Adapting the CMIM algorithm for multi-label feature selection. A comparison with existing methods.

Pablo Bermejo, José A. Gámez, José M. Puerta *

Abstract

The multi-label paradigm has recently attracted the attention of the machine learning community, multi-label problems being those which do not have only one class but several binomial classes instead. Although intensive research has been carried on lately into the multi-label classification paradigm, this is not the case of feature subset selection methods. In this work we propose an adaptation of the well-known CMIM feature selection algorithm, which is capable of approximating the conditional multivariate mutual information of each candidate attribute with respect to the whole set of labels. This capacity to search any degree of interaction among labels, is the reason why our proposal performs better than other state-of-the-art algorithms when the dataset on which it is run contains correlated labels.

This is the accepted version of:

Pablo Bermejo, José A. Gámez, José M. Puerta.
Adapting the CMIM algorithm for multilabel feature selection.
A comparison with existing methods.
Expert Systems. 35(1) (2018)
<https://doi.org/10.1111/exsy.12230>

Please, visit the provided url to obtain the published version.

*Department of Computer Systems, University of Castilla-La Mancha

1 Introduction

Given a set of features $N = \{x_1, \dots, x_{|N|}\}$, supervised feature subset selection (FSS) by means of machine learning methods is the process of automatically identifying the subset of features $S = \{s_1, \dots, s_{|S|}\} \subset N$ with the minimum-cardinality possible and which, marginally or in a multivariate way, best identifies the category that a record or instance belongs to, among a given set of categories. In traditional classification problems, all the categories that a record can belong to are mutually exclusive and are held in just one multinomial class feature $C = \{c_1, \dots, c_{|C|}\}$. This common scenario is known as *single-label* or *multi-class* classification. After running an FSS process on a given dataset, a predictive model only needs to be built using features from S instead of N , and thus it benefits from several advantages: the feature space is reduced so the database and the constructed model are more interpretable; learning the classifier function needs fewer computational resources; and, furthermore, the search method can be designed to detect redundant or irrelevant variables Bermejo et al. (2011), in which case the predictive model could be expected to perform better.

Multi-class FSS has been a well-known research topic over the last few decades, giving rise to a large variety of algorithms which can be classified according to different aspects. By looking at the metric which guides the algorithm through the search space, we find *filter* Yu & Liu (2004), Javed et al. (2010), Bermejo et al. (2010) and *wrapper* Kohavi & John (1997) algorithms. Filter metrics are computed from intrinsic properties of features in the databases, which is a fast process. On the other hand, wrapper methods use the performance of a classifier as metric; this is a relatively more time consuming way to guide the search, but it is commonly assumed that it provides more relevant subsets for the future classification task.

In the last few years, a new classification paradigm has attracted the attention of the machine learning community, namely multi-label classification Tsoumakas & Katakis (2007), in which the set of class variables are binomial and are called *labels*. Thus, given a set $L = \{\lambda_1, \dots, \lambda_{|L|}\}$ of binomial labels whose possible values are *Yes* or *No*, and a dataset whose records are described as $d_i = \{x_1, \dots, x_{|N|}, \lambda_1, \dots, \lambda_{|L|}\}$, the ultimate goal of multi-label classification is to learn a predictive model or function $f()$ which, given a new record $d = \{x_1, \dots, x_{|N|}, \lambda_1 = ?, \dots, \lambda_{|L|} = ?\}$, is capable of correctly predicting the value of all labels; that is, it is capable of predicting which labels can be used to tag d . The main difference from multi-class classification is that now not only the predictive variables might be related to the class(es), but also the labels might present dependencies among them. The nature of this paradigm comes from

real life situations, such as categorizing a text document about the Kyoto Protocol, which could be labeled as both *politics* and *ecology* Doquire & Verleysen (2011); or scene

classification, in which the picture of a beach could be tagged with several labels, such as *sea*, *sky*, *boats*, . . . Furthermore, tagging with a given picture might help the automated process to discard other tags or being more sure about some others; for instance, if a picture is tagged with label *boat*, it is more probable that the scene also belongs to tag *sea* rather than *street*.

In general, multi-labeling can be found in many applications such as bio-informatics Clare et al. (2001), Elisseff & Weston (2001*a*), scene classification Boutell et al. (2004) or mental health Bermejo et al. (2013), among others.

A wide variety of models can be found in the literature for dealing with multi-label classification and FSS. The most common taxonomy found in the literature is made up of 2 categories Tsoumakas & Katakis (2007): *problem transformation* and *algorithm adaptation* methods. *Transformation* of the problem, that is, of the database, is the process of swapping from multi-label to one or several multi-class tasks whose results are later combined. This has the advantage of making any multi-class legacy method available to be applied on multi-label domains. On the other hand, dependencies among labels may be not captured. Ideally, the effort behind the *adaptation* of existing methods Zhang & Zhou (2007), Elisseff & Weston (2001*b*), or the explicit creation of new ones Arias et al. (2014), M.L. & Z.H. (2006), mainly aims to handle labels dependencies.

However, depending on the type of transformation, this split of the database might mean ignoring the potential correlations among labels. There exist 3 fundamental transformations in the literature: Binary Relevance (BR) Sequeira et al. (2012), in which the dataset D is split into $|L|$ datasets $D_1, \dots, D_{|L|}$; Label Powerset (LP) Diplaris et al. (2005), Boutell et al. (2004): This creates a new dataset D' with only one multinomial class C , whose class values are all the existing subsets of labels used to tag records in the training dataset D ; and Pruned Problem Transformation (PPT) Read (2008): Based on LP, it aims to prune class values in D' based on their frequency of appearance.

Thus, both classification and the FSS preprocessing steps need to be transformed or adapted. Furthermore, common stratified cross-validation is not applicable anymore, and some of the frequently found solutions are: non stratified cross-validation Spolaôr et al. (2013), Zhang et al. (2009), using separate training and test datasets Doquire & Verleysen (2013, 2011), or 70%–30% hold-out validation Lee & Kim (2013).

