Neurocomputing 122 (2013) 148–155

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Letters

Mutual information-based feature selection for multilabel classification

Gauthier Doquire *,[1], Michel Verleysen

Machine Learning Group - ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 4 October 2012
Received in revised form
30 April 2013
Accepted 11 June 2013
Communicated by F. Rossi
Available online 5 July 2013

Keywords:
Feature selection
Mutual information
Multilabel classification
Problem transformation

ABSTRACT

This paper introduces a new methodology to perform feature selection in multi-label classification problems. Unlike previous works based on the χ^2 statistics, the proposed approach uses the multivariate mutual information criterion combined with a problem transformation and a pruning strategy. This allows us to consider the possible dependencies between the class labels and between the features during the feature selection process. A way to automatically set the pruning parameter is also proposed, based on the permutation test combined with a resampling strategy. Experiments carried out on both artificial and real-world datasets show the interest of our approach over existing methods.

1. Introduction

Unlike traditional single-label problems, multi-label classification assumes that each data point from a learning set can belong simultaneously to several classes. The problem of multi-label classification is thus more general than the single-label one and has been extensively studied due to its interest in numerous domains. Those include classification of visual scenes [1], text categorization [2], classification of music into emotions [3] or protein function classification [4]. As a simple example, an article about the Kyoto protocol in text classification can be obviously associated with both *politics* and *ecology* categories.

Two distinct approaches are generally followed to perform multi-label classification. The first one is to adapt existing classification algorithms to handle multi-label problems. Among the most popular methods, one can cite AdaBoost [2], support vector machines [5], C4.5 [6] and K nearest neighbors [7].

Another popular approach is to transform the multi-label problem into one or several single-label problems, which can be addressed using existing methods. The most simple transformation method is the binary relevance (BR) whose idea is to learn a separate classification model for each label. Said otherwise, there are as many prediction models as the maximum number of possible labels; each model decides whether or not a point belongs to a specific class, independent of the result of the other classifiers. The final label set is obtained by combining the decisions of all classifiers. The BR approach does not take into account the possible dependence that could exist between the labels, since the decision for each class is made separately. To address this problem, a solution is to consider each unique combination of labels in a training set as a label for a single-label classifier [8]; such an approach is called label power set (LP). Since the number of classes created this way can potentially be huge (2^c, where c is the original number of labels), Read et al. [8] suggested to prune the problem by getting rid of classes represented by a too small number of instances. Points belonging to these classes can either be removed or can be assigned to another class. This methodology is called pruned problem transformation (PPT) [8].

Feature selection is known to be an important preprocessing task for many pattern recognition applications, including classification. Indeed, the presence of irrelevant and/or redundant features can harm the performances of classification algorithms. Moreover, learning with high-dimensional data is a hard task in practice [9]. Eventually, the interpretation of the prediction models and the understanding of the considered problem can also greatly benefit from feature selection.

Traditionally, feature selection algorithms are divided into three main categories. First, wrappers use the classification algorithm to select a subset of features maximizing the performances of this algorithm. They are thus expected to lead to good prediction performances but they also require building many prediction models, which can be very time-consuming in practice. On the other hand, filters are independent of any prediction algorithm;

* Corresponding author. Tel.: +32 10 47 81 33.
E-mail addresses: gauthier.doquire@uclouvain.be (G. Doquire), michel.verleysen@uclouvain.be (M. Verleysen).

[1] Gauthier Doquire is funded by a Belgian F.R.I.A. grant.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.06.035

they are rather based on a relevance criterion, the most frequently used ones being the correlation coefficient [10] and the mutual information (MI) [11]. Filters are fast and can be used prior to any prediction model. Eventually, embedded methods such as the Lasso [12] perform simultaneously feature selection and prediction through regularization.

Only a few works address the problem of feature selection for multi-label classification. The most popular approach is to use the BR transformation and to evaluate the relevance of each feature for each of the labels independently using a χ^2 statistics [3,13]. The scores corresponding to the different labels are then combined to get a global ranking of the features. This strategy has mainly been used for text categorization [13]. The LP transformation combined with the χ^2 statistics has also been used in music classification [3]; this approach has been shown to outperform the work in [13], mainly because it takes the dependence between labels into account.

These approaches suffer from two major drawbacks. First, the χ^2 statistics is originally designed for discrete variables. When the training set is made of continuous features (as in [3]), it is necessary to discretize the features before evaluating their relevance; the results of the feature selection are obviously dependent on the discretization step. A criterion directly able to deal with continuous variables is thus to be preferred. More importantly, the χ^2 statistics is a univariate relevance criterion, meaning that all the features are scored individually. This criterion is not able to consider the possible redundancy between features; in the same way, it is not able to detect joint relevant features. A multivariate criterion such as the mutual information, able to score subsets of features, does not suffer from these drawbacks. Related works include [14], where a graphical model is used to represent the relationships among the labels and the features; it is designed for discrete datasets only. In [15], the authors combine PCA-based feature extraction with a wrapper feature selection method implemented through a genetic algorithm. However, this approach to dimensionality reduction is essentially a feature extraction method, in the sense that the original features are transformed, by opposition to feature selection where a subset of the original features is kept.

This paper proposes a MI-based feature selection algorithm designed for multi-label problems. The idea is to first transform the problem using the PPT. A greedy search algorithm based on the MI criterion is then conducted to select the most relevant features. The dependence between both the labels and the features is thus considered by this approach. A way to automatically select the pruning parameter for PPT is also proposed. This work extends preliminary results presented in [16].

