Feature Selection for Multi-label Classification Problems

Gauthier Doquire[*] and Michel Verleysen

Université catholique de Louvain, Machine Learning Group - ICTEAM
Place du Levant, 3, 1348 Louvain-la-Neuve, Belgium
{gauthier.doquire,michel.verleysen}@uclouvain.be
http://www.ucl.ac.be/mlg

Abstract. This paper proposes the use of mutual information for feature selection in multi-label classification, a surprisingly almost not studied problem. A pruned problem transformation method is first applied, transforming the multi-label problem into a single-label one. A greedy feature selection procedure based on multidimensional mutual information is then conducted. Results on three databases clearly demonstrate the interest of the approach which allows one to sharply reduce the dimension of the problem and to enhance the performance of classifiers.

Keywords: Feature selection, Multi-Label, Problem Transformation, Mutual Information.

1 Introduction

Multi-label Classification is the task of assigning data points to a set of classes or categories which are not mutually exclusive, meaning that a point can belong simultaneously to different classes. This problem is thus more general than the traditional single-label classification which assumes each point belongs to exactly one category; it is therefore often encountered in practice. As an example, in text categorization problems, an article about the Kyoto Protocol can be labelled with both *politics* and *ecology* categories. In scene classification, a picture could as well belong to different classes such as *beach* and *mountain* [1]. Other domains for which multi-label classification has proved useful also include protein function classification [2] and classification of music into emotions [3].

Due to its importance, multi-label classification has been studied quite extensively since a few years, leading to the development of numerous classification algorithms. Some of them are extensions of existing single-label classification methods such as AdaBoost [4], support vector machines (SVM) [5] or K nearest neighbors [6] among others.

Another popular approach to multi-label classification consists in transforming the problem into one or more single-label classification tasks. State of the art algorithms such as SVM can then be used directly. The most popular transformation method is the binary relevance (BR) which consists in learning a different

[*] Gauthier Doquire is funded by a Belgian FRIA grant.

J. Cabestany, I. Rojas, and G. Joya (Eds.): IWANN 2011, Part I, LNCS 6691, pp. 9–16, 2011.

classifier for each label. In other words, the original problem is transformed into C two classes single-label classification problems, where C is the number of possible labels. The i^{th} $(i = 1 \ldots l)$ classifier decides whether or not a point belongs to the i^{th} class. The union of predicted labels for each point is the final output. One of the major drawbacks of BR is that it does not take into account the dependence which could exist between labels.

Label powerset (LP) is a different problem transformation method which does consider this dependence. It treats each unique set of labels in the training set as a possible class of a single-label classifier. The number of classes created this way being potentially huge, Read et al. [7] recently proposed to prune the problem, by considering only classes represented by a minimum number of data points. Points with a too rare label are either removed from the training set or are given a new label and kept. They called this approach pruned problem transformation (PPT). See [7] for details.

Surprisingly, feature selection for multi-label classification has not received much attention yet. Indeed, to the best of our knowledge, one of the few proposed approach is the one by Trohidis et al. [3] which consists in transforming the problem with the LP method, before using the χ^2 statistic to rank the features. However, feature selection is an important task in machine learning and pattern recognition, as it can improve the interpretability of the problems, together with performances and learning time of prediction algorithms [8].

This paper proposes to use mutual information (MI) to achieve feature selection in multi-label classification problems. More precisely, the problem is first transformed with the PPT method and a greedy forward search strategy is then conducted with multidimensional MI as the search criterion. This approach thus considers dependencies between labels as well as dependencies between features, which ranking approaches such as [3] do not.

The remaining of the paper is organized as follows. Section 2 briefly recalls some basic concepts about MI, and introduces the estimator used. The methodology is described in Section 3, and the interest of the approach is experimentaly shown in section 4. Section 5 concludes the work and gives some future work perspectives.

2 Mutual Information

2.1 Definitions

MI [9] is a measure of the quantity of information two variables contain about each other. It has been widely used for feature selection [10] mainly because of its ability to detect non-linear relationships between variables. This not the case, as an example, for the correlation coefficient. Moreover, MI is naturally defined for groups of variables, which allows one to take feature dependence into account during the feature selection process.

MI of a couple of random variables X and Y is formally defined in terms of the probability density functions (PDF) of X, Y and (X, Y), respectively denoted as f_X, f_Y and $f_{X,Y}$:

$$I(X;Y) = \int\int f_{X,Y}(\zeta_X, \zeta_Y) \log \frac{f_{X,Y}(\zeta_X, \zeta_Y)}{f_X(\zeta_X) f_Y(\zeta_Y)} \, d\zeta_X \, d\zeta_Y. \tag{1}$$

In practice none of the PDF's are known for real-world problems, and MI has to be estimated from the dataset.