This work presents a new algorithm which focuses on the *FSS* part of the pipeline FSS+classification. Concretely, our proposal is an adaptation of the *CMIM* algorithm Fleuret (2004), which performs a greedy forward selection based on multivariate filter computations. Section 2 is devoted to introducing some of the most recent works proposed by the research community for adaptation of multi-label FSS methods. Section 3 explains in detail our contribution: to adapt the CMIM algorithm to perform feature selection in multi-label datasets. Next, Section 4 shows the results of comparing our proposal against other adapted

methods; and, finally, Section 5 discusses the main conclusions and proposes some future work based on the results obtained.

2 Multi-label FSS methods

As mentioned above, the multi-label paradigm has attracted the attention of the machine learning community from 2013 Sopolaôr et al. (2012) less than 60 related papers where found, which is a small number compared with traditional multi-class FSS methods.

The categorization of multi-label FSS methods can be carried out by following the same taxonomy as for multi-label classification: transformation of the database or adaptation of the algorithm. The latter type of algorithms commonly try to capture label dependencies, as newly proposed methods do too (methods explicitly created for multi-label contexts). That is why our proposal falls into this category. On the other hand, as happens in traditional FSS algorithms, methods from any category might also be filter or wrapper.

2.1 Transformation-based FSS

In the systematic review that Spolaor presents in Sopolaôr et al. (2012), most of the existing methods are transformation-based and use filter metrics, most commonly Information Gain and Chi-Square, but have only been tested on a few datasets. Thus, Spolaor evaluates Sopolaôr & Tsoumakas (2013) on a large benchmark BR approach which first makes a univariate filter (Chi-Square and Binary Split) ranking for each label, and then aggregates the rankings using different criteria: Mean value per label, maximum value per label, and two criteria proposed in Forman (2004) to deal with low frequency labels, namely Round Robin (gives one vote to the best feature per label turn) and also Random Robin (gives a vote to a feature per label randomly chosen, based on its frequency). Being based on filter rankings, it depends on the pre-fixed percentage of features to select. Combining the filter metrics and the aggregation methods, Spolaor evaluates (via a 10 fold cross-validation) 8 FSS methods, with the result that the best option is to aggregate rankings of Chi-Square using the mean value of each feature in all labels. This proposal not only does not take into account label dependences but it also does not capture dependencies, among predictive features.

A recent work in Sopolaôr et al. (2013) proposes to perform FSS using BR and LP methods together with ReliefF and Information Gain metrics. In this case, a pre-defined threshold is used to choose the cut-off point in the rankings. In the case of the LP transformation there exists only one ranking from which to apply the threshold, while in the case of BR

a ranking is created by averaging the scores obtained in each independent label ranking. Thus, 4 different combinations were evaluated (10 fold cross-validation) on a large benchmark and 6 performance metrics (example and label-based) were compared, concluding that the ReliefF metric performed the best, for both BR and LP transformations. In particular, LP-ReliefF performed the best, probably because it implicitly captures correlations among subsetting labels.

In Doquire & Verleysen (2011) Doquire proposes transforming the dataset using PPT, and to run a traditional greedy forward search using the Mutual Information (MI) between each candidate feature and the created class variable as metric to guide the search. This search has the particularity of using an MI estimator introduced in Gómez-Verdejo et al. (2009) which is based on nearest neighbours and does not need to discretize the predictive features to be computed. This proposal has a major drawback, which is the need to prefix 3 parameters: the number of attributes to select, the value of k for the MI estimator used and the pruning threshold of PPT. The latter is automatically set in a recent extension of this work Doquire & Verleysen (2013), which was evaluated on 3 datasets (holdout) and used 4 performance metrics. However, it is not compared against other existing methods.

The only transformation-based FSS which runs a wrapper search was found in the embedded feature selection internally performed when building classifier MLNB in Zhang et al. (2009), and is based on running a genetic algorithm search using a combination of 2 performance metrics as fitness function. Since it is an embedded search, it does not allow the separation of the FSS and classification steps from the pipeline mentioned in Section 1.

2.2 Adaptation-based FSS

The work in Lee & Kim (2013) claims to be the first FSS proposal in the literature which performs adaptation of a filter metric to capture the correlation among labels. The proposed method runs a greedy forward search using an approximation of multivariate Mutual Information as metric to guide the search. Starting from the empty set of selected features $S = \{\emptyset\}$, at each step it selects the feature $x^+ \in N \setminus S$; that is, at each step of the forward search we need to compute the joint mutual information of all features in the current subset S , all labels in L and, x^+ , for all $x^+ \in N \setminus S$. In terms of entropy, this joint information is expressed as

$$\begin{aligned} I(S'; L) &= H(S') - H(S', L) + H(L) = H(\{s_1, \dots, s_{|S|}, x^+\}) \\ &\quad - H(\{s_1, \dots, s_{|S|}, x^+, \lambda_1, \dots, \lambda_{|L|}\}) + H(\{\lambda_1, \dots, \lambda_{|L|}\}) \end{aligned} \tag{1}$$

with $S' = S \cup \{x^+\}$.

Since the computation of $I(S'; L)$ is in general intractable, Lee and Kim (2013) propose an approximation (Equation 2) which only needs to compute joint entropies of subsets of 2 or 3 elements, while taking into account 3 kinds of multivariate mutual information of the candidate attribute with: each label; each attribute in current S and each label; and pairs of labels.

$$\tilde{I}(S'; L) = \sum_{\lambda_i \in L} M(x^+, \lambda_i) - \sum_{s_i \in S} \sum_{\lambda_j \in L} M(x^+, s_i, \lambda_j) - \sum_{\lambda_i \in L} \sum_{\lambda_j \in L} M(x^+, \lambda_i, \lambda_j) \quad (2)$$

$M(T)$, the multivariate mutual information of elements in set T , is defined as:

$$M(T) = - \sum_{X \in T^*} (-1)^{|X|} H(X) \quad (3)$$

, where $\sum_{X \in T^*}$ represents a sum over all elements X drawn from T^* , and T^* denotes the powerset of T .