The paper is organized as follows. Section 2 recalls basic concepts about the MI criterion. The complete feature selection strategy is described in Section 3. Section 4 proposes a way to determine the value of the pruning parameter for PPT. Experiments on both artificial and real-world datasets are carried out in Section 5. Section 6 concludes the work.

2. Mutual information

This section briefly recalls basic definitions about the MI; next it presents a MI estimator specifically designed for classification problems.

2.1. Definitions

MI [17] is a symmetric measure of the amount of information that two variables X and Y contain about each other. MI has been widely used for feature selection since the seminal work by Battiti [11]. One of the main advantages of MI for feature selection is its ability to detect nonlinear relationships between variables, which is not the case, as an example, for the popular correlation coefficient. MI can also naturally be defined for groups of variables (or equivalently for multidimensional variables); this allows one to take the joint relevance and redundancy of features into account during the feature selection process.

The MI between X and Y is formally defined in terms of the probability density functions (PDF) of X, Y and (X,Y), respectively, denoted as p_X, p_Y and $p_{X,Y}$:

$$I(X;Y) = \iint p_{X,Y}(x,y)\log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\, dx\, dy. \tag{1}$$

The entropy of a random variable X is

$$H(X) = -\int p_X(x)\log\, p_X(x)\, dx; \tag{2}$$

the MI can be rewritten as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{3}$$

Since the entropy has a well-known interpretation in terms of the uncertainty of a random variable, (3) can be seen as the reduction of uncertainty about one variable once the other one is known. If X is a set of features and Y a class label vector, the last part of (3) gives a natural justification to the MI criterion for feature selection: maximizing $I(X;Y)$ with respect to a feature subset X is equivalent to searching for the subset of features reducing at most the uncertainty about the class label vector Y.

In practice none of the PDF's in (1) are known for real-world problems and MI has to be estimated from the data.

2.2. Estimation

In this paper, a MI estimator based on the nearest neighbors statistics and those introduced by Gomez et al. [18] is used. It is built from the Kozachenko–Leonenko estimator of entropy [19]:

$$\hat{H}(X) = -\psi(K) + \psi(N) + \log(c_d) + \frac{d}{N}\sum_{n=1}^{N}\log(\epsilon(n,K)) \tag{4}$$

where K is the number of nearest neighbors (a parameter of the estimator), N is the number of samples in X, d the dimensionality of these samples, c_d the volume of a unitary hypersphere of dimension d, $\epsilon(n,K)$ twice the Euclidean distance from the nth observation in X to its Kth nearest neighbor and ψ is the digamma function:

$$\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)} = \frac{d}{dk}\ln\Gamma(k), \quad \Gamma(k) = \int_0^\infty x^{k-1}e^{-x}\, dx. \tag{5}$$

The function ψ satisfies the following recursion: $\psi(x+1) = \psi(x) + 1/x$ and $\psi(1) = C$, $C = -0.5772\ldots$ being the Euler–Mascheroni constant.

Gomez et al. used Eq. (4) to derive a MI estimator specific to classification problems [18] (i.e. for problems where Y is a discrete variable). In such problems, the probability distribution of the (discrete) class variable Y can be estimated as $p(y = y_l) = n_l/N$, with n_l the number of points whose class value is y_l. Rewriting the estimated MI in terms of entropies, it gives as

$$\hat{I}(X;Y) = H(X) - H(X|Y) = H(X) - \sum_{l=1}^{L}\underbrace{\hat{p}(y=y_l)}_{n_l/N}H(X|Y=y_l), \tag{6}$$

where L is the total number of classes. This last equation indicates that only an estimation of $H(X)$ and $H(X|Y=y_l)$ is eventually needed to estimate $I(X;Y)$. More precisely, estimating $H(X|Y=y_l)$ is equivalent to estimating $H(X)$ using only those points whose class label is y_l. Plugging the entropy estimator (4) into Eq. (6), it is possible to derive

the following MI estimator:

$$\hat{I}(X;Y) = \psi(N) - \frac{1}{N}\sum_{l=1}^{L} n_l \psi(n_l) + \frac{d}{N}\left[\sum_{n=1}^{N} \log(\epsilon(n,K)) - \sum_{l=1}^{L}\sum_{n\in y_l} \log(\epsilon_l(n,K))\right]. \quad (7)$$

In (7), $\epsilon_l(n,K)$ has the same meaning as $\epsilon(n,K)$ but the set of possible neighbors for the nth observation is limited to the points whose class label is y_l.

The MI estimator (7) has the major advantage that it does not require the direct estimation of any PDF, which is a hard task when dealing with high-dimensional data, because of the *curse of dimensionality*. Indeed, histograms and kernel density estimators, which are traditionally used for MI estimation, suffer dramatically from an increase of dimensionality and are not likely to work well in high-dimensional spaces. By avoiding unreliable and imprecise PDF estimations, the estimator (7) is expected to be much less sensitive to the dimension of the data; it thus appears to be a reasonable choice to achieve multivariate MI estimation. Nearest neighbors-based MI estimators have already been used successfully for feature selection [18,20].

3. Feature selection algorithm

This section presents the complete methodology to perform feature selection for multi-label classification problems. Let us consider a dataset $D \in \Re^{N\times f}$, where f is the original number of features. Let us also consider, associated with D, a multi-label output vector $O \in \{0,1\}^{N\times c}$, where c is the number of possible labels. Each column of O thus codes for the samples in D to belong to one particular class.

