ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Multi-target feature selection through output space clustering

Konstantinos Sechidis[1,2] and Eleftherios Spyromitros-Xioufis[1,3] and Ioannis Vlahavas[1] *

1-Department of Computer Science, Aristotle University, Thessaloniki 54124, Greece

2-School of Computer Science, University of Manchester, Manchester M13 9PL, UK

3-Expedia, Geneva

Abstract. A key challenge in information theoretic feature selection is to estimate mutual information expressions that capture three desirable terms: the relevancy of a feature with the output, the redundancy and the complementarity between groups of features. The challenge becomes more pronounced in multi-target problems, where the output space is multi-dimensional. Our work presents a generic algorithm that captures these three desirable terms and is suitable for the well-known multi-target prediction settings of multi-label/dimensional classification and multivariate regression. We achieve this by combining two ideas: deriving low-order information theoretic approximations for the input space and using clustering for deriving low-dimensional approximations of the output space.

1 Introduction

Many real world applications generate huge amounts of data that create various new challenges, such as learning from high dimensional inputs (features). One way of dealing with big dimensionality is to ignore the irrelevant and redundant features by using a **feature selection** (FS) algorithm [1]. At the same time more and more applications need to predict multiple outputs (targets), instead of a single output. Depending of the type of the output variables there are various categories of **multi-target** problems, such as multi-label classification, multi-dimensional classification, and multivariate regression, when the outputs are binary, categorical, and continuous respectively [2].

In this paper we will focus on deriving novel information theoretic feature selection methods for multi-target problems. To do so we need to estimate **mutual information** (MI) expressions from finite sample data sets. As the number of selected features grows due to high dimensionality of the input space, and as the number of targets is high due to high dimensionality of the output space, the estimated MI expressions become less reliable. To overcome this problem, low-order criteria have been suggested.

Sechidis et al. [3] introduced a framework for generating such low-order FS criteria for multi-target problems by iteratively maximising different composite likelihood expressions, which make various assumptions about the input and output space. By exploring how the different assumptions compare, the authors have found that the best trade-off appears to assume partial independence in the

*This research is implemented through the Operational Program "Human Resources Development, Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek national funds.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

feature and full independence in the target space (a method known as Single-JMI, details in Sec. 2). While the partial independence of the feature space has been proven to be useful in deriving FS criteria for single-label data [1], the full independence in the label space ignores the useful information that the possible dependencies between the targets can provide. Our work introduces a novel algorithm (Group-JMI, details in Sec. 3) that uses the principles of the Single-JMI criterion but at the same time takes into account target dependencies.

2 Background on information theoretic multi-target FS

Let us assume that we have a multi-target problem where we observe N samples $\{\mathbf{x}^n, \mathbf{y}^n\}_{n=1}^N$. The feature vector $\mathbf{x} = [x_1 \dots x_d]$ is a realisation of the joint random variable $\mathbf{X} = X_1 \dots X_d$, while the output vector is a realisation of $\mathbf{Y} = Y_1 \dots Y_m$. When the variables of the output space are binary, i.e. the alphabet $\mathcal{Y}$ is $\{0,1\}^m$, the problem is known as multi-label classification, when they are categorical as multi-dimensional classification, while when they are continuous, i.e. $\mathcal{Y}$ is $\mathbb{R}^m$, as multivariate regression [2].

The problem of FS can be phrased as selecting a subset of features $\mathbf{X}_\theta \subset \mathbf{X}$ that contain as much useful information for our problem as possible. In *filter* FS, we firstly rank the features according to a score measure and then select the ones with the highest score. The score of each feature should be independent of any classifier and any evaluation measure. For single-output problems, Brown et al. [1] introduced a framework for generating information theoretic FS criteria by phrasing a clearly specified optimisation problem; maximising the conditional likelihood. A greedy forward selection to optimise this objective is: at each step k select the feature $X_k \in \mathbf{X}_{\widetilde{\theta}}$ that maximises the following conditional mutual information (CMI): $J_{\text{CMI}}(X_k) = I(X_k; Y|\mathbf{X}_\theta)$, where $\mathbf{X}_\theta$ is the set of the (k-1) features already selected, $\mathbf{X}_{\widetilde{\theta}}$ the unselected ones and Y the single-output target variable. It can be shown that optimising this objective leads to a criterion that assigns a score to each feature that increases if the relevancy of the feature with the targets is high, the redundancy with the existing features is low, and the complementarity with the existing features is high [1].