In a following work Lee & Kim (2014), Lee and Kim propose a generic equation to decide the number of interactions to seek among terms and labels. Thus, Equation 2 is referred to as *3Degree_FL* (3-degree interactions among features and labels); and Lee suggests as a better feature selection criteria an intermediate degree between 2 and 3, which is named *2Degrees_onlyF*, since they found it is not very likely to find 3 degree interactions between labels and/or other labels or features. *2Degrees_onlyF* basically consists on Equation 2 without the last term ($-\sum_{\lambda_i \in L} \sum_{\lambda_j \in L} M(x^+, \lambda_i, \lambda_j)$).

Finally, Lee & Kim (2015) found that the number of ineffective interactions increases exponentially with the number of low-entropy labels in L . Thus, they proposed an affective alleviation on their approach which consists on reducing the number of labels used in the computation of $\tilde{I}(S'; L)$ by discarding the worst labels in L from an entropy-based ranking. Thus, all computations are performed regarding a subset Q of labels, being $2 \leq Q \leq L$; in their experiments they propose as an optimal configuration setting $Q = 10$ and *2Degrees* (2-degree-interaction). It is important to not that both *2Degrees_onlyF*, *3Degree_FL* and *2Degrees_FL_{Q10}* proposals are not capable of detecting redundant features, since they do not compute conditional mutual information.

Correlation-based Feature subset Selection (CFS) Hall (2000) is a traditional multi-class FSS algorithm. It runs a greedy forward search guided by a filter metric which takes into account the redundancy between features and the correlation between features and the class; furthermore, it automatically stops the search without the need of a pre-fixed threshold.

The authors in Fernandes et al. (2013) propose 4 ways to apply CFS to multi-dimensional problems, so it can also be applied in a multi-label context. Two of the proposals are transformation-based. The other two adapt the formulation of the metric which guides the search of CFS, the best performing one being that called CFS_{sum} , which considers the sum of the correlations between each class variable and each feature of the subset evaluated. The formulation of the merit computed in CFS_{sum} is:

$$Merit(s_1, \dots, s_{|S'|}) = \frac{|S'| \sum_{i=1}^{|L|} t_{\lambda_i s}}{\sqrt{|S'| + |S'|(|S'| - 1)t_{ss}}} \quad (4)$$

where $t_{\lambda_i s}$ is the average correlation of all $\langle label_i, feature \rangle$ pairs, and t_{ss} is the average correlation of all $\langle feature, feature \rangle$ pairs; the correlation being measured with the Symmetrical Uncertainty score (SU) Hall (1999). These proposals were evaluated in a large set of experiments, with CFS_{sum} proving to be the best CFS adaptation. As it can be interpreted from Equation 4, CFS_{sum} is not designed to catch interactions among labels. A similar algorithm was proposed in Jungjit et al. (2012), namely $ML - CFS$, which considers the mean of the correlations between each class variable and each feature. This was later extended in Jungjit et al. (2014), but these extensions only outperformed $ML - CFS$ by using explicit domain-knowledge about the dataset used in the experiments. The experiments in Fernandes et al. (2013) show that CFS_{sum} outperforms $ML - CFS$. Just as was the case with $2Degrees_onlyF$ and $3Degree_FL$, we can see in Equation 4 that the adaptation of CFS to the multi-label paradigm is not capable of detecting redundant variables. Furthermore, it does not compute multivariate mutual information.

Again, we find that the adaptation of FSS methods to a multi-label context has been focused on filter methods. If using wrapper approaches is considered highly complex in transformation methods, it is even more complex when the feature selection method is adapted to the multi-label paradigm. To the best of the authors' knowledge, the studies based on Bayesian networks in Ruta (2014) and Zhang et al. (2009) are the only ones which adapt or create new wrapper FSS algorithms, the former being applied to sparse textual datasets, and the latter being an embedded FSS process in classification (inference) time.

3 Adapting CMIM for multi-label FSS

Marginal Mutual Information between two features $I(X; Y)$, or conditional $I(X; Y|Z)$, is a filter metric that is widely used in the FSS literature, being based on the Shannon entropy and capable of detecting non-linear relationships between variables. Exact marginal and conditional duples are easily computable, but increasing the cardinality of the tuples leads to an infeasible metric. Thus, it would not be possible to compute a reliable estimate of $I(x^+; C|S)$ in order to assess the redundancy of a new attribute x^+ with respect to those

already selected in S ; however, in the last years several approximations have appeared Brown et al. (2012), namely *CMIM*, *MRMR* and *MIFS* in Fleuret (2004), Peng et al. (2005), Battiti (1994), respectively, thus making it possible to implement greedy forward FSS methods, which seek correlated features with the class but non redundant among them. One of the mentioned approaches, namely *CMIM* (see Equation 5), is the only approximation for $I(x^+; C|S)$ which computes conditional duples, while the other two approximate by reducing computations to marginal duples.

$$\hat{I}(C; x^+ | \mathcal{S}) = \min_{s_i \in \mathcal{S}} I(C; x^+ | s_i) \quad (5)$$

Basically, *CMIM* consists in running a greedy forward subset selection process starting with $S = \{\emptyset\}$. On each iteration, the process selects the candidate which maximizes the merit computed using Equation 5. The search stops when a pre-fixed percentage of attributes in N are selected.