The first step consists in transforming the problem using the PPT approach. Every unique combination of class labels in O is thus considered as a single class label. The points belonging to classes containing less than p samples are then removed from the dataset. The transformed dataset and output vector are called D_t and O_t, respectively. D_t contains $N_t \leq N$ samples and is described by f features $f^i, i = 1 \ldots f$.

Another possibility is to keep the points whose class has less than p samples, but to duplicate them and to assign each copy a new label, chosen from the labels present in the dataset after the pruning [8]. This is however not acceptable when working with the nearest neighbors-based MI estimator (7). Indeed, the presence of points having the exact same feature values would make the determination of the exact Kth nearest neighbor of some points impossible and would therefore harm the MI estimation.

In the proposed methodology, the benefits of the problem transformation are two-fold. First, it simplifies the problem by limiting the possible number of classes. PPT thus prevents the MI estimator from being hurt by the presence of many rare classes and, in this sense, prevents from overfitting. Then, PPT ensures that each class has a minimum number of p samples. This is of major importance since the MI estimator (7) requires the distance between each point in D and its Kth nearest neighbor in the same class. Controlling the pruning parameter p allows us to ensure that condition $K < p$ is fulfilled. This would not be the case for the BR transformation for which we have no guarantee about the minimal cardinality of the classes.

Once D_t and O_t have been obtained, the feature selection procedure can be used. In this paper, the MI criterion is combined with a forward greedy search strategy. The procedure starts with an empty set of features. The feature from D_t whose individual MI with the output O_t is the highest is first selected. Then, sequentially, the (still unselected) feature whose addition to the subset of selected features leads to the subset having the highest MI with the output is selected. The procedure is stopped after a predetermined number of features has been selected or when another stopping criterion is met. Once a feature has been selected, this choice is never questioned again, hence the name forward. Obviously, other search procedures can be thought of, including the backward elimination which starts with all the features and recursively eliminates them one at a time.

At each step of the forward algorithm, a new feature is added to the feature subset whose dimension is consequently increased. This underlies the need for a MI estimator able to deal with high-dimensional data. Because the MI is estimated between a subset of features and the output, the above strategy is able to consider the possible interaction or redundancy between features, which is a great advantage in a feature selection context. Of course the computational cost of the proposed multivariate procedure is much higher than one of a simple ranking method ($O(f^2)$ multivariate MI estimations vs. f univariate MI estimations).

4. Determination of the pruning parameter

This section proposes a way to automatically set the pruning parameter p, i.e. the minimum number of samples belonging to a class after the transformation procedure. Intuitively, the objective is to find a compromise between two opposite requirements. First, it is important to keep as much as possible the original information carried by the data, by choosing a not too high value for p. Indeed, a too aggressive pruning would lead to the removal of many samples and would make the transformed dataset useless, as it would not be representative anymore of the original dataset; only a few classes would actually still be represented.

On the other hand, feature selection or classification algorithms will not be able to perform well in the presence of many different classes containing a very small number of samples. Such small classes represent extremely rare situations which are not relevant for the considered problem and lead to overfitting.

Since we are interested in MI-based feature selection, it is natural to use a feature selection process to determine the value of p. Regarding the above considerations, a good value of p should be such that the MI is effectively able to use as much relevant information as possible to determine the important features, without being harmed by too rare class labels. We are thus interested in values of p for which the MI criterion is the most able to discriminate between relevant and irrelevant features. As the values of the MI are not bounded to a known interval (contrarily to the absolute correlation coefficient always lying in $[0,1]$), one possibility is to compare the distribution of $I(f^i;O_t)$ with the distribution $I(U;O_t)$, where U denotes a useless feature [21]. The value of p according to which these distributions are best separated can be chosen. Intuitively, if there are too many small classes for the MI estimator to return relevant values, there will be no significant difference between the MI estimated from f^i and from U. The same conclusion will apply if only a small number of classes are considered, since the dataset would then contain a very limited amount of information.

In practice, the distribution of $I(f^i;O_t)$ is not known. One solution is therefore to use a permutation test, combined with a k-fold procedure to estimate the mean and the variance of the MI estimator. Based on the work in [21], the idea is the following. For a feature f^i, assumed to be relevant for the classification task, the distribution of $I(f^i;O_t)$ and $I(f^{i,\pi};O_t)$ is built, where $f^{i,\pi}$ denotes a randomly permuted version of f^i. Because of the permutation, $f^{i,\pi}$ can be assumed to be independent from O_t and thus irrelevant for the classification task. To build the necessary distributions, 20 estimations of $I(f^i;O_t)$ and $I(f^{i,\pi};O_t)$ are performed on non-overlapping subsets of the dataset, using a 20-fold cross-validation scheme.

Eventually the value of p can be chosen as the one best separating the two distributions in the sense of a Student measure

$$t_x^i = \frac{\mu_x - \mu_x^\pi}{\sqrt{\sigma_x^2 + (\sigma_x^\pi)^2}}. \quad (8)$$

In Eq. (8), μ_x and σ_x^2 are, respectively, the estimated mean and variance of the distribution of $I(f^i; O_t)$ obtained on the training set transformed with a pruning parameter p equal to x; μ_x^π and $(\sigma_x^\pi)^2$ are the corresponding quantities for the distribution of $I(f^{i,\pi}; O_t)$.

