Information Theoretic Feature Selection in Multi-label Data through Composite Likelihood

Konstantinos Sechidis, Nikolaos Nikolaou, and Gavin Brown

School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{sechidik,nikolaon,gavin.brown}@cs.manchester.ac.uk

Abstract. In this paper we present a framework to unify information theoretic feature selection criteria for multi-label data. Our framework combines two different ideas; expressing multi-label decomposition methods as *composite likelihoods* and then showing how feature selection criteria can be derived by maximizing these likelihood expressions. Many existing criteria, until now proposed as heuristics, can be reproduced from a single basis under the proposed framework. Furthermore we can derive new *problem-specific* criteria by making different independence assumptions over the feature and label spaces. One such derived criterion is shown experimentally to outperform other approaches proposed in the literature on real-world datasets.

1 Introduction

The problem of learning from multi-label data becomes increasingly interesting because of the large number of applications in many different areas [15]. In computer vision [2], multi-label data are used in automated image and video annotation, in situations where images can be associated with a number of semantic concepts. In bioinformatics [5], multi-label learning is used in functional genomics, where a gene or protein is associated with multiple functional labels, as an individual gene or protein usually performs a number of functions. In text mining [8], multi-label data are used in text categorization, as a news webpage can be associated with more than one category.

All of these areas have a common characteristic, a large number of features. High dimensional feature spaces are associated with a number of problems, such as over-fitting to irrelevant features and high computational complexity. The features can be divided in three categories: features that are 'relevant' to our task, features that are 'irrelevant' and features that are 'redundant' in the context of other features. The objective of feature selection is to find a minimal subset of features that provide us with maximal useful information about the data. In our work we focus on *filter methods* for feature selection, which operate under the assumption that the prediction and feature selection steps are independent [7].

More particularly the present work focuses on information theoretic feature selection techniques in multi-label datasets, a problem that has recently received

P. Fränti et al. (Eds.): S+SSPR 2014, LNCS 8621, pp. 143–152, 2014.

a lot of attention [4,10,9]. The starting point of our work is a recently proposed framework for single label data by Brown et al. [3], which shows that many existing criteria can be seen as iterative maximizers of a common objective function: the conditional likelihood of the true label given the selected features. We extend this work by incorporating the idea of expressing multi-label decomposition methods via composite likelihood, as presented by Zhang & Schneider [16]; we show that this leads naturally to the derivation of different feature selection criteria appropriate for multi-label data. By introducing this framework we provide insights into multi-label feature selection.

There are two main contributions in our work. First, we provide a theoretical foundation that unifies various multi-label criteria proposed in the literature by maximizing full and composite likelihood expressions that describe different independence assumptions over the feature and label spaces (Sections 3-4). Second, we derive and evaluate new multi-label criteria which we compare with the state-of-the-art in real-world datasets (Section 5).

2 Reviewing Likelihood Maximization Framework

In this section we review the single-label feature selection framework presented by Brown et al. [3]. We assume that we have an underlying independent and identically distributed (i.i.d.) process $p : \mathcal{X} \to \mathcal{Y}$, and N samples of this process are observed. The observations are pairs $\{\mathbf{x}^i, y^i\}_{i=1}^N$, where the features are d-dimensional vectors $\mathbf{x}^i = [x_1^i ... x_d^i]$. The features are drawn from the random variables $X_1, ..., X_d$, with their joint distribution being $X = X_1 X_2 ... X_d$ and the labels are drawn from the random variable Y. Following Brown et al. [3], in the feature selection procedure we define θ to be a d-dimensional binary vector, where the elements have a value of 1 if the feature is selected and 0 otherwise. Furthermore $\mathbf{x}_\theta$ is the vector of the chosen and $\mathbf{x}_{\tilde{\theta}}$ the vector of the unchosen features. We assume that the process p can be defined by a subset of features and so for an optimal vector θ^* we have $p(y|\mathbf{x}) = p(y|\mathbf{x}_{\theta^*})$, in other words the unselected features $\mathbf{x}_{\tilde{\theta}^*}$ are irrelevant or redundant given the selected ones. We approximate the process p using a hypothetical predictive model f. This model has two layers of parameters: θ, corresponding to the selected features, and τ, corresponding to the parameters used in the learning procedure in order to predict y. So the problem can be defined as searching for a minimal subset of features, whilst maximizing the conditional likelihood of the training labels. For single-label data the conditional likelihood ($\mathcal{L}$) and the *log*-likelihood (ℓ) have the form:

$$\mathcal{L}(\theta, \tau; y|\mathbf{x}) = \prod_{i=1}^{N} f(y^i|\mathbf{x}_\theta^i, \tau) \Leftrightarrow \ell(\theta, \tau; y|\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \log f(y^i|\mathbf{x}_\theta^i, \tau).$$