The study in Fleuret (2004) proposes a fast implementation of *CMIM* which is based on the fact that the merit of each candidate x^+ can only be reduced at each iteration, while at the end of such an iteration the chosen candidate is the one with the highest merit. Thus, when considering a candidate x^+ at iteration it , if the best candidate so far obtained a higher merit than x^+ at iteration $it - 1$, then the merit of x^+ does not need to be computed any more because it can only decrease, and so it is certain not to be selected at the end of the iteration. Since this is not an approximation, it provides the same results as in the original version of *CMIM* but with a much lower execution time. Thus, it seems appropriate to adapt it to the multi-label paradigm since the resulting increased complexity might be balanced by its fast implementation.

Thus, we propose to adapt the canonical *CMIM* to be a multi-label FSS method, namely *ML - CMIM*. First, this implies evolving the merit of candidates x^+ approximated in Equation 5 to Equation 6, thus obtaining computationally intractable calculation of mutual information, which is conditional and multivariate.

$$\hat{I}(L; x^+ | \mathcal{S}) = \min_{s_k \in \mathcal{S}} I(\overbrace{\{\lambda_1, \dots, \lambda_{|L|}\}}^L; x^+ | s_k) \quad (6)$$

Therefore, a second approximation is necessary. Roure Alcobé (2007) gives a recursive approximation for the *marginal* multivariate mutual information of a set of variables as shown in Equation 7 (terminology adapted to the notation used in this article).

$$\begin{aligned}\hat{I}_R(\{\lambda_1, \dots, \lambda_{|L|}\}; x^+) &= \hat{I}_R(\{\lambda_1, \dots, \lambda_{|L|-1}\}; x^+) + I(\lambda_{|L|}; x^+) \\ &+ \sum_{i=1}^{|L|-1} I(\lambda_{|L|}; \lambda_i | x^+) - \sum_{i=1}^{|L|-1} I(\lambda_{|L|}; \lambda_i)\end{aligned}\quad (7)$$

In order to approximate Equation 6, the expression in Equation 7 needs to be conditioned on one variable (s_k), so we obtain Equation 8.

$$\begin{aligned}\hat{I}_{RC}(\{\lambda_1, \dots, \lambda_{|L|}\}; x^+ | s_k) &= \hat{I}_{RC}(\{\lambda_1, \dots, \lambda_{|L|-1}\}; x^+ | s_k) + \\ &I(\lambda_{|L|}; x^+ | s_k) + \sum_{i=1}^{|L|-1} I(\lambda_{|L|}; \lambda_i | x^+, s_k) - \sum_{i=1}^{|L|-1} I(\lambda_{|L|}; \lambda_i | s_k)\end{aligned}\quad (8)$$

Thus, applying Equation 8 in Equation 6 we propose that our *ML – CMIM* algorithm uses Equation 9 to compute the merit of each candidate x^+ . To the best of our knowledge, our proposal is the first multi-label FSS algorithm which computes conditional mutual information; furthermore, it is the only one which takes into account any degree of interaction among labels without needing to set up such degree in advance.

$$\hat{I}(L; x^+ | \mathcal{S}) = \min_{s_k \in \mathcal{S}} \hat{I}_{RC}(\{\lambda_1, \dots, \lambda_{|L|-1}\}, \lambda_{|L|}; x^+ | s_k) \quad (9)$$

Thus, if at a given step of the forward search we have $L = \{l_1, l_2, l_3, l_4\}$, $\mathcal{S} = \{x_5, x_7\}$ and the current candidate feature is $x^+ = x_{10}$, then $\hat{I}_{RC}$ is computed for each selected feature in $\mathcal{S}$; that is, x_5 and x_7 . Then, the score assigned to candidate x_{10} is the minimum $\hat{I}_{RC}$ obtained. For instance, computations for x_5 are shown in Equation 10:

$$\begin{aligned}\hat{I}_{RC}(\{l_1, l_2, l_3, l_4\}; x_{10} | x_5) &= \hat{I}_{RC}(\{l_1, l_2, l_3\}; x_{10} | x_5) + \\ &I(l_4; x_{10} | x_5) + \sum_{l \in \{l_1, l_2, l_3\}} I(l_4; l_i | x_{10}, x_5) - \sum_{l \in \{l_1, l_2, l_3\}} I(l_4; l_i | x_5)\end{aligned}\quad (10)$$

As it can be seen in Equation 10, all exact $I()$ terms are easy to compute, while term $\hat{I}_{RC}(\{l_1, l_2, l_3\}; x_{10} | x_5)$ still needs to continue its recursive development until the base case $\hat{I}_{RC}(\{l_1\}; x_{10} | x_5)$ which can be computed without approximation but with exact conditional mutual information.

Finally, we show in Figure 1 the pseudo-code of the *ML – CMIM* algorithm used, which is adapted from Fleuret (2004).

Although the Roure conditional approximation may result in a large number of recursive calls (depending on the cardinality of L), it is possible to speed up the process by storing the $I(\lambda_i; \lambda_j)$ computations needed in the initialization part, and the $I(\lambda_i; \lambda_j | s_k)$ computations repeated inside each iteration on k .

Finally, Roure reports in Alcobé (2007) that this approximation may be either equal to, greater, or lower than the exact measure, and no analytical bound is established for the error. On the other hand, the author also observes that the proposed expression for mutual information still ‘measures different aspects of the dependencies between the variables’.

Comparing our proposal $ML - CMIM$ with $2Degrees_onlyF$, $3Degree_FL$, $2Degrees_{Q=10}$ and CFS_{sum} , we can see that $ML - CMIM$ is the only multi-label FSS algorithm which computes mutual information conditioned on the current set of selected variables; what makes it being the only algorithm, at least from a theoretical point of view, capable of detecting redundant features. Respect to the computation of mutual information in a multivariate manner, this helps find correlation among features and labels.

