entropy

MDPI

Article

Information Theoretic Multi-Target Feature Selection via Output Space Quantization †

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

1 Department of Computer Science, Aristotle University, 54124 Thessaloniki, Greece
2 School of Computer Science, University of Manchester, Manchester M13 9PL, UK
3 Expedia, 1207 Geneva, Switzerland
* Correspondence: konstantinos.sechidis@manchester.ac.uk

† This Paper is an Extended Version of Our Paper Published in the 27th European Symposium on Artificial Neural Networks (ESANN), Computational Intelligence and Machine Learning, Bruges (Belgium), 24–26 April 2019.

Received: 5 August 2019; Accepted: 29 August 2019; Published: 31 August 2019

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 an 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 quantization algorithms for deriving low-dimensional approximations of the output space. Under the above framework we derive a novel criterion, *G*roup-JMI-Rand, which captures various high-order target interactions. In an extensive experimental study we showed that our suggested criterion achieves competing performance against various other information theoretic feature selection criteria suggested in the literature.

Keywords: feature selection; mutual information; multi-target; multi-label; clustering

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]. In our work we will focus on information theoretic FS criteria, which quantify the importance of each feature by estimating mutual information terms to capture—the relevancy, the redundancy and the complementarity [2]. Choosing a subset of features that has the highest relevancy with the output space, the minimum redundancy between them and the highest complementarity, helps us to reduce the input space and at the same time keep as much useful information as possible.

At the same time more and more applications need to predict multiple outputs (targets), instead of a single one. Depending on 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 [3]. For example, in computer vision [4], multi-label data are used in automated image annotation, since an image can be associated with a number of semantic concepts. In bioinformatics [5], multi-dimensional learning is used in functional genomics, where a gene or protein is associated with multiple functional labels, since an individual gene or protein usually performs a number of functions. Finally, multivariate regression has

been used in ecological modeling in order to predict various target variables that capture the quality of the vegetation [6].

In this paper we focus on deriving novel information theoretic FS 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. [7] 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 feature and full independence in the target space, a method known as Single-JMI (Joint Mutual Information), details in Section 2. While the partial independence of the feature space has been proven to be useful in deriving FS criteria for single-label data [8], the full independence in the label space ignores the useful information that the possible dependencies between the targets can provide.

Our work, which is an extension of the conference paper in Reference [9], introduces an algorithm that uses the principles of the Single-JMI criterion but at the same time takes into account target dependencies. In the current work, we expanded the preliminary conference paper, by extending the discussion of related work (Section 2), by providing a novel theoretical and sensitivity analysis (Sections 3.2 and 3.3 respectively), by providing a larger empirical study for multi-label classification (Section 4), including more datasets and competing methods and by providing a novel empirical study on multivariate regression problems (Section 5). The software related to this paper, including implementations of our novel FS criteria, is available at: https://github.com/sechidis.

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, that is, the alphabet $\mathcal{Y}$ is $\{0,1\}^m$, the problem is known as multi-label classification, when they are categorical as multi-dimensional classification $\mathcal{Y}$ is $\{0,\dots,c\}^m$, while when they are continuous, that is, $\mathcal{Y}$ is $\mathbb{R}^m$, as multivariate regression [3].

The problem of FS can be phrased as selecting a subset of K features $\mathbf{X}_\theta \subset \mathbf{X}$, where $|\mathbf{X}_\theta| = K$, that contain as much useful information for our problem as possible. With a slight abuse of notation, in the rest of our work, we interchange the symbol for a set of variables and for their joint random variable. FS methods can be categorized in three groups [10]—filters, wrappers and embedded. Filters are independent of the classifier and they define a scoring criterion (or relevance index) by which they produce a ranking of the features. Wrappers are classifier dependent; they use an evaluation measure to check the performance of the different subsets of features with a particular classifier and they choose the subset with the best performance. Finally, embedded methods are again classifier dependent, since they are part of the learning algorithm and the FS is applied in the training procedure. From the above descriptions we can find the strengths and the weaknesses of each approach. Filters are classifier independent, they are fast and they are less likely to overfit but on the other hand the performance is worse than the classifier specific methods (some of the filters may underfit the data). Embedded methods require some model, which introduces additional assumptions and may be slower than filters but may result to better performance and tend to overfit less than wrappers. Wrappers, because they are classifier dependent, may achieve better performance but on the other hand, they are computationally intensive and tend to overfit more than the other techniques [1,8].

In our work we focus on *filter* methods for FS, which operate under the assumption that the prediction and FS steps are independent [1] or in other words, the selection of features is independent of the classifier or the regressor used. 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 and it is desirable to increase 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 [8].

2.1. Deriving Criteria via Maximum Likelihood Maximization Framework

For single-output problems, that is, the output space is a single variable Y, Brown et al. [8] 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. CMI criterion can be written in the following way:

$$J_{\text{CMI}}(X_k) = I(X_k; Y) - I(X_k|\mathbf{X}_\theta) + I(X_k; \mathbf{X}_\theta|Y).$$

The first term of the above expression corresponds to the relevancy of a feature with the target, the second to the redundancy of a feature with the set of features already selected and the last term to the complentarity (or conditional redundancy) of the feature with the set of selected features. While the importance of the first two terms is pronounced in the FS literature, the last term has not been traditionally accounted [8]. This term has opposite sign than redundancy, which means that dependent features can be useful, as long the dependence within class is stronger than the overall dependence.