Once t_x^i is obtained for every feature f^i and every parameter value x in a chosen range, the best value for p can be selected as the one corresponding to the highest value of t_x^i among all the features. This ensures that no useless feature is taken into account to choose the value of p. Considering useless features would indeed make no sense, if they can be considered as independent from O_t. Of course, one could also decide to consider the m highest values of t_x^i, or the features for which t_x^i is above a fixed threshold. This would however require the introduction (and possibly the tuning) of an additional parameter.

It is worth noting that parameter K of the MI estimator could as well be set according to the same methodology. To do so, t_x^i should simply be made dependent to K, to get $t_x^i(K)$ (with $K < p$). The maximum value of $t_x^i(K)$ over all features then gives the best couple of parameters for (p,K). Of course, the computational cost of the procedure is multiplied by K when compared to the case where only p has to be determined. In this paper, the value of K is set arbitrarily, as preliminary experiments have shown that it does not influence significantly the feature selection process when chosen in a reasonable range.

5. Experimental results

This section presents experimental results illustrating the shortcomings of existing multi-label feature selection algorithms and showing the interest of the proposed approach. Experiments are carried out on both artificial and real-world datasets. The proposed methodology is compared to the one introduced in [3], detailed in Section 1 and later denoted as χ^2. It is also compared to the use of the univariate MI to rank the features, without taking any joint redundancy or relevance into account. The discretization needed for the χ^2 approach follows [22]. Parameter K is set to 4 in the MI estimator, while the value of the pruning parameter p is chosen between 5 and 20. As it is needed that $p > K$, we have chosen a relatively small value for K. Moreover, we set a maximum value of 20 for p in order to limit the amount of removed samples.

5.1. Artificial datasets

Two artificial multi-label datasets whose characteristics are given in Table 1 are considered. For the first one, ten features $(f^1 \ldots f^{10})$ are drawn from a uniform distribution on the $[0, 1]$ interval. Five supplementary features are then constructed: $f^{11} = (f^1 - f^2)/2$, $f^{12} = (f^1 + f^2)/2$, $f^{13} = f^3 + 0.1$, $f^{14} = f^4 - 0.2$ and $f^{15} = 2 \times f^5$.

Table 1
Characteristics of the artificial datasets.

	Samples	Attributes	Labels	Classes
Problem 1 (9)	1000	15	4	8
Problem 2 (10)	1000	8	4	8

The multi-label output $O = [O^1 \ldots O^{c=4}]$ is then built as follows:

$$\begin{cases} O^1 = 1 & \text{if } f^1 > f^2 \\ O^2 = 1 & \text{if } f^4 > f^3 \\ O^3 = 1 & \text{if } O^1 + O^2 = 1 \\ O^4 = 1 & \text{if } f^5 > 0.8 \\ O^i = 0 & \text{otherwise } (i = 1 \ldots 4). \end{cases} \quad (9)$$

Obviously, only features f^{11} (or f^1 and f^2), f^3 (or f^{13}), f^4 (or f^{14}) and f^5 (or f^{15}) are needed to entirely determine the output.

20 artificial datasets of sample size 1000 have been randomly generated. Using the MI-based algorithm, feature f^{11} is always selected first. The next best ranked features are then f^5 or f^{15}, f^3 or f^{13} and f^4 or f^{14}. Thus only relevant and non-redundant features are always ranked in the top four positions, as could be expected from an efficient feature selection algorithm. When the χ^2−based method is used, the first three best ranked features are always f^5, f^{15} and f^{11} in a random order. Then f^1 or f^2 is ranked in the 4th position in half of the experiments (and are redundant with f^{11}). From one experiment to another, 6 to 8 features have to be considered to select the 4 necessary ones. The irrelevant features are, however, always ranked in the last places. This simple example clearly shows the interest of considering the redundancy when looking for small subsets of relevant features. Indeed, the proposed algorithm leads to features subsets of up to half the size of those returned by the χ^2−based strategy for an identical quantity of information. The results obtained with the univariate MI are extremely similar to the ones obtained using the χ^2 approach

The second artificial dataset consists in 8 features $(f^1 \ldots f^8)$ randomly drawn from a uniform distribution on the $[0, 1]$ interval. The output vector is built as follows:

$$\begin{cases} O^1 = 1 & \text{if } (f^1 > 0.5 \text{ and } f^2 > 0.5) \text{ or if } (f^1 < 0.5 \text{ and } f^2 < 0.5) \\ O^2 = 1 & \text{if } (f^3 > 0.5 \text{ and } f^4 > 0.5) \text{ or if } (f^3 < 0.5 \text{ and } f^4 < 0.5) \\ O^3 = 1 & \text{if } (f^1 > 0.5 \text{ and } f^4 > 0.5) \text{ or if } (f^1 < 0.5 \text{ and } f^4 < 0.5) \\ O^4 = 1 & \text{if } (f^2 > 0.5 \text{ and } f^3 > 0.5) \text{ or if } (f^2 < 0.5 \text{ and } f^3 < 0.5) \\ O^i = 0 & \text{otherwise } (i = 1 \ldots 4). \end{cases} \quad (10)$$

Hence, only features f^1–f^4 are relevant. Moreover, these features are relevant only if considered in pairs. Said otherwise, f^1 alone carries no information about O while f^1 and f^2 together do; they completely determine O^1. The same observation is true for the (f^1, f^4), (f^2, f^3) and (f^3, f^4) couples of features.