2.2 Estimation

In this paper, an MI estimator introduced by Gomez et al. [11] is used. It is based on the Kozachenko-Leonenko estimator of entropy [12]:

$$\hat{H}(X) = -\psi(K) + \psi(N) + \log(c_d) + \frac{d}{N} \sum_{n=1}^{N} \log(\epsilon(n, K)) \tag{2}$$

where ψ is the digamma function, K the number of nearest neighbors considered, N the number of samples in X, d the dimensionality of these samples, c_d the volume of a unitary ball of dimension d and $\epsilon(n, K)$ twice the distance from the n^{th} observation in X to its K^{th} nearest neighbor. Throughout this paper, the metric used in the X space is the Euclidean distance.

Basing their developments on (2), Kraskov et al. previously proposed two estimators for regression problems. See [13] for details.

For classification problems, the probability distribution of the (discrete) class vector Y is $p(y = y_l) = n_l/N$, whith n_l the number of points whose class value is y_l. Rewriting MI in terms of entropies:

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

it is possible to derive the following estimator:

$$\begin{aligned}\hat{I}_{cat}(X;Y) = \psi(N) - \frac{1}{N} n_l \psi(n_l) + \\ \frac{d}{N}\left[\sum_{n=1}^{N} \log(\epsilon(n, K)) - \sum_{l=1}^{L} \sum_{n \in y_l} \log(\epsilon_l(n, K))\right]\end{aligned} \tag{4}$$

where L is the total number of classes. $\epsilon_l(n, K)$ has the same meaning as $\epsilon(n, K)$ but the set of possible neighbors for the n^{th} observation is limited to the points whose class label is y_l.

This estimator has the crucial advantage that it does not require directly the estimation of any PDF which is a particularly hard task when the dimension of the data increases, due to the so-called *curse of dimensionality.* This curse reflects the fact that the number of points needed to sample a space increases exponentially with the dimension of the space. Histograms or kernel density estimators [14] are thus not likely to work well in high dimensional spaces. Because it avoids such unreliable estimations, (4) is expected to be less sensitive to the dimension of the data; this family of estimators has already been used successfully for feature selection [11,15].

3 Methodology

This section describes the methodology followed to achieve feature selection. First, the multi-label problem is transformed using the PPT method defined above [7], and the data points with a class label encountered less than t times in the training set are discarded. The result of this transformation is thus a multi-class *single-label* classification problem. The pruning has a double interest here; it leads to a simplified version of the problem and ensures that all classes are represented by at least t points. This last observation is crucial since the MI estimator (4) requires the distance between each point and its K^{th} nearest neighbor from the same class. It is thus needed that $K < t$.

Once the problem has been transformed, traditional feature selection techniques can be used. In this paper, a greedy forward feature selection algorithm based on MI is employed. It begins with an empty set of features and first selects the feature whose MI with the class vector is the highest. Then, sequentially, the algorithm selects the feature not yet selected whose addition to the current subset of selected features leads to the set having the highest MI with the output. This choice is never questionned again, hence the name *forward*. Of course, other search strategies could also be considered such as backward elimination, which starts with the set of all features and recursively removes them one at a time.

The procedure can be ended when a predefined number of features have been chosen. Another strategy is to rank all the features and then to choose the optimal number to keep on a validation set.

4 Experiments and Results

This section begins by introducing the performance criterias considered since they differ from those used for single-label classification. The databases used in the experiments are then briefly described and the results are eventually shown.

4.1 Evaluation Criteria

Two very popular evaluation criterias are considered in this study: the Hamming loss and the accuracy. Let $|M|$ be the number of points in a test set M, Y_i, $i = 1 \dots |M|$, the sets of true class labels and $\hat{Y_i}$ the sets of labels predicted by a multi-label classifier h.

The Hamming loss is then defined as:

$$HL(h, M) = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{1}{|C|} |Y_i \,\Delta\, \hat{Y_i}| \tag{5}$$

where Δ denotes the the symmetric difference between two sets, i.e. the difference between the union and the intersection of the two sets. $|C|$ is the number of possible labels.

The accuracy is defined as:

$$Accuracy(h, M) = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}. \tag{6}$$

Of course, the smaller the Hamming loss and the higher the accuracy, the better the classifier's performances. It is important to note that all data points are assumed to belong to at least one class. If it was not the case, the accuracy as defined above (6) would be infinite.

4.2 Datasets

Three datasets are used for experiments in this paper.