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 popular 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 [8]:

$$\begin{aligned} J_{\text{JMI}}(X_k) &= \sum_{X_j \in \mathbf{X}_\theta} I(X_j X_k; Y) \\ &\propto I(X_k; Y) - \frac{1}{|\mathbf{X}_\theta|} \sum_{X_j \in \mathbf{X}_\theta} \left(I(X_k; X_j) - I(X_k; X_j|Y) \right), \end{aligned}$$

where the symbol $\propto$ indicates a ranking equivalent expression for the criterion. The proof for this ranking equivalence can be found in Appendix A.1 of Reference [8]. From the last expression we can see that JMI takes into account all three desirable terms—the score increases when the relevance of a feature is high, when the average redundancy with the features already selected is low and when the average complementarity with the selected features is high.

Sechidis et al. [7] derived two versions of the JMI criterion suitable for multi-output problems, that is, the output space is a joint variable $\mathbf{Y} = Y_1 \dots Y_m$. Their approach was based on the idea of expressing multi-label decomposition methods as composite likelihoods and then showing how FS criteria can be derived by greedily maximising these likelihood expressions. Different decomposition

methods lead naturally to different FS criteria. The scoring functions for the two multi-output criteria suggested by Sechidis et al. [7] are the following:

$$J^{\text{Joint}}_{\text{JMI}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} I(X_j X_k; \mathbf{Y}), \tag{1}$$

$$J^{\text{Single}}_{\text{JMI}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} \sum_{Y_i \in \mathbf{Y}} I(X_j X_k; Y_i). \tag{2}$$

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 (LP) 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 [11], 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 it is independent from the rest (Single-JMI).

In a small experimental study, using only two datasets, Sechidis et al. [7] showed that Single-JMI, even though it assumes full independence between the targets, outperforms Joint-JMI, whicht 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. Before that we will review other information theoretic criteria suggested in the literature, while a systematic review on multi-label FS methods can found in Reference [12].

2.2. Other Information Theoretic Criteria

Yang & Pedersen [13] introduced the first information theoretic multi-label FS method, which ranks the features using the criterion: $J^{BR}_{\text{MIM}}(X_k) = \sum_{Y_i \in \mathbf{Y}} I(X_k; Y_i)$. *MIM-BR* ranks the features only on their relevancy with each target independently and it does not take into account possible correlations between features (i.e., redundancy/complementarity). *AMI* [14] is an extension that takes into account redundancy terms but still treats each label independently. *ELA+CHI* [15] uses an Entropy-based Label Assignment, which assigns the labels weights based on label entropy, to transform the label space and then uses the χ^2 statistic, a quantity that is asymptotically equivalent to the MI [16], to rank the features. Lee & Kim [17] proposed *PMU*, a criterion that uses the multivariate MI and avoids the computational cost by restricting the number of variables to three. The same authors suggested *FIMF* [18], an algorithm for a computationally efficient information theoretic FS and more recently *SCLS* [19] that introduces a novel way of measuring feature redundancy.

All the above methods were proposed for solving the classification problem (i.e., multi-label) and to the best of our knowledge our work is the first that suggests an information theoretic algorithm that can be used for any kind of multi-target tasks, even on multivariate regression using the default plug-in MI estimator.

At this point we should clarify that in information theoretic FS the scoring criterion, for example, Equations (1) and (2), is combined with a search method which describes how the candidate feature sets are selected. All of the FS algorithms presented so far use greedy forward search, testing each

feature in turn for inclusion and adding the one with the highest score. Using a greedy search to present the capabilities of a criterion it is a widely used strategy in the information theoretic FS literature [8]. Apart from the greedy (forward or backward) methods to optimize a scoring criterion, more advanced methods can be used, such as genetic algorithms ([1], Chapter 4). For the remainder of this paper, we will use greedy forward search to test our suggested novel scoring criteria.

3. A Novel Framework to Take into Account Target Dependencies

3.1. Transforming Output Space via Quantization to Account for Target Dependencies

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*:

$$J_{\text{JMI}}^{\text{Group}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} I(X_j X_k; \widetilde{Y}_i). \tag{3}$$

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, quantization-based strategy:

- 1st Step—Generate Groups of Target Variables, Using PoT Parameter

In this step we create m groups of variables $\mathbf{Z}_1, ..., \mathbf{Z}_m$, where each group is a random subset of the targets, that is, $\mathbf{Z}_i \subset \mathbf{Y} \; \forall \; i = 1, \dots, m$. Each group is generated by sampling the set of target variables without replacement and by 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 [20], a state of the art method for learning from multi-label data.

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 $\mathbf{Z}_1, \dots, \mathbf{Z}_{20}$ will be generated, each one consisting of $6 (= 20 \times 0.30)$ 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 Quantization, Using NoC Parameter

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 Cluster (NoC) is provided 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:

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

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

In this work, we use the K-medoids clustering algorithm ([21], Section 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.

Algorithm 1 provides a greedy forward FS algorithm using our Group-JMI criterion. In Line 9 we need to estimate the JMI between two features, that is, X_j and X_k and the transformed target variable $\widetilde{Y}_j$ from our sample data. Any MI estimator can be used for this task [22]. In our work we use the plug-in estimator for the MI:

$$\hat{I}(X_jX_k;\widetilde{Y}_i) = \sum_{x_j\in\mathcal{X}_j}\sum_{x_k\in\mathcal{X}_k}\sum_{\widetilde{y}_i\in\widetilde{\mathcal{Y}}_i} \hat{p}(x_j,x_k,\widetilde{y}_i)\ln\frac{\hat{p}(x_j,x_k,\widetilde{y}_i)}{\hat{p}(x_j,x_k)\hat{p}(\widetilde{y}_i)}, \tag{4}$$

where, for example, $\hat{p}(x_j,x_k,\widetilde{y}_i)$ is the maximum likelihood estimate of the joint probability that the random variable X_j takes the value x_j, the random variable X_k takes the values x_k and the random variable $\widetilde{Y}_i$ takes the values $\widetilde{y}_i$. Estimating these probabilities with categorical features is straightforward, while continuous features can be discretised, for example equal-width discretisation is used often in the FS literature [8,17].