Brown et al. [3] showed that this likelihood decomposes as so:

$$\lim_{N\to\infty} -\ell = E_{XY}\left\{\log \frac{p(y|\mathbf{x}_\theta)}{f(y|\mathbf{x}_\theta, \tau)}\right\} + I(X_{\tilde{\theta}}; Y|X_\theta) + H(Y|X).$$

From the above three terms, the first describes how well the model f approximates p given the selected features, the second term depends on the choice of the selected features, while the third term is an irreducible constant which forms a bound on the Bayes error rate. More details regarding this decomposition can be found in Brown et al. [3]. The main assumption of filter methods is that the classification and the feature selection steps are independent [7]. Under this assumption and ignoring the constant term, the value of θ that maximizes the conditional likelihood is the same as the value of θ that minimizes the conditional mutual information

$$\arg\max_{\theta} \mathcal{L}(\theta; y|\mathbf{x}) = \arg\min_{\theta} I(X_{\tilde{\theta}}; Y|X_{\theta}). \tag{1}$$

As we see in Brown et al. [3] a greedy optimization process to minimize the conditional mutual information in eq. (1) will select a feature X_k that maximizes the following scoring function

$$J_{CMI}(X_k) = I(X_k; Y|X_{\theta}) \quad \text{with} \quad X_k \in X_{\tilde{\theta}}, \tag{2}$$

where the subscript *CMI* stands for *Conditional Mutual Information*.

Since X_{θ} is high-dimensional, the estimates of the mutual information become less reliable as we increase the number of selected features; this can lead to poorly selected subsets. For that reason there have been proposed in the literature low-dimensional approximations of this conditional mutual information, such as the *Mutual Information Maximization (MIM)* [1] and the *Joint Mutual Information maximization (JMI)* [12]. The respective criteria are given by

$$J_{MIM}(X_k) = I(X_k; Y), \qquad J_{JMI}(X_k) = \sum_{j=1}^{|X_{\theta}|} I(X_{\theta_j} X_k; Y),$$

where we used the notation $X_{\theta_j}\ \forall\ j \in \{0, ..., |X_{\theta}|\}$ to represent the j^{th} feature already selected, while $|X_{\theta}|$ is the number of selected features so far. As we can see the *MIM* criterion selects the features independently and so it has the ability to observe relevant features, but not to detect redundant ones. On the other hand the *JMI* also controls the redundancy of the selected features, as it examines the joint random variable $X_{\theta_j} X_k$. Brown et al. [3] present the assumptions made by each approximation, and derive these criteria from first principles by incorporating the assumptions in eq. (2). Furthermore they show in a large empirical study that assuming independence in the feature space (i.e. with *JMI*/*MIM*) has major benefits over the full dependence case of *CMI*. In the following section we will extend the above framework to multi-label data, exploring independence assumptions in the *label* space.

3 Extending the Framework to Multi-label Data

The key difference between single and multi-label classification is that in binary single-label classification, for example, the label space $\mathcal{Y}$ is $\{0, 1\}$, while in multi-label classification the space $\mathcal{Y}$ is $\{0, 1\}^q$ where q represents the number of labels.

The labeling of the i-th instance is a q-dimensional binary vector $\mathbf{y}^i = [y_1^i \dots y_q^i]$, with $y_l^i = 1$ if the example i is positive to the label l and $y_l^i = 0$ if it is negative. The labels are drawn from the random variables $Y_1, \dots, Y_q$ with their joint distribution denoted $Y_{1:q}$.

3.1 Label-Powerset Transformation

When learning from multi-label data, the most general approach is to not assume any label independencies [15]. This transforms the multi-label problem into a multi-class single label one by combining each different label set into a different "meta-class". This approach is known as the *Label Powerset (LP)* transformation, and the maximum number of classes is 2^q. Figure 1 represents the probabilistic graphical model for *LP* transformation, according to the framework presented in Section 2.

