oard, D.W.: Combining feature selectors for text classification. In: Proceedings of the 15th ACM international conference on Information and knowledge management. pp. 798–799 (2006)
109. Oscar Gabriel Reyes Pupo, C.M.&.S.V.t.S.: Relieff-ml: An extension of relieff algorithm to multi-label learning. pp. 528–535 (11 2013). https://doi.org/10.1007/978-3-642-41827-3_66
110. Pacharawongsakda, E., Theeramunkong, T.: Multi-label classification using dependent and independent dual space reduction. The Computer Journal **56**(9), 1113–1135 (Feb 2013). https://doi.org/10.1093/comjnl/bxs169, http://dx.doi.org/10.1093/comjnl/bxs169
111. Pacharawongsakda, E., Theeramunkong, T.: A two-stage dual space reduction framework for multi-label classification. Unpublished (2013). https://doi.org/10.13140/2.1.1814.0166, http://rgdoi.net/10.13140/2.1.1814.0166
112. Pacharawongsakda, E., Theeramunkong, T.: A Comparative Study on Single and Dual Space Reduction in Multi-label Classification, p. 389–400. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-19090-7_29
113. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (Nov 1901). https://doi.org/10.1080/14786440109462720
114. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence **27**(8), 1226–1238 (2005)
115. Pereira, R.B., Plastino, A., Zadrozny, B., de Campos Merschmann, L.H.: Information gain feature selection for multi-label classification. J. Inf. Data Manag. **6**, 48–58 (2015)
116. Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H.C.: Categorizing feature selection methods for multi-label classification. Artificial Intelligence Review **49**(1), 57–78 (Sep 2016). https://doi.org/10.1007/s10462-016-9516-4
117. Pereira, R.M., Costa, Y.M., Silla Jr., C.N.: Mltl: A multi-label approach for the tomek link undersampling algorithm. Neurocomputing **383**, 95–105 (Mar 2020). https://doi.org/10.1016/j.neucom.2019.11.076, http://dx.doi.org/10.1016/j.neucom.2019.11.076
118. Pyle, D.: Data preparation for data mining. morgan kaufmann (1999)
119. Qiao, L., Zhang, L., Sun, Z., Liu, X.: Selecting label-dependent features for multi-label classification. Neurocomputing **259**, 112–118 (Oct 2017). https://doi.org/10.1016/j.neucom.2016.08.122
120. Rao, T.R., Mitra, P., Bhatt, R., Goswami, A.: The big data system, components, tools, and technologies: a survey. Knowledge and Information Systems **60**(3), 1165–1245 (Sep 2018). https://doi.org/10.1007/s10115-018-1248-0, http://dx.doi.org/10.1007/s10115-018-1248-0

121. Read, J.: A pruned problem transformation method for multi-label classification (01 2008)
122. Reyes, O., Morell, C., Ventura, S.: Scalable extensions of the relieff algorithm for weighting and selecting features on the multi-label learning context. Neurocomputing **161**, 168–182 (Aug 2015). https://doi.org/10.1016/j.neucom.2015.02.045
123. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Machine learning **53**(1-2), 23–69 (2003)
124. Rosipal, R., Krämer, N.: Overview and Recent Advances in Partial Least Squares, p. 34–51. Springer Berlin Heidelberg (2006). https://doi.org/10.1007/11752790_2
125. Rouhi, A., Nezamabadi-Pour, H.: Feature Selection in High-Dimensional Data, p. 85–128. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-34094-0_5, http://dx.doi.org/10.1007/978-3-030-34094-0_5
126. Roweis, S.T.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (Dec 2000). https://doi.org/10.1126/science.290.5500.2323
127. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis, p. 583–588. Springer Berlin Heidelberg (1997). https://doi.org/10.1007/bfb0020217
128. Sechidis, K., Nikolaou, N., Brown, G.: Information Theoretic Feature Selection in Multi-label Data through Composite Likelihood, p. 143–152. Springer Berlin Heidelberg (2014). https://doi.org/10.1007/978-3-662-44415-3_15
129. Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., Wang, Z.: A novel feature selection algorithm for text categorization. Expert systems with applications **33**(1), 1–5 (2007)
130. Shao, H., Li, G., Liu, G., Wang, Y.: Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine. Science China Information Sciences **56**(5), 1–13 (Jan 2012). https://doi.org/10.1007/s11432-011-4406-5
131. Shu, X., Lai, D., Xu, H., Tao, L.: Learning shared subspace for multi-label dimensionality reduction via dependence maximization. Neurocomputing **168**, 356–364 (Nov 2015). https://doi.org/10.1016/j.neucom.2015.05.090