Again, 20 datasets of sample size 1000 are randomly generated. For each of the experiments, the χ^2−based approach returns a zero score for every feature, meaning that it considers 8 features as equally and totally irrelevant. Similarly, the univariate Mi criterion gives a very low score to each feature. When the proposed MI-based forward feature selection approach is considered, things are quite different. The first (few) feature(s) are of course chosen at random and can thus be irrelevant. However, as soon as one of the relevant features is chosen, the three other ones are always consecutively selected. The multivariate MI criterion is thus able to detect relevant subsets of features.

Moreover, we also used a backward search procedure using the multivariate MI criterion on the same 20 artificial datasets; for each run of the experiment, the relevant features were always eliminated last, showing the interest and the flexibility of the proposed methodology. As a ranking method which does not consider any interaction between features, the χ^2 and the univariate MI-based method are obviously never able to detect any jointly relevant or redundant features.

Table 2
Characteristics of the real-world datasets.

	Samples	Attributes	Labels	Classes
Yeast	2417	103	14	198
Scene	2407	294	6	15
Emotions	593	72	6	37

5.2. Real-world datasets

To further demonstrate the interest of the proposed approach for feature selection, this section shows experiments carried out on three real-world multi-label datasets.

5.2.1. Datasets

The first dataset is called Yeast. The objective is to associate genes with a set of functional classes. The dataset has been preprocessed by Elisseeff and Weston, as detailed in [5], in order to consider only the known structure of the functional classes. The dataset has 103 features and 14 distinct labels. The sample size is 1500 for the training set and 917 for the test set. Note that an interesting reference for the use of MI in gene selection is [23], where the authors are concerned with single label classification.

The Scene dataset [1] is concerned with the semantic classification of pictures. There are 294 features and 6 labels. The sample size is 1211 for the training set and 1196 for the test set.

The last dataset is named Emotions. The goal is to classify pieces of music according to the kind of emotions they raise. The number of features is 72 and 6 labels are possible. There are 391 samples in the training set and 202 in the test set. The characteristics of the three datasets are summarized in Table 2. The proposed divisions of the samples into the training and test sets are the ones suggested on the website of the Mulan project, where the three datasets can be downloaded in ARFF format [25].

5.2.2. Performance criteria

For real-world datasets, relevant features are not known in advance as they were for artificially built datasets. The quality of feature selection algorithms cannot thus be directly evaluated but can be measured by the performances of a classification model using the selected features. Four popular multi-label performance criteria are considered in this work. Let us call M the number of points in the test set, O_i $(i=1\ldots M)$ the set of true class labels for sample i, $\hat{O}_i$ the set of class labels predicted by a multi-label classifier h for sample i and $\hat{N}_i$ the set of non-predicted labels. Let us also define, for sample i, $r_i(o)$ as the position of label o in the predicted ranking.

The Hamming loss is defined as

$$HL(h,M)=\frac{1}{|M|}\sum_{i=1}^{|M|}\frac{1}{c}|O_i\Delta\hat{O}_i|, \tag{11}$$

where Δ is the symmetric difference between two sets, i.e. the difference between the union and the intersection of the two sets. Again, c is the maximum number of possible labels. The Hamming loss counts the number of times a label not associated to a sample is predicted or a label associated to a sample is not predicted. Of course, the smaller the Hamming loss, the better the performances of the model.

The second criterion is the accuracy defined as

$$Accuracy(h,M)=\frac{1}{|M|}\sum_{i=1}^{|M|}\frac{|O_i\cap\hat{O}_i|}{|O_i\cup\hat{O}_i|}. \tag{12}$$

It is important to note that all samples in the dataset are assumed to belong to at least one class. If it was not the case, the accuracy as defined above (12) could be infinite. Of course, the higher the accuracy, the better the classification performances.

The ranking loss is defined as

$$RL(h,M)=\frac{1}{|M|}\sum_{i=1}^{|M|}\frac{|(o,o')\in\hat{O}_i\times\hat{N}_i|r_i(o)>r_i(o')|}{|\hat{O}_i\quad||\quad\hat{N}_i|} \tag{13}$$

and corresponds to the mean proportion of pairs of labels which are not correctly ordered.

Eventually, coverage is defined as

$$Co(h,M)=\frac{1}{|M|}\sum_{i=1}^{|M|}\max_{o\in O_i}r_i(o)-1. \tag{14}$$

It corresponds to the average number of predicted labels required to cover the complete set of true labels. The lower the ranking loss and the coverage, the better the performances of the model.

The four presented criteria are very popular in multi-label classification and are more detailed in [24], where the interested reader will find a very general and complete survey of multi-label classification.

5.3. Results and discussions

To compare the feature selection algorithms, ML-KNN, a K nearest neighbors-based multi-label classifier introduced in [7] has been used. The basic idea is to first identify the K nearest neighbors of the sample to be classified. The maximum a posteriori principle is then used to predict the label set. This algorithm has been chosen for both its simplicity and its high sensitivity to the presence of irrelevant features. Indeed, a K nearest neighbors algorithm gives the same weight to each feature and is not able to perform any kind of embedded feature selection. It is important to note that the feature transformation method has only been used to perform feature selection. After feature selection, the original dataset with all instances is considered for the classification step. This way, we address the same classification problem and use the same learning set as for the χ^2-based algorithm. Parameter K of the MI estimator is again set to 4, while the pruning parameter p is set as described in Section 4 and is chosen between 5 and 20. The values 8, 12 and 9 have been obtained for p for the Yeast, Scene and Emotions datasets, respectively; they confirm that the range of tested values was reasonable since they lie in the middle of it.