We also test in our proposal the optimization by Lee & Kim (2015) explained in Section 2.2, and so we denote this version of our proposal as $ML - CMIM_{Q10}$. The CFS_{sum} algorithm is the only FSS method which does not compute multivariate mutual information, so it does not detect correlation between labels. Furthermore, algorithms $2Degrees_onlyF$ and $3Degree_FL$ force the maximum degree of interactions between labels and attributes to search for; however, the recursive computation of $ML - CMIM$ searches up to the maximum degree of interactions among labels.

4 Experiments

In this section we describe the properties of the 4 datasets used in our experiments, and then we use these properties to explain the results obtained in the experiments.

Our proposal is implemented using the Mulan API Tsoumakas et al. (2011), which extends the Weka Hall et al. (2009) toolkit to allow for multi-label pre-processing and classification tasks. For the sake of reproducibility, our code is available via the following link:

<http://dsi.uclm.es/simd/downloads/ML-CMIM.zip>

4.1 Datasets and metrics

In order to evaluate our proposal and compare it with state-of-the-art multi-label FSS algorithms, we selected 7 datasets with numerical attributes: four datasets from the Mulan repository¹: *cal500*, *scene*, *enron* and *medical*; one dataset from the Meka repository: *slash-dot*²; and 2 versions of *CLEF 2014*³: one with 8 labels and the other with 18 labels. We selected these two datasets because the correlation between labels is very different comparing the former against *CLEF* datasets. Thus, we can show the superior capacity of our proposal

¹<http://mulan.sourceforge.net/datasets-mlc.html>

²<http://meka.sourceforge.net/#download>

³<http://www.rovit.ua.es/dataset/vidrilo/downloads.html>

to detect such interactions.

This evaluation corpus is described in Table 1: name of the dataset; number of instances ($|D|$); number of attributes (N) and number of labels ($|L|$).

Respect to the *CLEF14-8labels* dataset, its 8 labels represent objects which can be found inside a building. Some of these objects can be near other objects (such as *chair* and *printer*), while it is quite unlikely for others (for instance, the *urinal* and *fridge* labels). Thus, we can expect to find a certain degree of correlation between some pairs of labels. In the case of *CLEF14-18labels*, 10 extra labels are included and they relate to the room in which those objects are, so we can find correlated triples of labels such as $\langle \textit{phone}, \textit{bookshelf}, \textit{professoroffice} \rangle$; thus again we can expect to find some correlation between labels. Figure 2 shows a histogram of the mutual information shared between pairs of labels. It is clear that the *CLEF* datasets contain several pairs of labels correlated between them; on the other hand, some datasets such as *cal500*, *enron* or *medical* do not contain almost any pair of labels which share mutual information. Respect to *scene*, it does have some pairs of labels correlated in pairs. It should be noted that none histogram in Figure2 catches triple interactions between labels, and it should be interpreted just as a heuristic explanation of the correlation among labels.

4.2 Performance metrics

As explained in Zhang & Zhou (2014), we cannot use traditional performance metrics such as Accuracy or F-measure since, since performance evaluation in multi-label learning is more complicated than traditional classification, as each record can be associated with multiple labels simultaneously. There exist a wide variety of wrapper-based performance metrics, the most frequently used in the multi-label FSS context being: Hamming Loss and Example-based Accuracy. These two metrics range $[0-1]$; but, while the goal with the former is to be reduced as much as possible, the Example-based Accuracy is the greater the better. These wrapper metrics are defined in Equations 11 and 12, where $|V|$ is the number of records in the test set, Y_i is the set of true labels and Z_i is the set of predicted labels. For a complete definition of multi-label classification metrics, the reader is referred to Zhang & Zhou (2014) and Tsoumakas et al. (2010).

$$HammingLoss = 1/|V| \sum_{i=1}^{|V|} |Y_i \Delta Z_i| / |L| \quad (11)$$

where Δ is the symmetric difference between two sets. This metric counts the times an error is made by predicting an incorrect label for a record, or by not predicting a label which should have been predicted for such a record. The range of *Loss* metrics range is $[0 - 1]$, and

the lower the better the performance.

With respect to Example-based Accuracy,

$$Example - based Accuracy = 1/|V| \sum_{i=1}^{|V|} |Y_i \cap Z_i| / |Y_i \cup Z_i| \quad (12)$$

4.3 Evaluation and comparison framework

We ran our proposal $ML - CMIM$ and the other methods to select from 1 to 100 attributes. For comparison with other multi-label FSS adapted methods, we included in our evaluation the methods proposed by Lee in Lee & Kim (2013) and its extensions proposed in Lee & Kim (2014) and Lee & Kim (2015); we also included the best configuration of the CFS adaptation by Fernandes in Fernandes et al. (2013) (see Section 3). These 4 methods are denoted as $3Degree_FL$, $2Degrees_onlyF$, $2Degrees_FL_{Q10}$ and CFS_{sum} , respectively. We also extended our algorithm with the proposal in Lee & Kim (2015), namely $ML - CMIM_{Q10}$.

An adapted classifier was used for validation, namely $ML - kNN$ Zhang & Zhou (2007), which, first, identifies its K nearest neighbors in the training set; then, based on the number of neighboring records which belong to each possible label, the maximum a posteriori (MAP) principle is used at classification time to determine the label set for a new record. This classifier is suggested in Doquire & Verleysen (2013) for evaluation due to its high sensitivity to the presence of irrelevant features. $ML - KNN$ classifier was used as provided by the Mulan package, with default value $K = 10$. For completion purposes, we also run validation using a transformation-based classifier, namely MLNB Zhang et al. (2009), which splits the database in L datasets and computes predictions using the Naive Bayes classifier.

The methods from the state of the art used for comparison ($3Degree_FL$, $2Degrees_onlyF$, $2Degrees_FL_{Q10}$ and CFS_{sum}) were self-coded. The validation scheme was a 10-fold cross-validation with random seed = 1.