Algorithm 1 Forward FS with our Group-JMI criterion

Input: Dataset $\{\mathbf{x}^n,\mathbf{y}^n\}_{n=1}^N$, parameters PoT and NoC and the number of features to be selected K.

Output: List of top-K features $\mathbf{X}_\theta$

1: $\mathbf{X}_{\widetilde{\theta}} = \mathbf{X}$ ▷ Set of candidate features

2: Set $\mathbf{X}_\theta$ to empty list ▷ List of selected features

3: **for** $i := 1$ to m **do** ▷ Output transformation (where m is the number of target variables)

4: Use PoT to generate a random subset of targets: $\mathbf{Z}_i \subset \mathbf{Y}$

5: Derive $\widetilde{Y}_i$, from the cluster indices: $\widetilde{Y}_i = \text{Clustering}(\mathbf{Z}_i, \text{NoC})$

6: **end for**

7: **for** $k := 1$ to K **do**

8: Let $X_k^* \in \mathbf{X}_{\widetilde{\theta}}$ maximise:

9: $$J_{\text{JMI}}^{\text{Group}}(X_k) = \sum_{X_j\in\mathbf{X}_\theta}\sum_{\widetilde{Y}_i\in\widetilde{\mathbf{Y}}} I(X_jX_k;\widetilde{Y}_i)$$ ▷ Our scoring criterion

10: $\mathbf{X}_\theta(k) = X_k^*$ ▷ Add feature X_k^* to the list

11: $\mathbf{X}_{\widetilde{\theta}} = \mathbf{X}_{\widetilde{\theta}} \backslash X_k^*$ ▷ Remove feature X_k^* from the candidate set

12: **end for**

3.2. Theoretical Analysis

Now we will show that our suggested criterion, Group-JMI, captures all three desirable characteristics of an information theoretic FS criterion—relevancy, redundancy and complementarity. Let us start from Equation (3):

$$J_{\text{JMI}}^{\text{Group}}(X_k) = \sum_{X_j\in\mathbf{X}_\theta}\sum_{\widetilde{Y}_i\in\widetilde{\mathbf{Y}}} I(X_jX_k;\widetilde{Y}_i). \tag{5}$$

Using the chain rule for mutual information, $I(AB;C) = I(A;C) + I(B;C|A)$, the criterion can be written as follows:

$$J_{\text{JMI}}^{\text{Group}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} \left(I(X_j; \widetilde{Y}_i) + I(X_k; \widetilde{Y}_i | X_j) \right). \tag{6}$$

The term $\sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} I(X_j; \widetilde{Y}_i)$ in the above is constant with respect to the X_k argument that we are interested in, so can be omitted and the criterion gets the following ranking equivalent form:

$$J_{\text{JMI}}^{\text{Group}}(X_k) \propto \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} I(X_k; \widetilde{Y}_i | X_j). \tag{7}$$

By using the information theoretic identity $I(A;B|C) = I(A;B) - I(A;C) + I(A;C|B)$, the criterion can be written as follows:

$$\begin{aligned}
J_{\text{JMI}}^{\text{Group}}(X_k) &\propto \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} \left(I(X_k; \widetilde{Y}_i) - I(X_k; X_j) + I(X_k; X_j | \widetilde{Y}_i) \right). \\
J_{\text{JMI}}^{\text{Group}}(X_k) &\propto |\mathbf{X}_\theta| \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} I(X_k; \widetilde{Y}_i) - \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} \left(I(X_k; X_j) - I(X_k; X_j | \widetilde{Y}_i) \right) \\
J_{\text{JMI}}^{\text{Group}}(X_k) &\propto \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} I(X_k; \widetilde{Y}_i) - \frac{1}{|\mathbf{X}_\theta|} \sum_{X_j \in \mathbf{X}_\theta} \sum_{\widetilde{Y}_i \in \widetilde{\mathbf{Y}}} \left(I(X_k; X_j) - I(X_k; X_j | \widetilde{Y}_i) \right)
\end{aligned} \tag{8}$$

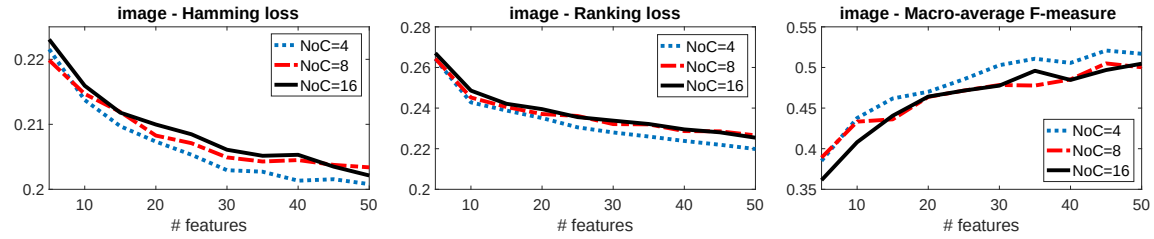
Interestingly, by the decomposition of Equation (8), the first term of *rhs* captures the relevancy of the feature X_k and each transformed target variable $\widetilde{Y}_j$, the second term the average redundancy between the feature X_k and the already selected features $X_j \in \mathbf{X}_\theta$, while the final term captures the average complementarity between the feature X_k and the already selected features, given each transformed target variable $\widetilde{Y}_j$. The first and the third have positive contribution, while the second negative.