Figs. 1–3 show the accuracy, the Hamming loss, the ranking loss and the coverage of the ML-KNN classifier as a function of the number of selected features for the three datasets. The results have been obtained on the test set, independent of the training set. The proposed method (denoted as MI forward) is compared with the one in [3] and with a ranking of the features based on the individual MI between the features and the outptut (denoted MI).

The results confirm the superiority of the proposed methodology, already observed with the artificial datasets. First, the forward MI-based approach is the only one leading to improved classification performances for the three datasets and according to the four performance criteria, when compared to the case where all features are used. This is of course a very desirable quality for a feature selection algorithm, which demonstrates that the method is effectively able to detect relevant features. For the three datasets, it is also the only method that consistently achieves good performances. Indeed, the univariate MI fails for the Scene dataset, while the χ^2 approach leads to poor results on the Yeast dataset. Both univariate methods are comparable on the Emotions dataset. The performances of the two MI-based approaches are similar for the Yeast dataset, possibly because the relevant features in this dataset are not redundant. On the Scene dataset, the univariate MI fails while the multivariate forward procedure

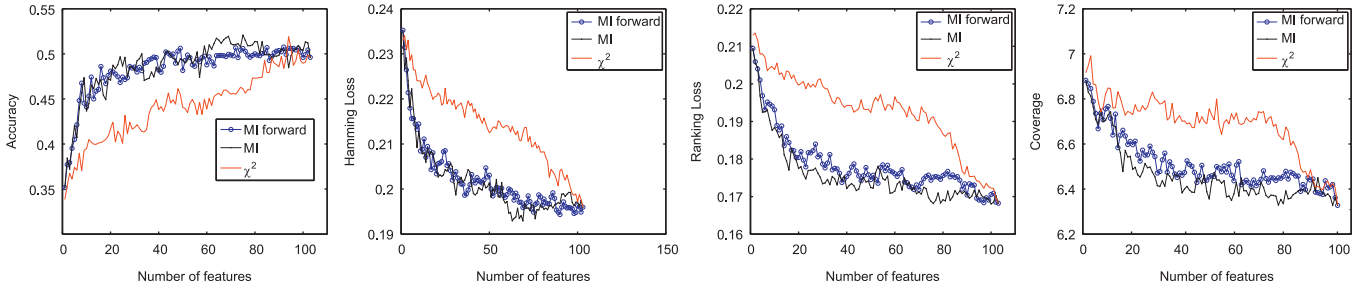

Fig. 1. Four quality criteria of the K nearest neighbors classifier as a function of the number of selected features for the Yeast dataset.

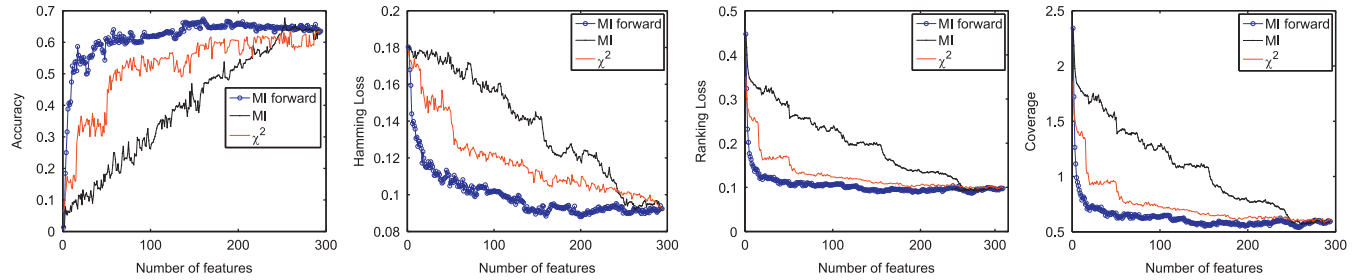

Fig. 2. Four quality criteria of the K nearest neighbors classifier as a function of the number of selected features for the Scene dataset.

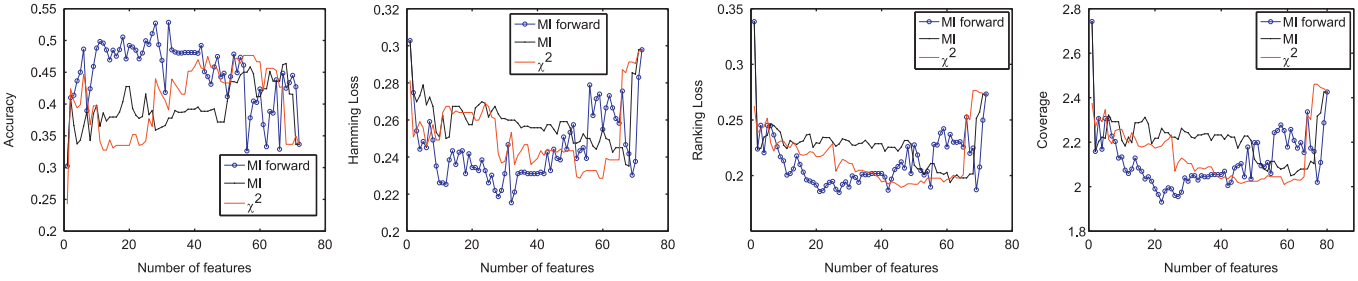

Fig. 3. Four quality criteria of the K nearest neighbors classifier as a function of the number of selected features for the Emotions dataset.

actually leads to good results. This indicates that the MI criterion has interest by itself but that univariate procedures can get stuck in selecting huge groups of relevant but redundant features. This can obviously degrade the classification performances. Considering multivariate procedures is thus extremely important in practice.