Regarding the discretization of numerical variables (needed to compute entropies), an unsupervised discretization with 2 bins of equal-width was used, as in Lee & Kim (2013).

4.4 Results

In Figures 3 and 4, we show the 2 performance metrics computed for the 7 datasets using the $ML - KNN$ classifier. By looking first at the two *CLEF* datasets, which we know they do have many correlated labels, our proposal $ML - CMIM$ and its optimized version $ML - CMIM_{Q10}$ are the ones which perform the most, converging with 100 attributes to

the performance of not using feature selection. It should be noted as well that $ML - CMIM$ converges before than $2Degrees_onlyF$, $3Degree_FL$ and CFS_{sum} ; probably because it is the only algorithm sensitive to redundant features. Respect to CFS_{sum} , which is the only feature selection algorithm which does not seek for interactions between labels, is the one which performs the worst with these two datasets with correlated labels.

Respect to *cal500*, which contains a few pairs of labels correlated (see Figure 2), we found that $ML - CMIM$ and $ML - CMIM_{Q10}$ are again the ones which use to perform the best; and in this case most of the feature selection algorithms outperform the performance using the original database.

The performance of feature selection on *scene* again shows that is necessary to select many features to approximate the performance of the original database. This dataset contains labels which are correlated, as seen in Figure 2, and so our proposal $ML - CMIM$ and $3Degree_FL$ perform the best when selecting very few features (5 to 10 features). However, the results of CFS_{sum} show that in this dataset it is not worth seeking for interactions between labels. Maybe one reason could be a probable lack of higher degree interactions in the *scene* dataset, which are the aim of $ML - CMIM$, $2Degrees_onlyF$ and $3Degree_FL$.

Respect to the three more high-dimensional datasets, (*enron*, *medical* and *slashdot*), feature selection does improve the performance better than using the original set of attributes, except in some times algorithms $2Degrees_FL_{Q10}$ and $3Degree_FLf$. The three best performing algorithms are: our proposal $ML - CMIM$ and its optimized version $ML - CMIM_{Q10}$, and, CFS_{sum} . Although these 3 datasets do not contain pairs of correlated labels, it seems that our proposal works well in the case of high-dimensional datasets. And CFS_{sum} shows again that it behaves quite well when the dataset does not contain related pairs of labels.

In Figures 5 and 6 we show the performance metrics with the $ML - NB_{Basic}$ classifier. This is a classifier which performs transformation of the dataset by splitting it in $|L|$ binary-class datasets. Thus, applying any effort to identify important attributes by measuring the interaction among 2 or more labels seems to be, a priori, useless since the learned model will not handle any dependency among labels. This explains the results obtained using the $ML - NB_{Basic}$ classifier; as we can see in Figures 5 and 6, the FSS algorithm CFS_{sum} is usually the one which performs the best. On the other hand, by looking at the max values of Hamming Loss and Example-based Accuracy obtained by classifiers $ML - KNN$ and $ML - NB_{Basic}$, we can see that using a classifier which implicitly/explicitly takes label de-

dependencies into account is worthwhile. For instance, the maximum Hamming Loss obtained in *medical* using the $ML - NB_{Basic}$ classifier is 0.040. However, when using the adapted classifier $ML - KNN$, this is reduced to a maximum Hamming Loss of 0.020.

Finally, in Figure 7 we show the mean value of both performance metrics (Hamming-Loss and Example-Based Accuracy) for all methods, fixing the number of selected attributes to 30, 50 and 80. In this comparison, a method is better the nearer it is to the left-upper corner. Having selected a low number of attributes, not performing feature selection performs better than any other method. However, as the number of selected features increases, it is clear that the proposed methods $ML - CMIM$ and $ML - CMIM_{Q10}$ outperform, on average, the rest in both metrics, except in one case for method CFS_{sum} .

5 Conclusions and future work

We proposed an adaptation of the multi-class $CMIM$ algorithm so that it may be used for feature subset selection in multi-label datasets. Up to the author’s knowledge, this is the first multi-label FSS algorithm which computes conditional multivariate mutual information.

We compared our proposal, and an optimized version, on 7 datasets and against 4 other state-of-the-art methods. We found that the performance of the classifier on these multi-label is not always improved, although it is when dealing with high-dimensional datasets. In the other cases. the benefit obtained is the reduction of the dimensionality.

We interpreted the results based on the mutual information between labels of each dataset. We found that in the case of datasets with many labels correlated between them, our proposal $ML - CMIM$ behaves better than the other methods; that is, our method performs better on datasets with interactions between labels, which is what our proposal is designed for. It also behaves well in high-dimensional datasets. Furthermore, $ML - CMIM$ does not set any limit to the number of interactions to search, as it can be interpreted from Equation 9. On the other hand, if the dataset used does not contain correlated labels, as it is the case of the *cal500* dataset, we found that more simple methods which do not try to catch labels interactions, such as CFS_{sum} , do behave better than the rest.

One last conclusion we obtained is that the use of adapted classifiers, which do not split apart the datasets in separate sets, helps feature selection methods to perform better, even more when they take into account interaction among labels.

Several issues were found which could be solved in future work:

First, it was difficult to find multi-label datasets which contain a great percentage of labels correlated at least in pairs. Thus, it would be of interest to search for more datasets of the kind or nature of CLEF, in order to prove the superior capacity of methods which search for higher-degree interactions among labels.

Second, we needed to use unsupervised discretization in order to compute mutual information to solve the Roure approximation in Equation 9. We plan to skip this discretization process by computing mutual information directly from numerical attributes. In the case of marginal expressions such as $I(X_{num}; Y_{cat})$, a proposal can be found in Pérez et al. (2006) to solve this without discretization. However, Equation 9 still needs to solve expressions of the type $I(X_{cat}; Y|S)$, where Y may be categorical or numerical, and S may be 1 or 2 numerical attributes. Thus this needs further investigation.