As the number of selected features grows, the dimensionality of $\mathbf{X}_\theta$ also grows, making the estimates less reliable. To overcome this issue a number of methods have been proposed for deriving low-order criteria. A famous criterion that controls relevancy, redundancy and complementarity, providing a good trade-off between accuracy, stability and flexibility is the joint mutual information (JMI), with scoring function [1]: $J_{\text{JMI}}(X_k) = \sum\limits_{X_j \in \mathbf{X}_\theta} I(X_j X_k; Y)$.

Sechidis et al. [3] derived two versions of the JMI criterion suitable for multi-output problems. Their approach was based on the idea of expressing multi-label decomposition methods as composite likelihoods. The scoring functions for the two multi-output criteria are the following:

$$J_{\text{JMI}}^{\text{Joint}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} I(X_j X_k; \mathbf{Y}), \qquad J_{\text{JMI}}^{\text{Single}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} \sum_{Y_i \in \mathbf{Y}} I(X_j X_k; Y_i).$$

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

The superscripts denote the assumptions over the output space:

Joint-JMI does not make any assumptions and deals with the joint random variable $\mathbf{Y}$. This corresponds to the Label Powerset transformation in the multi-label literature. The main *limitation* of this method is that $\mathbf{Y}$ is high dimensional. For example, in multi-label problems we have up to $\min(N, 2^m)$ distinct labelsets [4], which makes it difficult to estimate MI expressions reliably.

Single-JMI deals with each variable $Y_i, i = 1 \dots m$, independently of the others. This corresponds to the Binary Relevance (BR) transformation in the multi-label literature. The main *limitation* of this method is that by making the full independence assumption it ignores possible useful information on how the targets interact with each other.

These two versions of the JMI criterion can be seen as the two extreme cases; assuming no independence at all (Joint-JMI) and assuming every outcome is independent from the rest (Single-JMI).

In their experimental study, Sechidis et al. [3] showed that Single-JMI, even if it assumes full independence between the targets, outperforms Joint-JMI, the variant that makes no assumptions about the targets. This is happening because the low-dimensional MI expressions in Single-JMI are estimated more reliably from small datasets than the high dimensional MI expressions in Joint-JMI. Next section introduces a novel algorithm that accounts for target dependencies and at the same time keeps the dimensionality of the MI expressions low.

3 Proposed methodology

The main idea behind our approach is to derive a novel representation of the output space $\widetilde{\mathbf{Y}} = \widetilde{Y}_1 \dots \widetilde{Y}_m$, where each variable $\widetilde{Y}_i$ captures the joint information of some group of target variables. After deriving this representation, we will use the following criterion, which we call **Group-JMI**:

$$J_{\mathrm{JMI}}^{\mathrm{Group}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} I(X_j X_k; \widetilde{Y}_i),$$

Group-JMI can be seen as the modification of Single-JMI criterion using $\widetilde{Y}_i$ instead of the initial targets Y_i. By doing this we keep estimating low dimensional MI expressions, but at the same time we take into account target dependencies; each $\widetilde{Y}_i$ captures the information that is shared in a group of target variables.

The main challenge is to derive the projected space $\widetilde{\mathbf{Y}}$ from the initial space $\mathbf{Y}$. Here, we solve this challenge using the following two-step, clustering-based strategy:

1st Step: Generate groups of target variables

In this step we create m groups of variables $\mathbf{Z}_1, \dots \mathbf{Z}_m$, where each group is a random subset of the targets, i.e. $\mathbf{Z}_i \subset \mathbf{Y} \ \forall \ i = 1 \dots m$. Each group is generated by sampling the set of target variables without replacement, but allowing overlap between the groups. Randomly sampling groups of targets has been extensively used for deriving learning algorithms but not for FS. A famous example is RAKEL [5], a state of the art method for learning from multi-label data.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Similarly to RAKEL, the number of targets in each group is controlled by a parameter that specifies the Proportion of Targets (PoT) randomly sampled to generate each group. Given, for example, a multi-target problem with $m = 20$ targets and PoT=0.30, 20 groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_{20}$ will be generated, each one consisting of 6 randomly selected target variables. Assuming binary targets the joint variable in each group may take up to $2^6 = 64$ distinct values, a dimensionality that prevents reliable density estimation unless a very large amount of data is available. To overcome this issue, we introduce a way to derive low-dimensional approximations in the following step.