3.3. Sensitivity Analysis

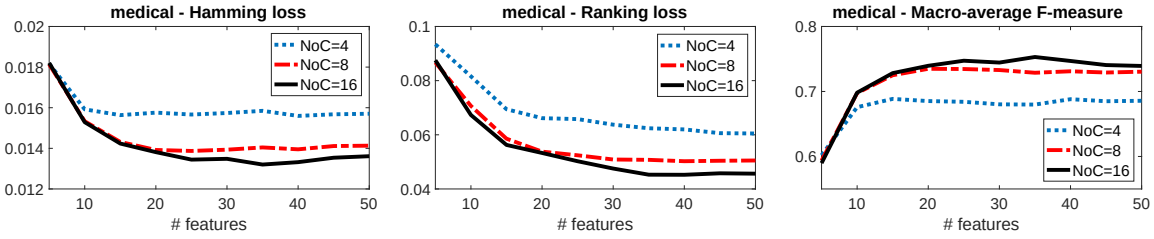
This section presents the sensitivity of the proposed algorithm, with respect to the PoT and NoC parameters. We will focus on three multi-label datasets (*image*, *medical*, *genbase*), using three evaluation measures (hamming loss, ranking loss, macro-average F-measure) and in various numbers of selected features ($K = 1, \ldots, 50$). More details on the experimental setting will be given in Section 4.

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* (Figure 1a), 16 for *medical* (Figure 1b), while for *genbase* there is no clear winner between 8 and 16 (Figure 1c). 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* (Figure 2a), by groups that contain 25% of the targets in *medical* (Figure 2b), while for *genbase* there is no clear winner between 50% and 75% (Figure 2c).

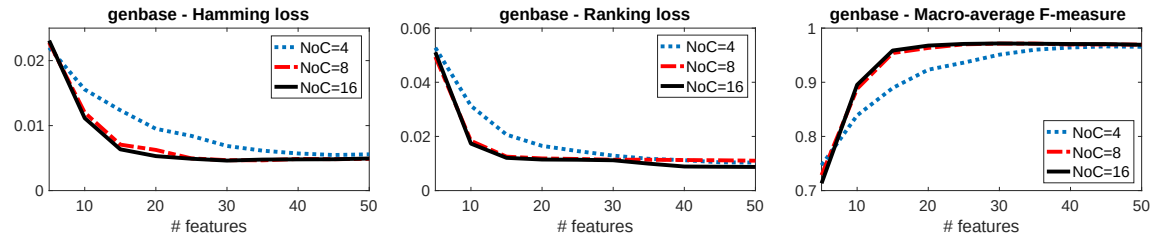
These results highlight the power of our novel parametrisation and the fact that the optimal parameters depend on the intrinsic characteristics of each dataset. For example, the *image* dataset has few labels and distinct label combinations, as a result NoC = 4 is a good approximation, which is not the case for *medical*, a dataset with many labels. On the other hand, the larger the number of labels, the smaller the best PoT. For example in the *medical* dataset, using a PoT = 0.25 means that in each combination we have $\sim$11 labels, which is already much higher than the total labels of *image* (5 labels). As a result, in *image* we achieve better performance with high values of PoT, while in *medical* with lower.



(**a**) Results for *image* dataset



(**b**) Results for *medical* dataset



(**c**) Results for *genbase* dataset

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

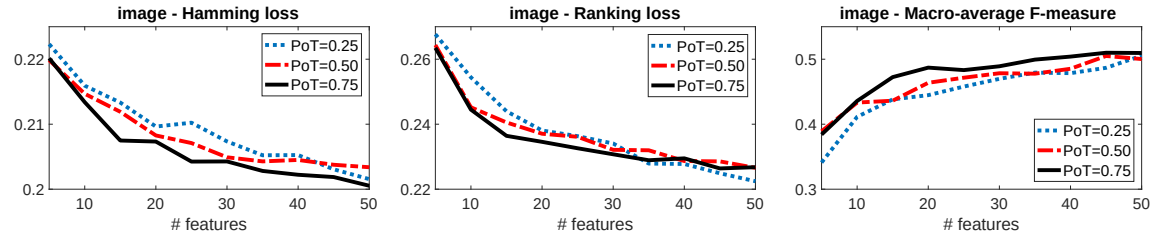
3.4. A Group-JMI Criterion That Captures Various High-Order Target Interactions

One approach to estimate the optimal parameters is by using grid-search on a hold-out set to optimize a specific evaluation measure. However, this approach assumes that a specific multi-target classification/regression algorithm will be used. Unfortunately, this conflicts with the filter assumption—select features independently from the classification/regression algorithm (more details in Section 2).