Finally, the Roure approximation itself presents several problems. The first, as mentioned above, is that it requires some conditional mutual information computations that are hard to perform without discretization, and its recursive expression may slow down the fast version of *CMIM*. Furthermore, this approximation is not score-equivalent and has no upper nor lower bound to its error. Thus, we think that the *CMIM* algorithm may benefit from a lighter approximation of mutual information conditioned to a large subset of variables, or from an approximation with its related error bound.

Acknowledgments

This work has been supported by the JCCM under regional project PEII-2014-049-P, and the MECD under national projects TIN2013-46638-C3-3-P and TIN2016-77902-C3-1-P.

References

- Alcobé, J. R. (2007), *Advances in Probabilistic Graphical Models*, Vol. 214/2007 of *Studies in Fuzziness and Soft Computing*, Springer Berlin / Heidelberg, chapter Learning Bayesian Networks with an Approximated MDL Score, pp. 215–234.
- Arias, J., Gámez, J. A., Nielsen, T. D. & Puerta, J. M. (2014), A pairwise class interaction framework for multilabel classification, *in* ‘Probabilistic Graphical Models - 7th European Workshop, PGM 2014, Utrecht, The Netherlands, September 17-19, 2014. Proceedings’, pp. 17–32.
- Battiti, R. (1994), ‘Using mutual information for selecting features in supervised neural net learning’, *IEEE Transactions on Neural Networks* **5**, 537–550.

- Bermejo, P., Gámez, J. A. & Puerta, J. M. (2010), Improving incremental wrapper-based feature subset selection by using re-ranking, *in* ‘Proceedings of the The Twenty Third International Conference on Industrial, Engineering & Industrial, Engineering & IEA-AIE’, LNCS/LNAI, pp. 580–589.
- Bermejo, P., Gmez, J. A. & Puerta, J. M. (2011), Improving incremental wrapper-based subset selection via replacement and early stopping., pp. 605–625.
- Bermejo, P., Lucas, M., Rodríguez-Montes, J. A., Tárraga, P. J., Lucas, J., Gámez, J. A. & Puerta, J. M. (2013), Single- and multi-label prediction of burden on families of schizophrenia patients., *in* ‘AIME’, pp. 115–124.
- Boutell, M. R., Luo, J., Shen, X. & Brown, C. M. (2004), ‘Learning multi-label scene classification’, *Pattern Recognition* **37**(9), 1757 – 1771.
- Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. (2012), ‘Conditional likelihood maximisation: A unifying framework for information theoretic feature selection’, *J. Mach. Learn. Res.* **13**, 27–66.
- Clare, A., Clare, A. & King, R. D. (2001), Knowledge discovery in multi-label phenotype data, *in* ‘In: Lecture Notes in Computer Science’, Springer, pp. 42–53.
- Diplaris, S., Tsoumakas, G., Mitkas, P. A. & Vlahavas, I. P. (2005), Protein classification with multiple algorithms, *in* P. Bozanis & E. N. Houstis, eds, ‘Proceedings of the 10th Panhellenic Conference in Informatics (PCI-05)’, Springer-Verlag, pp. 448–456.
- Doquire, G. & Verleysen, M. (2011), Feature selection for multi-label classification problems, *in* ‘Advances in Computational Intelligence - 11th International Work-Conference on Artificial Neural Networks, IWANN 2011, Torremolinos-Málaga, Spain, June 8-10, 2011, Proceedings, Part I’, pp. 9–16.
- Doquire, G. & Verleysen, M. (2013), ‘Mutual information-based feature selection for multilabel classification’, *Neurocomputing* **122**, 148–155.
- Elisseeff, A. & Weston, J. (2001*a*), A kernel method for multi-labelled classification, *in* ‘Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]’, pp. 681–687.
- Elisseeff, A. & Weston, J. (2001*b*), A kernel method for multi-labelled classification, *in* ‘In Advances in Neural Information Processing Systems 14’, MIT Press, pp. 681–687.

- Fernandes, J. A., Lozano, J. A., Inza, I. n., Irigoien, X., Pérez, A. & Rodríguez, J. D. (2013), ‘Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting’, *Environ. Model. Softw.* **40**, 245–254.
- Fleuret, F. (2004), ‘Fast binary feature selection with conditional mutual information’, *Journal of Machine Learning Research* **5**, 1531–1555.
- Forman, G. (2004), A pitfall and solution in multi-class feature selection for text classification, in ‘Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004’.
- Gómez-Verdejo, V., Verleysen, M. & Fleury, J. (2009), ‘Information-theoretic feature selection for functional data classification’, *Neurocomputing* **72**(16-18), 3580–3589.
- Hall, M. A. (1999), Correlation-based Feature Subset Selection for Machine Learning, PhD thesis, University of Waikato, Hamilton, New Zealand.
- Hall, M. A. (2000), Correlation-based feature selection for discrete and numeric class machine learning, in ‘Proceedings of the Seventeenth International Conference on Machine Learning’, pp. 359–366.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009), ‘The weka data mining software: an update’, *SIGKDD Exploration Newsletter* **11**, 10–18.
- Javed, K., Babri, H. A. & Saeed, M. (2010), ‘Feature selection based on class-dependent densities for high-dimensional binary data’, *IEEE Transactions on Knowledge and Data Engineering* **99**(PrePrints).
- Jungjit, S., Freitas, A. A., Michaelis, M. & Cinatl, J. (2012), A multi-label correlation-based feature selection method for the classification of neuroblastoma microarray data, in ‘Advances in Data Mining, 12th Industrial Conference, ICDM 2012’, pp. 149–157.
- Jungjit, S., Michaelis, M., Freitas, A. A. & Cinatl, J. (2014), Extending multi-label feature selection with KEGG pathway information for microarray data analysis, in ‘IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2014, Honolulu, HI, USA, May 21-24, 2014’, pp. 1–8.
- Kohavi, R. & John, G. H. (1997), ‘Wrappers for feature subset selection’, *Artificial Intelligence* **97**(1-2), 273–324.
- Lee, J. & Kim, D.-W. (2013), ‘Feature selection for multi-label classification using multivariate mutual information’, *Pattern Recogn. Lett.* **34**(3), 349–357.