The determination of the pruning parameter is also an important aspect of the proposed feature selection algorithm. It is thus necessary to check in practice that it leads to good feature selection performances. Since no other solution currently exists in the literature, we compare the results obtained by the proposed method with the results obtained using other potential pruning parameter values. More precisely, Figs. 4 and 5 show, respectively, for the Yeast and Emotions datasets, the mean accuracy and Hamming loss obtained using all pruning parameter values between 5 and 15 but the one obtained with the proposed criterion. Those classification performances are compared with the ones resulting from the proposed approach. As can be observed, choosing a pruning parameter value using the suggested permutation test leads to better classification performances than what is obtained on average using reasonable values. Indeed, in the four cases illustrated in Figs. 4 and 5, the proposed criterion leads to the global best performances (i.e. smallest Hamming loss or highest accuracy). Even if the differences in performances are not huge, this indicates the interest of the suggested methodology compared to the random choice of what is believed to be a good value.

6. Conclusion

This paper addresses the feature selection problem in the context of multi-label classification. The proposed approach first suggests transforming the problem using the PPT into a single-label problem. The mutual information criterion is then combined with a greedy search strategy to select a relevant set of features. The mutual information is estimated through a nearest neighbor based estimator that is robust in high-dimensional spaces. The interest of the chosen transformation is twofold. First, it leads to a simplified version of the problem where too rare class labels are not taken into account. Then, it ensures that the mutual information estimator is able to work correctly, by being able to find a sufficient number of nearest neighbors of each class for each sample of the training set. While being quite straigthforward, the PPT approach requires the determination of parameter p (the minimum number of samples per class), for which an adequate value cannot be easily guessed in practice.

Consequently, a sound criterion to determine a good value of the pruning parameter for the problem transformation is also presented. The idea is to choose the value for which the distribution of the mutual information between a relevant feature and the output is best separated from the distribution of the mutual information between the same output and an irrelevant feature. To this end, a resampling strategy combined with a k-fold cross-validation procedure has been considered.

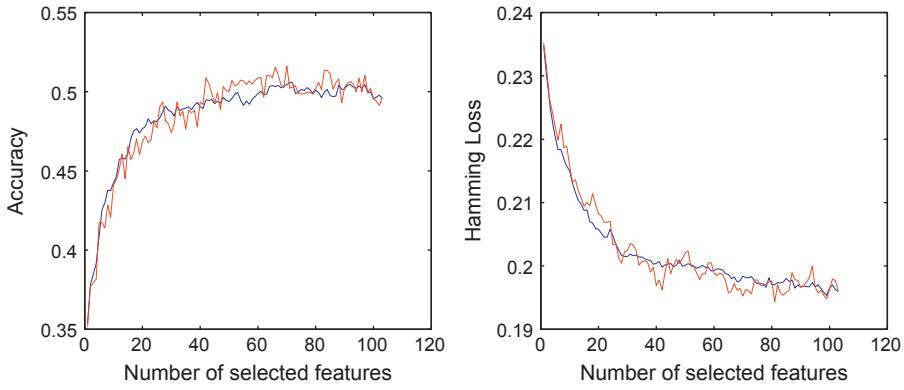

Fig. 4. Accuracy and Hamming loss obtained using the proposed permutation test (red) and average accuracy and Hamming loss obtained using all pruning parameter values between 5 and 15 but the one given by the permutation test (blue) on the Yeast dataset. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

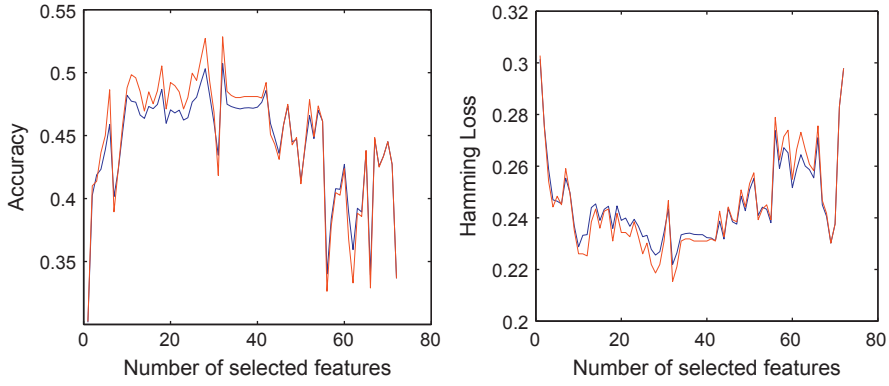

Fig. 5. Accuracy and Hamming loss obtained using the proposed permutation test (red) and average accuracy and Hamming loss obtained using all pruning parameter values between 5 and 15 but the one given by the permutation test (blue) on the Emotions dataset. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

The main limitation of this feature selection algorithm concerns the problem transformation. If each combination of labels only appears a limited number of times in the training set, the cardinality of each class after the problem transformation will be low (and a small value for parameter p will have to be used). This could lead to poor MI estimation and low learning performances.