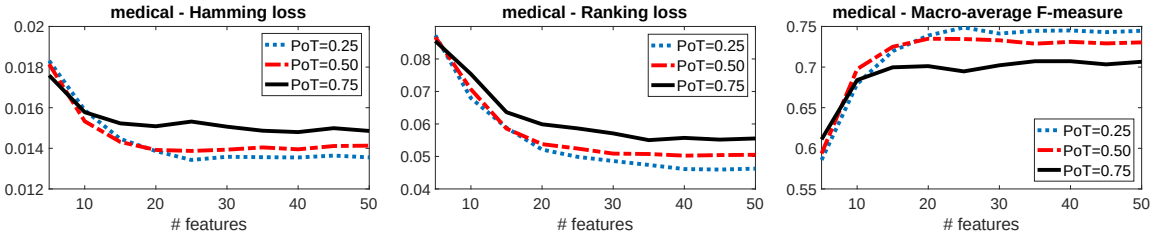
To overcome this issue, we suggest Group-JMI-Rand, which chooses the parameters for generating each $\widetilde{Y}_i$, uniformly at random from the following pre-specified set:

Group-JMI-Rand: PoT chosen randomly from [0.25–0.75] and NoC from {4,...,16}.

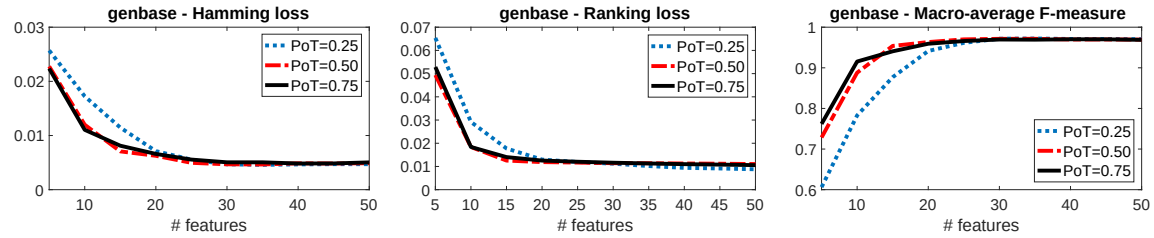
By this parametrisation Group-JMI-Rand uses a large number of targets, since to generate each group we sample at random 25–75% of the targets. At the same time clustering keeps the dimensionality of the estimated densities low. To achieve this we are randomly choosing in each group the number of clusters to be between 4–16. In the next section we will show that the above criterion achieves state-of-the-art performance in various datasets and evaluation measures.



(**a**) Results for *image* dataset



(**b**) Results for *medical* dataset



(**c**) Results for *genbase* dataset

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

4. Experiments with Multi-Label Data

We focus on various multi-label datasets with diverse characteristics, shown in Table 1 [23].

Table 1. Characteristics of the multi-label datasets.

Name	Application	Examples	Features	Labels	Distinct Labelsets
CAL500	Music	502	68	174	502
emotions	Music	593	72	6	27
enron	Text	1702	1001	53	752
genbase	Biology	662	1186	27	32
image	Images	2000	294	5	20
languagelog	Text	1460	1004	75	1241
medical	Text	978	1449	45	94
scene	Images	2407	294	6	15
yeast	Bioinformatics	2417	103	14	198

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 four measures—hamming loss, ranking loss, normalised coverage and macro-average F-measure [11]. Following the FS literature [8], we used a nearest neighbour classifier, which makes as few assumptions as possible about the data and we avoid the need for parameter tuning. For our work we used the multi-label nearest neighbour classifier introduced by Zhang and Zhou [24] and, following their recommendation we set the number of neighbours to 7. We conducted a holdout balanced cross-validation for each experiment—50% of the examples in a given dataset were randomly chosen as the training set for multi-label FS

and classifier training and the remaining 50% were used as the test set to obtain the multi-label classification performance to be reported. Each experiment was repeated 30 times and the average testing performance was reported.

To take into account the performance over various values of selected features, we select top-K with $K = 1, \ldots, 50$. For each K the method with the best performance (i.e., lowest loss) is assigned ranking score 1, the second best 2 and so forth, and at the end we average the scores across all K. This score provides an indication on how well each method performs across a range of K values. Finally, for estimating MI the default plug-in estimator was used, while continuous features were discretised into 5 bins, using an equal-width strategy [8].

4.1. Comparing Group-JMI-Rand with Other JMI Criteria

Firstly, we will compare our novel JMI criterion, Group-JMI-Rand, with the two multi-label JMI criteria that have been suggested in Reference [7]—Single-JMI and Joint JMI (more details in Section 2). Table 2 presents the ranking score of each FS method averaged across all possible FS sizes (top-$K = 1, \ldots, 50$). Overall, we see that our method achieves the best performance in 20 out of 36 settings, while Joint-JMI in 13 and Single-JMI in 3. Each setting is a combination of an evaluation loss measure and a particular dataset.

From this set of experiments we can conclude that our initial idea, to derive a criterion that is a trade-off between the two extremes, Single-JMI (assumes independent targets, thus needs to estimate low-dimensional probability distributions) versus Joint-JMI (no assumption at all, thus needs to estimate high-dimensional probability distributions) outperforms both of them. This is happening because Group-JMI-Rand, by using the parameter PoT, randomly groups the labels and as a result it does not assume full independence between the labels. At the same time, by using a quantization algorithm the probability distribution is compressed in a low density specified by the NoC parameter. Interestingly, even choosen PoT and NoC at random from a large pre-specified set of values outperforms the competing methods.

Table 2. Comparing the three JMI based criteria in terms of the average ranking score using five evaluation measures: (**a**) hamming loss, (**b**) ranking loss, (**c**) normalised coverage and (**d**) macro-average F-measure. The best method for each combination of evaluation measure and dataset is highlighted in bold.