- Lee, J. & Kim, D.-W. (2014), ‘Mutual information-based multi-label feature selection using interaction information’, *Expert Systems with Applications* **42**(4), 2013–2025.
- Lee, J. & Kim, D.-W. (2015), ‘Fast multi-label feature selection based on information-theoretic feature ranking’, *Pattern Recognition* **48**(9), 2761–2771.
- M.L., Z. & Z.H., Z. (2006), ‘Multi-label neural networks with applications to functional genomics and text categorization’, *IEEE Transactions on Knowledge and Data Engineering* **18**, 1338–1351.
- Peng, H., Long, F. & Ding, C. (2005), ‘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.
- Pérez, A., Larraaga, P. & Inza, I. (2006), ‘Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes’, *International Journal of Approximate Reasoning* **43**, 1–25.
- Read, J. (2008), A pruned problem transformation method for multi-label classification, in ‘Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008)’, pp. 143–150.
- Ruta, D. (2014), Robust method of sparse feature selection for multi-label classification with naive bayes, in ‘Federated Conference on Computer Science and Information Systems’, pp. 375–380.
- Sequeira, J., Gonçalves, T. & Quaresma, P. (2012), Semantic role labeling for portuguese - A preliminary approach -, in ‘Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings’, pp. 193–203.
- Sopolaôr, N., Monard, M. C. & Lee, H. D. (2012), A systematic review to identify feature selection publications in multi-labeled data, Technical Report 374, Institute of Mathematics and Computer Science, University of São Paulo.
- Spolaôr, N., Cherman, E. A., Monard, M. C. & Lee, H. D. (2013), ‘A comparison of multi-label feature selection methods using the problem transformation approach’, *Electr. Notes Theor. Comput. Sci.* **292**, 135–151.
- Spolaôr, N. & Tsoumakas, G. (2013), Evaluating feature selection methods for multi-label text classification, in ‘Proceedings of the first Workshop on Bio-Medical Semantic Indexing

and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) , Valencia, Spain, September 27th, 2013.’.

Tsoumakas, G. & Katakis, I. (2007), ‘Multi label classification: An overview’, *International Journal of Data Warehouse and Mining* **3**(3), 1–13.

Tsoumakas, G., Katakis, I. & Vlahavas, I. (2010), Mining multi-label data, *in* ‘In Data Mining and Knowledge Discovery Handbook’, pp. 667–685.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. & Vlahavas, I. (2011), ‘Mulan: A java library for multi-label learning’, *Journal of Machine Learning Research* **12**, 2411–2414.

Yu, L. & Liu, H. (2004), ‘Efficient feature selection via analysis of relevance and redundancy’, *Journal of Machine Learning Research* **5**, 1205–1224.

Zhang, M.-L., Peña, J. M. & Robles, V. (2009), ‘Feature selection for multi-label naive bayes classification’, *Inf. Sci.* **179**(19), 3218–3229.

Zhang, M.-L. & Zhou, Z.-H. (2007), ‘Ml-knn: A lazy learning approach to multi-label learning’, *Pattern Recogn.* **40**(7), 2038–2048.

Zhang, M. & Zhou, Z. (2014), ‘A review on multi-label learning algorithms’, *IEEE Trans. Knowl. Data Eng.* **26**(8), 1819–1837.

Table 1: Description of our evaluation corpus.

Datasets	 D 	N	 L
cal500	502	68	174
scene	2407	294	6
enron	1702	1001	53
medical	978	1449	44
slashdot	3782	1079	22
CLEF14-8labels	9500	360	8
CLEF14-18labels	9500	360	18

In	K : number of attributes to select; L : set of labels; N : original set of attributes; $S = \emptyset$
Out	subset S of selected attributes

```

//initiation for selecting the first attribute
1  for  $n = 1$  to  $N$  do
2     $x^+ \leftarrow N[n]$ 
3     $ps[n] \leftarrow \hat{I}_R(L; x^+)$ 
4     $m[n] \leftarrow 0$ 
    //conditioned search starts here
5  for  $k = 1$  to  $K$  do
6     $s^* \leftarrow 0$ 
7    for  $n = 1$  to  $N$  do
8      while  $ps[n] > s^*$  and  $m[n] < k - 1$  do
9         $m[n]++$ 
10        $ps[n] \leftarrow \min(ps[n], \hat{I}_{RC}(L; x^+ | S[m[n]]))$ 
11      if  $ps[n] > s^*$  then
12         $s^* \leftarrow ps[n]$ 
13       $S[k] \leftarrow x^*$ 

```

Figure 1: Pseudo-code for $ML - CMIM$. Please note that auxiliary variables ps , m and s^* used in the pseudo-code are needed for the fast implementation of $CMIM$. For further explanation about their use, the reader is referred to Fleuret (2004).

Figure 2: Histogram of the number of label pairs and their mutual information.

Figure 3: Hamming Loss performance of the 7 datasets with $ML - KNN$ classifier.

Figure 4: Example-Based Accuracy performance of the 7 datasets with $ML-KNN$ classifier.

Figure 5: Hamming Loss performance of the 7 datasets with $MLNB_{Basic}$ classifier.

Figure 6: Example-Based Accuracy performance of the 7 datasets with $MLNB_{Basic}$ classifier.

Figure 7: Mean value of both performance metrics for each method, using the adapted classifier $ML - KNN$.