The first one is the Yeast dataset. It is concerned with associating each gene a set of functional classes. For the sake of simplicity, the data have been preprocessed by Elisseeff and Weston [5] to consider only the known structure of the functionnal classes. Eventually, the sample sizes of the training set and the test set are 1500 and 917 respectively. There are 103 features and 14 possible labels. The Scene dataset is also considered [1]. The goal here is the semantic indexing of scenes. The number of samples is 1211 for the training set and 1196 for the test set. There are 294 features and 6 labels. The last dataset is called Emotions and is about the classification of music into emotions [3]. Among the 593 samples, 391 are used as the training set and the 202 other as the test set. The number of features and of labels is 72 and 6 respectively. The proposed training set/testing set splittings are the ones traditionally used in the multi label classification litterature.

The three datasets are available for download in ARFF format at the web page of the Mulan project[1] .

4.3 Experimental Results

The k nearest neighbors based multi-label classification algorithm introduced by Zhang and Zhou [6] is used to illustrate the interest of the proposed approach. To determine the set of labels of a new instance, the algorithm first identifies its k nearest neighbors. Then, based on their labels, the maximum a posteriori principle is used to predict the label set of the new instance. As suggested by the authors, the value of the parameter k has been set to 7.

The problem transformation is only used to achieve feature selection. Once the features have been ranked, the original multi-label problem is considered again with all the samples. The number of neighbors considered in the MI estimator (4) is $K = 4$ and the t parameter for the PPT is fixed to 6. These values have been

[1] http://mulan.sourceforge.net/datasets.html

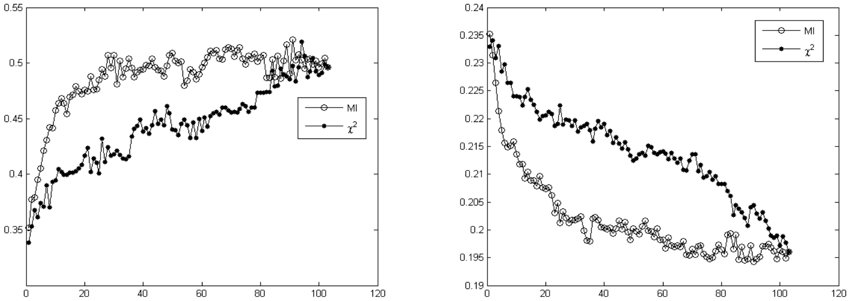

Fig. 1. Accuracy (left) and Hamming loss (right) of the k nearest neighbors classifier as a function of the number of selected features for the Emotions dataset

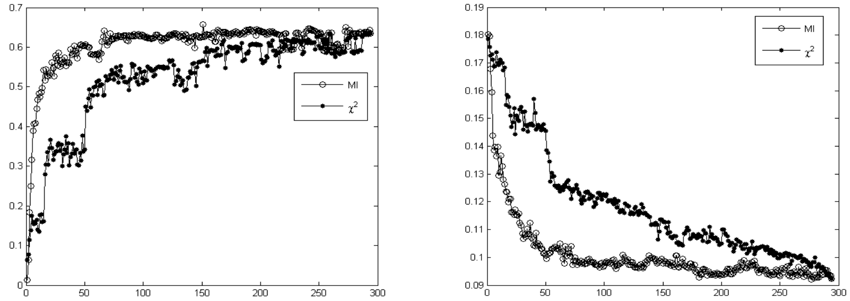

Fig. 2. Accuracy (left) and Hamming loss (right) of the k nearest neighbors classifier as a function of the number of selected features for the Emotions dataset

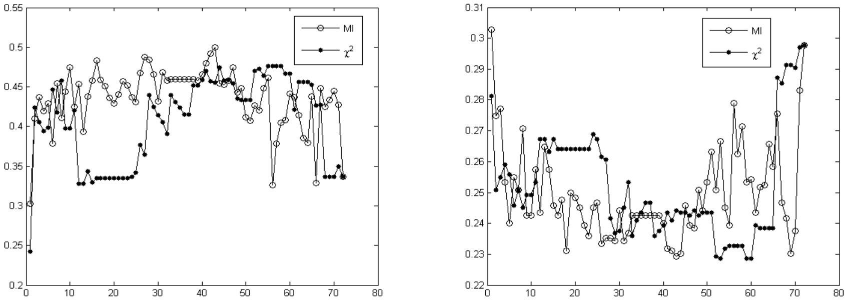

Fig. 3. Accuracy (left) and Hamming loss (right) of the k nearest neighbors classifier as a function of the number of selected features for the Emotions dataset

chosen as a compromise between the need to consider a sufficiently large number of neighbors in the MI estimator and the fact that in [7], better performances are usually obtained with a small value of t.