(a) Hamming Loss			
	Single-JMI [7]	**Joint-JMI [7]**	**Group-JMI-Rand (Our Method)**
CAL500	2.05	2.10	**1.85**
emotions	2.33	1.85	**1.82**
enron	2.10	**1.00**	2.90
genbase	2.11	**1.71**	2.17
image	1.90	2.73	**1.38**
medical	2.01	2.86	**1.12**
scene	1.80	**1.25**	2.95
yeast	1.57	3.00	**1.43**
languagelog	1.60	**1.40**	3.00
Total wins	0	4	**5**
(b) Ranking Loss			
	Single-JMI [7]	**Joint-JMI [7]**	**Group-JMI-Rand (Our Method)**
CAL500	2.20	1.93	**1.88**
emotions	**1.57**	2.40	2.02
enron	1.75	**1.30**	2.95
genbase	2.29	**1.66**	2.05

Table 2. *Cont.*

image	1.90	2.77	**1.32**
medical	2.11	2.79	**1.10**
scene	1.90	**1.15**	2.95
yeast	1.52	3.00	**1.48**
languagelog	2.62	2.38	**1.00**
Total wins	1	3	**5**

(c) Normalised Coverage

	Single-JMI [7]	**Joint-JMI [7]**	**Group-JMI-Rand (Our Method)**
CAL500	**1.75**	2.35	1.90
emotions	1.95	2.80	**1.25**
enron	1.82	**1.25**	2.92
genbase	2.14	**1.44**	2.42
image	2.08	2.50	**1.43**
languagelog	2.40	2.60	**1.00**
medical	1.96	2.84	**1.20**
scene	1.57	**1.48**	2.95
yeast	1.62	3.00	**1.38**
Total wins	1	3	**5**

(d) Macro-average F-measure

	Single-JMI [7]	**Joint-JMI [7]**	**Group-JMI-Rand (Our Method)**
CAL500	1.92	2.25	**1.82**
emotions	2.10	2.08	**1.82**
enron	2.00	**1.00**	3.00
genbase	2.34	**1.49**	2.17
image	1.77	2.80	**1.43**
languagelog	**1.43**	1.65	2.92
medical	2.01	2.84	**1.15**
scene	1.77	**1.30**	2.92
yeast	1.75	3.00	**1.25**
Total wins	1	3	**5**

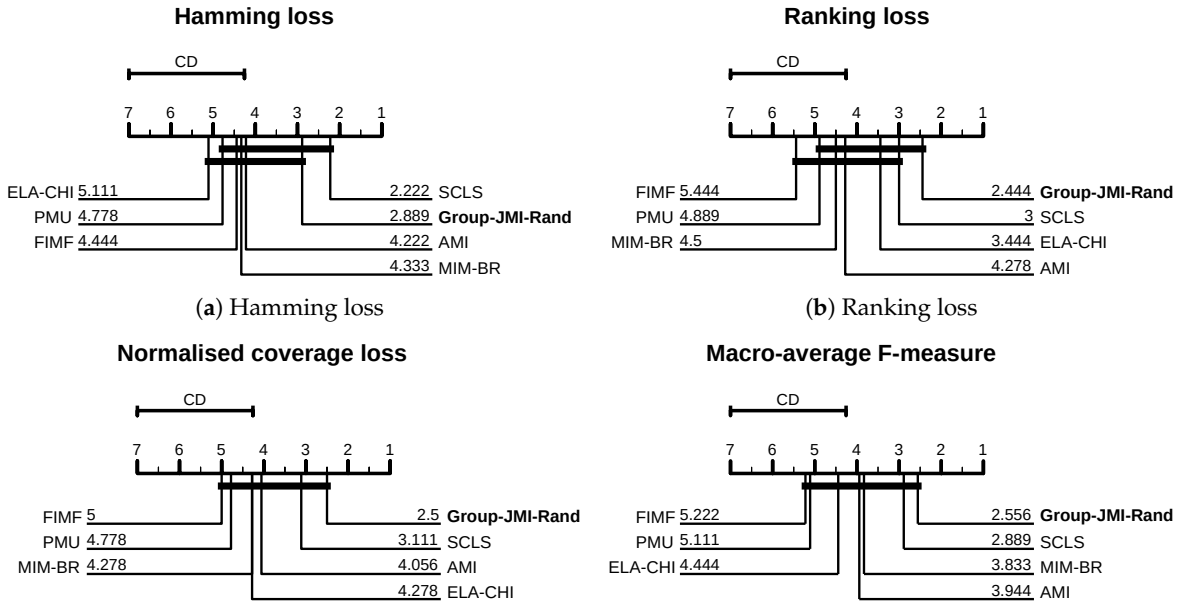
4.2. Comparing Group-JMI-Rand with State-of-the-Art Information Theoretic FS Criteria

To test the efficiency of the proposed criterion Group-JMI-Rand, we will compare its performance against six information theoretic FS suggested in the literature—MIM-BR [13], ELA-CHI [15], AMI [14], PMU [17], FIMF [18] and SCLS [19] (arranged in chronological order). More details on the competing methods can be found in Section 2.

In the literature on data mining and machine learning there are various ways on performing statistically sound comparisons between different methods [25–27]. In our work we will use the critical difference diagrams (CD), introduced by Demšar [25] and Figure 3 presents our results. For all the CD diagrams of this work, groups of methods that are not significantly different at level $\alpha = 0.05$ (using the Nemenyi post-hoc test) are connected. The method that achieves the best performance is given a rank of 1, the second best a rank of 2 and so forth.