Fortunately, the advantages of the proposed methodology over ranking methods based on the MI or the χ^2 are numerous and largely compensate for this drawback. First, the proposed procedure is multivariate, which makes it possible to take into account a possible joint relevance or a joint redundancy between the features regarding the output to predict. Then, it does not require the discretization of the continuous features, which can greatly harm the selection process. Eventually, the same mutual information relevance criterion can be combined with other search procedures if needed. For instance, the backward strategy has shown its interest for problems where individual features carry a low amount of information about the class labels. However, this procedure starts with all the features and the MI estimation can become unreliable if the original dimension of the dataset is too high. Experimental results clearly illustrate how the mentioned advantages actually translate into better classification performances in practice.

References

[1] M. Boutell, J. Luo, X. Shen, C. Brown, Learning multi-label scene classification, Pattern Recogn. 37 (2004) 1757–1771.
[2] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2000) 135–168.
[3] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, Multi-label classification of music into emotions, in: 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, 2008, pp. 325–330.
[4] S. Diplaris, G. Tsoumakas, P. Mitkas, I. Vlahavas, Protein classification with multiple algorithms, in: 10th Panhellenic Conference On Informatics (PCI 2005), Lecture Notes in Computer Science, vol. 3746, Springer, Heidelberg, 2005, pp. 448–459.
[5] A. Elisseeff, J. Weston, A Kernel method for multi-labelled classification, in: Advances in Neural Information Proceesing Systems, vol. 14, 2001, pp 681–687.
[6] A. Clare, R. King, Knowledge discovery in multi-label phenotype data, in: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001), Freiburg, Germany, 2001, pp. 42–53.
[7] M.-L. Zhang, Z.-H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recogn. 40 (2007) 2038–2048.
[8] J. Read, A pruned problem transformation mathod for multi-label classification, in: New Zealand Computer Science Research Student Conference (NZCSRS 2008), 2008, pp. 143–150.
[9] M. Verleysen, Learning high-dimensional data, Limitations and Future Trends in Neural Computation, NATO Science Series 186 (2003) 141–162.
[10] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the International Conference on Machine Learning, Stanford University, CA, 2000, pp. 359–366.
[11] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (1994) 537–550.
[12] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1) (1996) 267–288.
[13] W. Chen, J. Yan, B. Zhang, Z. Chen, Q. Yang, Document transformation for multi-label feature selection in text categorization, in: Proceedings of 7th IEEE International Conference on Data Mining, Los Alamitos, CA, USA, 2007, pp. 451–456.
[14] G. Lastra, O. Luaces, J.R. Quevedo, A. Bahamonde, Graphical feature selection for multilabel classification tasks, in: Proceedings of 10th International Conference on Advances in Intelligent Data Analysis, Porto, Portugal, 2011, pp. 246–257.
[15] M.-L. Zhang, J.M. Peña, V. Robles, Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19) (2009) 3218–3229.

[16] G. Doquire, M. Verleysen, Feature selection for multi-label classification problems. In: Proceedings of the International workshop on Artificial Neural Networks (IWANN 2011), Lecture Notes in Computer Science, vol. 6691, 2011, pp. 9–16.
[17] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423 623–656.
[18] V. Gomez-Verdejo, M. Verleysen, J. Fleury, Information-theoretic feature selection for functional data classification, Neurocomputing 72 (2009) 3580–3589.
[19] L.F. Kozachenko, N. Leonenko, Sample estimate of the entropy of a random vector, Probl. Inf. Transm. 23 (1987) 95–101.
[20] N. Benoudjit, D. François, M. Meurens, M. Verleysen, Spectrophotometric variable selection by mutual information, Chemometr. Intell. Lab. 74 (2004) 243–251.
[21] D. Francois, F. Rossi, V. Wertz, M. Verleysen, Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70 (2007) 1276–1288.
[22] H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, in: Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, 1995, pp. 388–391.
[23] P.E. Meyer, C. Schretter, G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, IEEE J. Sel. Top. Signal Process. 2 (2008) 261–274.
[24] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining Multi-label Data, 2010, pp. 667–685 (Chapter 34).
[25] The Mulan project: ⟨http://mulan.sourceforge.net/datasets.html⟩.

Gauthier Doquire was born in 1987 in Belgium. He received the M.S. in Applied Mathematics from the Université catholique de Louvain (Belgium) in 2009. He is currently a Ph.D. student at the Machine Learning Group of the same university. His research interests include machine learning, feature selection and mutual information estimation.

Michel Verleysen was born in 1965 in Belgium. He received the M.S. and the Ph.D. degrees in Electrical Engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (E.P.F.L.), Switzerland in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris 1–Panthéon-Sorbonne in 2002–2004. He is now a Research Director of the Belgian Fonds National de la Recherche Scientifique (F.N.R.S.) and Lecturer at the Université catholique de Louvain. He is Editor-in-Chief of the Neural Processing Letters journal, Chairman of the Annual European Symposium on Artificial Neural Networks (ESANN) Conference, Associate Editor of the IEEE Transactions on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is the author or the co-author of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the co-author of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?", in French. His research interests include machine learning, artificial neural networks, self-organization, time-series fore-casting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.