Figures 1 to 3 show the Hamming loss and the accuracy of the classifier on the test set of the three datasets with the MI based feature selection (denoted as MI). For comparison, the results obtained with the approach by Trohidis et al. [3] are also presented (and are denoted as χ^2).

The results clearly demonstrate the interest of the MI based approach and its advantage over the method based on the χ^2 statistic. Particularly, the proposed approach always leads to an increase in performance both for the Hamming loss and the accuracy compared with the case no feature selection is considered. This is not the case for the Hamming loss with the χ^2 based approach on the Yeast and Scene datasets and the differences between the two approach performances are particularly obvious for those two datasets. The results are much more comparable for the Emotions dataset for which both methods result in a large improvement of the classifier performances.

As already stated, the good behaviour of the proposed methodology can be explained by the use of a powerful criterion combined with an approach taking relations between features into account. Indeed, the forward selection procedure described above is expected to better handle redundancy between features than simple ranking methods do. This is fundamental since a feature, even with a high predictive power, is useless if it carries the same information about the output than another selected feature; it should therefore not be selected.

The same experiments have been carried out with a SVM classifier working directly on the transformed and pruned problem. The results obtained confirm those presented in this paper. However, due to space limitations, they are not presented here.

5 Conclusions

This paper is concerned with feature selection for multi-label classification, a problem which has up to now received little attention despite its great importance and the amount of work recently proposed on multi-label classification.

It is suggested to use multidimensional mutual information after the transformation of the multi-label problem to a single-label one through the pruned problem transformation method. To this end, a nearest neighbors based MI estimator is used, as it is believed to behave well when dealing with high-dimensional data. The estimator is combined with a simple greedy forward search strategy to achieve feature selection.

Results on three real-world datasets coming from different domains show the interest of this new approach compared with a strategy based on the χ^2 statistic in terms of the Hamming loss and the accuracy of a classifier.

Future work should include the study of the influence of the pruning parameter; it could be possible to tune this parameter to maximise the performances of the classifiers. A trade-off should then be found between the increase in performances and the computation load of validation procedures such as k-fold cross validation.

Besides the basic stopping criteria proposed in Section 3, much sophisticated strategies can be thought of. As an example, Damien et al. proposed a stopping criterion based on resampling and the permutation test [16]. The basic idea is to halt the procedure when the addition of a new feature does not improve significantly the MI between the selected features an the output. This approach could be applied to the problem considered in this paper.

References

1. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning Multi-Label Scene Classification. Pattern Recogn. 37, 1757–1771 (2004)
2. Diplaris, S., Tsoumakas, G., Mitkas, P., Vlahavas, I.: Protein Classification with Multiple Algorithms. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 448–459. Springer, Heidelberg (2005)
3. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-Label Classification of Music into Emotions. In: 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, pp. 325–330 (2008)
4. Schapire, R.E., Singer, Y.: Boostexter: A Boosting-Based System for Text categorization. Machine Learning 39, 135–168 (2000)
5. Elisseeff, A., Weston, J.: A Kernel method for Multi-Labelled Classification. Advances in Neural Information Proceesing Systems 14, 681–687 (2001)
6. Zhang, M.-L., Zhou, Z.-H.: ML-KNN: A Lazy Learning Approach to Multi-Label Learning. Pattern Recogn. 40, 2038–2048 (2007)
7. Read, J.: A Pruned Problem Transformation Mathod for Multi-label Classification. In: New Zealand Computer Science Research Student Conference (NZCSRS 2008), pp. 143–150 (2008)
8. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. J. Mach. Lear. Res. 3, 1157–1182 (2003)
9. Shannon, C.E.: A mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423, 623–656 (1948)
10. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE T. Neural. Networ. 5, 537–550 (1994)
11. Gomez-Verdejo, V., Verleysen, M., Fleury, J.: Information-Theoretic Feature Selection for Functional Data Classification. Neurocomputing 72, 3580–3589 (2009)
12. Kozachenko, L.F., Leonenko, N.: Sample Estimate of the Entropy of a Random Vector. Problems Inform. Transmission 23, 95–101 (1987)
13. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating Mutual Information. Phys. Rev. E 69, 066138 (2004)
14. Parzen, E.: On Estimation of a Probability Density Function and Mode. Ann. Math. Statist. 33, 1065–1076 (1962)
15. Benoudjit, N., François, D., Meurens, M., Verleysen, M.: Spectrophotometric Variable Selection by Mutual Information. Chemometr. Intell. Lab. 74, 243–251 (2004)
16. Francois, D., Rossi, F., Wertz, V., Verleysen, M.: Resampling Methods for Parameter-free and Robust Feature Selection with Mutual Information. Neurocomputing 70, 1276–1288 (2007)