Our suggested criterion, Group-JMI-Rand, performs better than the competitors in three evaluation measures—ranking loss (Figure 3b), normalized coverage (Figure 3c) and Macro-average F-measure (Figure 3d), while for hamming loss (Figure 3a), a measure that does not take into account label dependencies, the SCLS [19] method performs better. Another interesting conclusion is that our method and SCLS are always in the top-2 positions and in all four evaluation measures there is no statistically significant difference between them. Due to the quantization of the output space,

Group-JMI-Rand is more flexible and apart from multi-label data it can be also used to multi-variate regression problems and the next section focuses on this type of data.



(**a**) Hamming loss (**b**) Ranking loss

(**c**) Normalized coverage (**d**) Macro-average F-measure

Figure 3. Comparing our suggested FS criterion, Group-JMI-Rand, with state-of-the-art approaches across four evaluation measures: (**a**) Hamming loss, (**b**) Ranking loss, (**c**) Normalized coverage and (**d**) Macro-average F-measure.

5. Experiments with Multivariate Regression Data

In this section we focus on various multi-variate regression datasets, shown in Table 3 [28].

Table 3. Characteristics of the multi-variate regression datasets.

Name	Application	Examples	Features	Targets
atp1d	Airline Ticket Price	337	411	6
atp7d	Airline Ticket Price	296	411	6
oes97	Occupational Employment Survey	334	263	16
oes10	Occupational Employment Survey	403	298	16
osales	Online Product Sales	639	413	12
scm20d	Supply Chain Management	8966	61	16

As we already mentioned in Section 2, there are no information theoretic FS criteria tailored to multivariate regression data suggested in the literature. For that reason we compare the performance of our proposed algorithm Group-JMI-Rand (using the Euclidean distance for clustering, since we have continuous variables this time instead of binary), against a popular filter FS method, tailored to regression problems—*RReliefF* (Regressional ReliefF) [29]. RReliefF is a nearest neighbor-based feature weighting method for univariate regression problems. In a multivariate regression context, we apply RReliefF separately for each target to get an importance weight per feature and target and then rank the features based on their average importance weight across all targets. We compare the performance of two different variations of RReliefF, *RReliefF 10* and *RReliefF 50* setting the number of neighbours to 10 and 50 respectively.

To compare the performance of the different FS methods, making as few assumptions as possible, we used again a nearest neighbors regression model and predict each target independently. In this set of experiments we set the number of neighbours to be 10, same number of neighbours as in *RReliefF 10*.

Finally, the evalutation measure we used is the average Relative Root Mean Squared Error (RRMSE) across all targets, a measure widely used in the multi-target regression literature [28].

Figure 4 shows that our proposed method *JMI-Group-Rand* achieves the best performance in four out of six datasets (*atp1d atp7d*, *oes10*, *osales*). In *oes97* it achieves the same performance as RReliefF 50, while in *scm20d* the RReliefF methods outperform our information theoretic criterion.

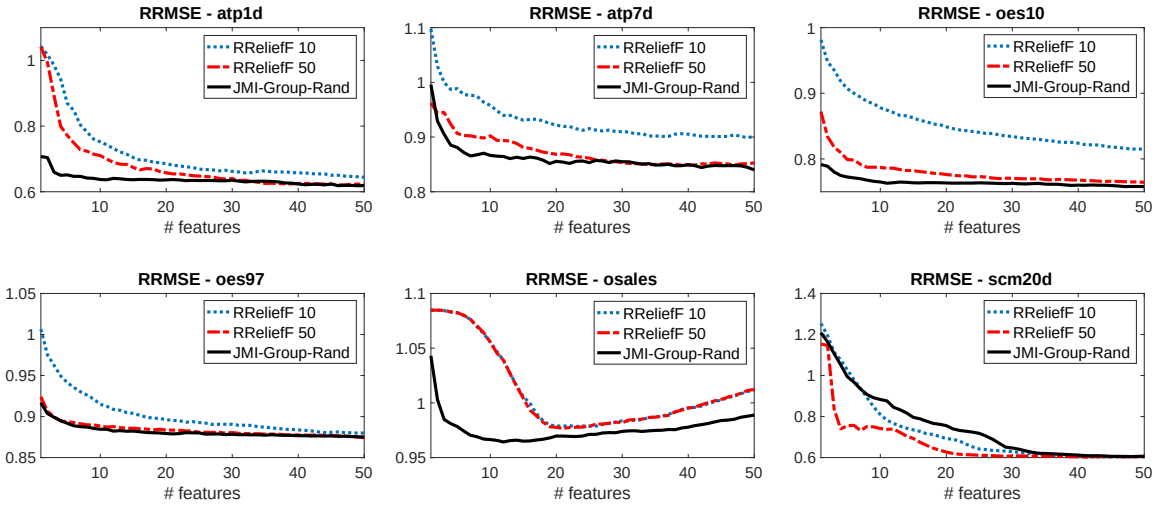


Figure 4. Multi-variate regression experiments. Comparing our suggested criterion with two different versions of RReliefF.

6. Conclusions

In this work we presented a FS algorithm suitable for multi-target problems, such as multi-label classification and multivariate regression. Our criterion, Group-JMI, uses 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 that capture target correlations. 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 try to estimate. Under our framework, we suggest the Group-JMI-Rand criterion, which chooses these two parameters at random from a prespecified set of values. On an extensive empirical study across 15 real-world datasets, 10 competing methods and 5 evaluation measures, our proposed criterion Group-JMI-Rand achieves a competitive performance against various other information theoretic FS criteria.