2nd Step: Low-dimensional approximations via clustering

To derive low dimensional representations for each group, we will use the idea of clustering together examples with "similar" output vectors. In the most common case we assume the Number of Clusters (NoC) is known a priori. For each group $\mathbf{Z}_i$, we derive a novel categorical variable $\widetilde{Y}_i$, with the alphabet $\{1, ..., \text{NoC}\}$, that describes the cluster indices of each observation:

$$\widetilde{y}_i^n = \text{Clustering}(\mathbf{z}_i^n, \text{NoC}), \forall\ i = 1 \ldots m\ , n = 1 \ldots N,$$

where the inputs of the clustering algorithm are the target variables of the $\mathbf{Z}_i$ group and the parameter NoC.

In this work, we use the K-medoids clustering algorithm [6, Sec. 14.3.10] - mainly due to its robustness to outliers - but any clustering algorithm that is compatible with the target variables could be used instead. Furthermore, the distance metric can be chosen according to the multi-target problem at hand (e.g. Hamming distance for multi-label classification and Euclidean distance for multivariate regression).

At this point, the problem of estimating the joint (high-dimensional) density of the targets in each group becomes a problem of estimating a discrete distribution of NoC categories. The trade-off is between making no approximations and estimating high-dimensional densities, which leads to poor and unreliable estimates of the MI, or deriving lower dimensional approximations through clustering, which leads to more reliable estimates of the MI.

4 Experiments

In this work, we evaluate our multi-target feature selection approach on multi-label classification problems[1]. We focus on three datasets with diverse characteristics: emotions (N=593, d=72, m=6), image (N=2000, d=294, m=5) and yeast (N=2417, d=103, m=14) [4]. To compare the performance of the different FS methods, we train a multi-label classifier using the selected features and evaluate its performance on the testing data using two measures: hamming and ranking loss [4]. We used the ML-kNN classifier, setting the number of nearest neighbours to 7 as suggested in [7] and the experimental protocol that we follow is 50x2-fold cross-validation. To estimate MI we use the normalised maximum-likelihood plug-in estimator, discretising continuous features into 5 bins by the equal width strategy.

[1] A more extensive evaluation that includes multivariate regression is left for future work.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

In the first set of experiments we analyse the sensitivity of the proposed algorithm, with respect to the PoT and NoC parameters using the hamming loss metric[2] and for various numbers of selected features (k=10...40). Figure 1 shows the performance for different numbers of clusters (NoC) when PoT is fixed to 0.50. We notice that the optimal number is 4 for image and yeast and 8 for emotions. Figure 2 shows the performance for different proportions of targets when NoC is fixed to 8. We notice that the best performance is achieved by groups that contain 75% of the targets in image and by groups that contain 50% of the targets in emotions, while for yeast there is no clear winner. These results highlight the importance of correct parametrisation, and the fact that the optimal parameters depend on the intrinsic characteristics of each dataset.

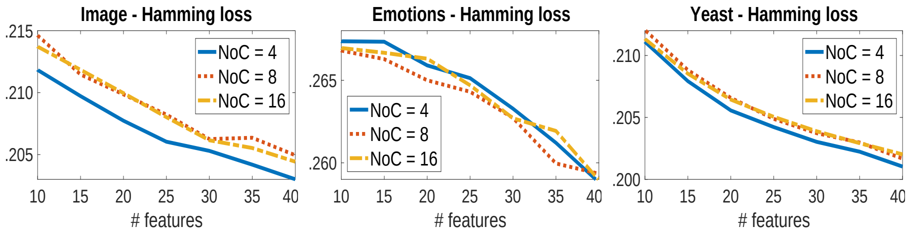

Fig. 1: Comparing Group-JMI for various values of NoC with PoT fixed to 0.50.