132. Siblini, W., Kuntz, P., Meyer, F.: A review on dimensionality reduction for multi-label classification. IEEE Transactions on Knowledge and Data Engineering p. 1–1 (2019). https://doi.org/10.1109/tkde.2019.2940014
133. Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A review of unsupervised feature selection methods. Artificial Intelligence Review **53**(2), 907–948 (Jan 2019). https://doi.org/10.1007/s10462-019-09682-y, http://dx.doi.org/10.1007/s10462-019-09682-y
134. Sorower, M.S.: A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis **18**, 1–25 (2010)
135. Sorzano, C.O.S., Vargas, J., Montano, A.P.: A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877 (2014)
136. Spolaor, N., Monard, M.C., Tsoumakas, G., Lee, H.: Label construction for multi-label feature selection. In: 2014 Brazilian Conference on Intelligent Systems. IEEE (Oct 2014). https://doi.org/10.1109/bracis.2014.52
137. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science **292**, 135–151 (Mar 2013). https://doi.org/10.1016/j.entcs.2013.02.010
138. Spolaôr, N., Monard, M.C., Tsoumakas, G., Lee, H.D.: A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing **180**, 3–15 (Mar 2016). https://doi.org/10.1016/j.neucom.2015.07.118

139. Szymański, P., Kajdanowicz, T., Chawla, N.: Lnemlc: Label network embeddings for multi-label classifiation (12 2018)
140. Tai, F., Lin, H.T.: Multilabel classification with principal label space transformation. Neural Computation **24**(9), 2508–2542 (Sep 2012). https://doi.org/10.1162/necoa00320
141. Tang, J., Alelyani, S., Liu, H.: Data classification: algorithms and applications. Data Mining and Knowledge Discovery Series, CRC Press pp. 37–64 (2014)
142. Tang, L., Liu, L., Gan, J.: An overview of label space dimension reduction for multi-label classification. In: Proceedings of the 2nd International Conference on Intelligent Information Processing - IIP'17. ACM Press (2017). https://doi.org/10.1145/3144789.3144807
143. Tomek, I.: Two modifications of cnn. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(11), 769–772 (Nov 1976). https://doi.org/10.1109/tsmc.1976.4309452, http://dx.doi.org/10.1109/TSMC.1976.4309452
144. Trochidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music into emotions. vol. 2011, pp. 325–330 (01 2008). https://doi.org/10.1186/1687-4722-2011-426793
145. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music by emotion. EURASIP Journal on Audio, Speech, and Music Processing **2011**(1) (Sep 2011). https://doi.org/10.1186/1687-4722-2011-426793
146. Tsoumakas, G., Katakis, I.: Multi-label classification. International Journal of Data Warehousing and Mining **3**(3), 1–13 (Jul 2007). https://doi.org/10.4018/jdwm.2007070101
147. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data, p. 667–685. Springer US (2009). https://doi.org/10.1007/978-0-387-09823-4_34
148. Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S., Moore, J.H.: Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics **85**, 189–203 (Sep 2018). https://doi.org/10.1016/j.jbi.2018.07.014, http://dx.doi.org/10.1016/j.jbi.2018.07.014
149. Vafaie, H., Imam, I.F.: Feature selection methods: genetic algorithms vs. greedy-like search. In: Proceedings of the international conference on fuzzy and intelligent control systems. vol. 51, p. 28 (1994)
150. Van Loan, C.F.: Generalizing the singular value decomposition. SIAM Journal on numerical Analysis **13**(1), 76–83 (1976)
151. Vinh, L.T., Lee, S., Park, Y.T., d'Auriol, B.J.: A novel feature selection method based on normalized mutual information. Applied Intelligence **37**(1), 100–120 (Aug 2011). https://doi.org/10.1007/s10489-011-0315-y
152. Wang, H., Ding, C., Huang, H.: Multi-label linear discriminant analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. pp. 126–139. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
153. Wang, H., Nie, F., Huang, H.: Globally and locally consistent unsupervised projection. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
154. Wang, X., Li, J., Xu, J.: A Label Embedding Method for Multi-label Classification via Exploiting Local Label Correlations, p. 168–180. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-36802-9_19, http://dx.doi.org/10.1007/978-3-030-36802-9_19
155. Wang, X., Li, J., Xu, J.: A Label Embedding Method for Multi-label Classification via Exploiting Local Label Correlations, pp. 168–180 (12 2019). https://doi.org/10.1007/978-3-030-36802-9_19

156. Warmuth, M., Kuzmin, D.: Randomized pca algorithms with regret bounds that are logarithmic in the dimension. vol. 19, pp. 1481–1488 (11 2006)