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 [30]. Finally, by connecting the problem of multi-target FS with the problem of biomarker discovery in clinical trials with multiple endpoints, we can potentially use Group-JMI-Rand for deriving prognostic and predictive biomarkers in multiple endpoint trials [31].

Author Contributions: Formal analysis, K.S.; Funding acquisition, K.S., E.S.-X. and I.V.; Methodology, K.S. and E.S.-X.; Software, K.S.; Supervision, I.V.; Validation, K.S.; Visualization, K.S.; Writing—original draft, K.S.; Writing—review & editing, K.S., E.S.-X. and I.V.

Funding: 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.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FS	Feature Selection
MI	Mutual Information
CMI	Conditional Mutual Information
JMI	Joint Mutual Information
NoC	Number of Clusters
PoT	Proportion of Targets

References

1. Guyon, I.M.; Gunn, S.R.; Nikravesh, M.; Zadeh, L. (Eds.) *Feature Extraction: Foundations and Applications*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2006.
2. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
3. Waegeman, W.; Dembczynski, K.; Huellermeier, E. Multi-target prediction: A unifying view on problems and methods. *arXiv* **2018**, arXiv:1809.02352.
4. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning multi-label scene classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [CrossRef]
5. Elisseeff, A.; Weston, J. A Kernel Method for Multi-Labelled Classification. In *Advances in Neural Information Processing Systems (NIPS) 14*; MIT Press: Cambridge, MA, USA, 2001; pp. 681–687.
6. Kocev, D.; Džeroski, S.; White, M.D.; Newell, G.R.; Griffioen, P. Using single-and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecol. Model.* **2009**, *220*, 1159–1168. [CrossRef]
7. Sechidis, K.; Nikolaou, N.; Brown, G. Information Theoretic Feature Selection in Multi-label Data through Composite Likelihood. In *S+SSPR 2014*; Springer: Berlin/Heidelberg, Germany, 2014.
8. Brown, G.; Pocock, A.; Zhao, M.J.; Lujan, M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res. (JMLR)* **2012**, *13*, 27–66.
9. Sechidis, K.; Spyromitros-Xioufis, E.; Vlahavas, I. Multi-target feature selection through output space clustering. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 24–26 April 2019.
10. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
11. Tsoumakas, G.; Katakis, I.; Vlahavas, I. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2009; pp. 667–685.
12. Spolaôr, N.; Monard, M.C.; Tsoumakas, G.; Lee, H.D. A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing* **2016**, *180*, 3–15. [CrossRef]
13. Yang, Y.; Pedersen, J.O. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML), Nashville, TN, USA, 8–12 July 1997; pp. 412–420.
14. Lee, J.; Lim, H.; Kim, D.W. Approximating mutual information for multi-label feature selection. *Electron. Lett.* **2012**, *48*, 929–930. [CrossRef]
15. Chen, W.; Yan, J.; Zhang, B.; Chen, Z.; Yang, Q. Document transformation for multi-label feature selection in text categorization. In Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, USA, 28–31 October 2007; pp. 451–456.
16. Sechidis, K.; Sperrin, M.; Petherick, E.S.; Luján, M.; Brown, G. Dealing with under-reported variables: An information theoretic solution. *Int. J. Approx. Reason.* **2017**, *85*, 159–177. [CrossRef]

17. Lee, J.; Kim, D.W. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognit. Lett.* **2013**, *34*, 349–357. [CrossRef]
18. Lee, J.; Kim, D.W. Fast multi-label feature selection based on information-theoretic feature ranking. *Pattern Recognit.* **2015**, *48*, 2761–2771. [CrossRef]
19. Lee, J.; Kim, D.W. SCLS: Multi-label Feature Selection based on Scalable Criterion for Large Label Set. *Pattern Recognit.* **2017**, *66*, 342–352. [CrossRef]
20. Tsoumakas, G.; Katakis, I.; Vlahavas, I. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1079–1089. [CrossRef]
21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
22. Brillinger, D.R. Some data analyses using mutual information. *Braz. J. Probab. Stat.* **2004**, *18*, 163–182.
23. Tsoumakas, G.; Spyromitros-Xioufis, E.; Vilcek, J.; Vlahavas, I. Mulan: A Java Library for Multi-Label Learning. *J. Mach. Learn. Res.* **2011**, *12*, 2411–2414.
24. Zhang, M.L.; Zhou, Z.H. A k-nearest neighbor based algorithm for multi-label classification. In Proceedings of the IEEE International Conference on Granular Computing, Beijing, China, 25–27 July 2005; Volume 2, pp. 718–721.
25. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
26. García, S.; Fernández, A.; Luengo, J.; Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.* **2010**, *180*, 2044–2064. [CrossRef]
27. Fernandez-Lozano, C.; Gestal, M.; Munteanu, C.R.; Dorado, J.; Pazos, A. A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ* **2016**, *4*, e2721. [CrossRef]
28. Spyromitros-Xioufis, E.; Tsoumakas, G.; Groves, W.; Vlahavas, I. Multi-target regression via input space expansion: Treating targets as inputs. *Mach. Learn.* **2016**, *104*, 55–98. [CrossRef]
29. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
30. Sechidis, K.; Brown, G. Simple strategies for semi-supervised feature selection. *Mach. Learn.* **2018**, *107*, 357–395. [CrossRef]
31. Sechidis, K.; Papangelou, K.; Metcalfe, P.D.; Svensson, D.; Weatherall, J.; Brown, G. Distinguishing prognostic and predictive biomarkers: An information theoretic approach. *Bioinformatics* **2018**, *34*, 3365–3376. [CrossRef] [PubMed]