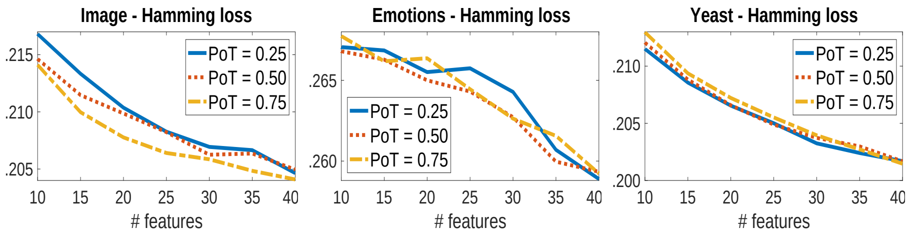

Fig. 2: Comparing Group-JMI for various values of PoT with NoC fixed to 8.

We now compare Group-JMI against the methods presented in Sec. 02, Joint-JMI and Single-JMI. To show the robustness of Group-JMI, we will create two versions by randomly choosing the parameters for generating each $\widetilde{Y}_i$:
Group-JMI-Small (G-S): PoT chosen from (0-0.50] and NoC from {4,...,16},
Group-JMI-Large (G-L): PoT chosen from (0.50-1] and NoC from {4,...,16}.
By this parametrisation G-S uses a small number of targets to generate each group (less than half), while G-L a large (more than half). Table 1 presents the average ranking score of each FS method. We select various values of top-k selected features, k=10, 15, 20, ...,40, for each k the method with the lowest loss is assigned ranking score 1, the second best 2, etc., and at the end we average the scores across all k. Overall, we see that our methods achieve the best performance in five out of six settings, where each setting is a combination of an evaluation loss measure and a particular dataset.

[2]Similar results hold for ranking loss, but we have omitted them due to space limitations.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

	Hamming loss					Ranking loss				
Dataset	Single	Joint	G-S	G-L		Single	Joint	G-S	G-L	
emotions	2.00	4.00	2.43	**1.57**	✓	**1.86**	4.00	2.29	**1.86**	✓
image	2.14	3.29	3.43	**1.14**	✓	2.29	3.71	3.00	**1.00**	✓
yeast	**1.57**	4.00	1.71	2.71		1.86	4.00	**1.43**	2.71	✓

Table 1: The average ranking score with two losses and three datasets. The best method for each loss is highlighted in bold and ✓ indicates a setting (a combination of evaluation loss and dataset) where our methods outperform competitors.

5 Conclusions and future work

In this work we presented a generic FS algorithm suitable for multi-label classification and multivariate regression problems. Our method is using the JMI principle to derive low-order approximations of the input space, and it clusters similar targets to derive low-order approximations of the output space. Group-JMI has two parameters: the PoT that controls the number of targets that interact in each group, and the NoC that controls the dimensionality of the density that we will try to estimate.

Our future work will focus on providing methods for optimising these parameters. One approach is to use a validation set and minimise a loss of a particular classifier, but this violates the filter assumption: selecting the features independently of any classifier or evaluation measure. To overcome this issue our current line of work splits in two directions. For PoT we explore ways of automatically grouping the targets that share some minimum amount of information measured by multi-variate MI. For optimising NoC we explore ways to determine the maximum number of clusters we can have to estimate reliably MI from the available data. This can be done by performing sample size determination for observing given MI quantities with a particular statistical power [8]. Lastly, a more extensive evaluation that includes multi-label and multivariate regression with large output spaces is left for future work.

References

[1] G. Brown, A. Pocock, M.-J.. Zhao, and M. Lujan. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *JMLR*, 2012.

[2] Willem Waegeman, Krzysztof Dembczynski, and Eyke Huellermeier. Multi-target prediction: A unifying view on problems and methods. *arXiv preprint arXiv:1809.02352*, 2018.

[3] K. Sechidis, N. Nikolaou, and G. Brown. Information theoretic feature selection in multi-label data through composite likelihood. In *S+SSPR*. 2014.

[4] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009.

[5] G. Tsoumakas, I. Katakis, and I. Vlahavas. Random k-labelsets for multilabel classification. *IEEE Trans. on Knowledge and Data Engineering*, 23(7):1079–1089, 2011.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. 2009.

[7] Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *IEEE International Conference on Granular Computing*, 2005.

[8] K. Sechidis and G. Brown. Simple strategies for semi-supervised feature selection. *Machine Learning*, 2018.