157. Wicker, J., Pfahringer, B., Kramer, S.: Multi-label classification using boolean matrix decomposition. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12. ACM Press (2012). https://doi.org/10.1145/2245276.2245311
158. Willoughby, R.A.: Solutions of ill-posed problems (an tikhonov and vy arsenin). SIAM Review **21**(2), 266 (1979)
159. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics (3), 408–421 (1972)
160. Xu, C., Tao, D., Xu, C.: Robust extreme multi-label learning. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (Aug 2016). https://doi.org/10.1145/2939672.2939798
161. Xu, J., Liu, J., Yin, J., Sun, C.: A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously. Knowledge-Based Systems **98**, 172–184 (Apr 2016). https://doi.org/10.1016/j.knosys.2016.01.032
162. Xu, J., Mao, Z.H.: Multilabel feature extraction algorithm via maximizing approximated and symmetrized normalized cross-covariance operator. IEEE Transactions on Cybernetics p. 1–14 (2019). https://doi.org/10.1109/tcyb.2019.2909779, http://dx.doi.org/10.1109/tcyb.2019.2909779
163. Yan, R., Tesic, J., Smith, J.R.: Model-shared subspace boosting for multi-label classification. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07. ACM Press (2007). https://doi.org/10.1145/1281192.1281281
164. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Icml. vol. 97, p. 35. Nashville, TN, USA (1997)
165. Yilmaz, T., Yazici, A., Kitsuregawa, M.: Relief-mm: effective modality weighting for multimedia information retrieval. Multimedia systems **20**(4), 389–413 (2014)
166. Yu, K., Yu, S., Tresp, V.: Multi-label informed latent semantic indexing. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05. ACM Press (2005). https://doi.org/10.1145/1076034.1076080
167. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 856–863 (2003)
168. Zhang, J.J., Fang, M., Wang, H., Li, X.: Dependence maximization based label space dimension reduction for multi-label classification. Engineering Applications of Artificial Intelligence **45**, 453–463 (Oct 2015). https://doi.org/10.1016/j.engappai.2015.07.023
169. Zhang, L., Duan, Q.: A feature selection method for multi-label text based on feature importance. Applied Sciences **9**(4), 665 (Feb 2019). https://doi.org/10.3390/app9040665, http://dx.doi.org/10.3390/app9040665
170. Zhang, L., Jiang, L., Li, C.: A new feature selection approach to naive bayes text classifiers. International Journal of Pattern Recognition and Artificial Intelligence **30**(02), 1650003 (2016)
171. Zhang, L., Jiang, L., Li, C., Kong, G.: Two feature weighting approaches for naive bayes text classifiers. Knowledge-Based Systems **100**, 137–144 (2016)
172. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences **179**(19), 3218–3229 (2009)

173. 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 (Oct 2006). https://doi.org/10.1109/tkde.2006.162, http://dx.doi.org/10.1109/TKDE.2006.162
174. 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)
175. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering **26**(8), 1819–1837 (2013)
176. Zhang, P., Liu, G., Gao, W.: Distinguishing two types of labels for multi-label feature selection. Pattern Recognition **95**, 72–82 (Nov 2019). https://doi.org/10.1016/j.patcog.2019.06.004, http://dx.doi.org/10.1016/j.patcog.2019.06.004
177. Zhang, S., Ma, Z., Zhang, G., Gan, W.: Dimensionality reduction based on multilocal linear pattern preservation. IEEE Transactions on Knowledge and Data Engineering p. 1–1 (2020). https://doi.org/10.1109/tkde.2020.2999504
178. Zhang, Y., Zhou, Z.H.: Multilabel dimensionality reduction via dependence maximization. ACM Transactions on Knowledge Discovery from Data **4**(3), 1–21 (Oct 2010). https://doi.org/10.1145/1839490.1839495
179. Zhang, Z., Zha, H.: Nonlinear Dimension Reduction via Local Tangent Space Alignment, p. 477–481. Springer Berlin Heidelberg (2003). https://doi.org/10.1007/978-3-540-45080-166
180. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th international conference on Machine learning. pp. 1151–1157 (2007)
181. Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H.: Advancing feature selection research. ASU Feature Selection Repository Arizona State University pp. 1–28 (01 2010)
182. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter **6**(1), 80–89 (2004)
183. Zhou, T., Tao, D., Wu, X.: Compressed labeling on distilled labelsets for multi-label learning. Machine Learning **88**(1–2), 69–126 (Jan 2012). https://doi.org/10.1007/s10994-011-5276-1