The framework presented in Section 2 can be extended to multi-label data just by substituting the single label output variable Y with the multi-label joint random variable $Y_{1:q}$. By making this substitution we arrive at the following multi-label filter:

$$J_{CMI}^{LP}(X_k) = I(X_k; Y_{1:q}|X_\theta). \quad (3)$$

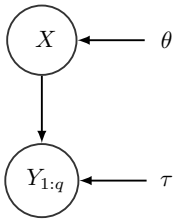


Fig. 1. Label-powerset transformation

The superscript *LP* denotes the assumption over the label space (i.e. none) and the subscript *CMI* stands for the assumptions over the feature space (i.e. also in this case, none). Using the chain rule of mutual information $I(X_k X_\theta; Y_{1:q}) = I(X_\theta; Y_{1:q}) + I(X_k; Y_{1:q}|X_\theta)$ we rewrite the *CMI* criterion as

$$X_k = \arg\max_{X_k \in X_{\bar{\theta}}} I(X_k; Y_{1:q}|X_\theta) = \arg\max_{X_k \in X_{\bar{\theta}}} I(X_k X_\theta; Y_{1:q}),$$

which is *exactly* the multi-label criterion heuristically proposed by Doquire & Verleysen [4]. In our work we derived this criterion by maximising an explicit objective function: the conditional likelihood of the training labels under the probabilistic model presented in Figure 1.

3.2 Binary-Relevance Transformation

The number of distinct label combinations is 2^q, increasing exponentially with the number of labels. Thus we need a large amount of data to have reliable estimates for the probabilities under the *LP* transformation. There have been proposed various transformation approaches to deal with this problem, a detailed exposition of these can be found in Zhang & Zhou [15]. The simplest transformation is to ignore any dependencies between the labels and predict each label independently, this method is known as *Binary Relevance (BR)* or one-vs-all transformation. The graphical model for the *BR* transformation can be seen in

Figure 2. The conditional likelihood for this model, which has been given in Zhang & Schneider [16] in the context of *composite likelihood*, has the form

$$\mathcal{L}_{BR}(\theta, \tau; \mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} \prod_{l=1}^{q} f(y_l^i | \mathbf{x}_\theta^i, \tau_l).$$

By maximising this likelihood and following the same procedure as in Section 2 we can derive the following multi-label criterion

$$J_{CMI}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l | X_\theta). \quad (4)$$

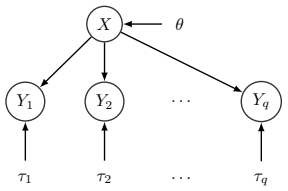


Fig. 2. Binary-relevance transformation

Again, the superscript BR represents the assumption of the conditionally independent labels, while the subscript represents the assumptions made in the feature space. This can be seen as the BR version of eq. (3); to the best of our knowledge this is the first time this has been proposed in the literature.

At this point, we can observe two types of assumption behind different criteria – in the feature space, and in the label space. Table 1 represents the different types of assumption we explore in this work. From now on we will follow this notation to describe the criteria.

Table 1. Choices in the design of multi-label feature selection criteria

		Feature space independence assumptions		
		CMI (none)	JMI (partial)	MIM (full)
Label space independence assumptions	Label Powerset (none)	$J^{\text{Y:none}}_{\text{X:none}}$	$J^{\text{Y:none}}_{\text{X:partial}}$	$J^{\text{Y:none}}_{\text{X:full}}$
	Binary Relevance (full)	$J^{\text{Y:full}}_{\text{X:none}}$	$J^{\text{Y:full}}_{\text{X:partial}}$	$J^{\text{Y:full}}_{\text{X:full}}$

In the following section we will make no assumptions on feature space, and explore the effect of label space assumptions. We will thus compare the criteria described by eqs. (3) and (4), in our new notation $J^{\text{Y:none}}_{\text{X:none}}$ and $J^{\text{Y:full}}_{\text{X:none}}$ respectively.

3.3 Empirical Comparison of the Assumptions in the Label Space

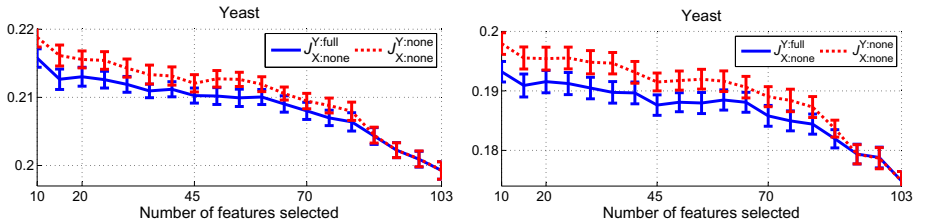
The experiments are performed on real-world multi-label datasets — *yeast* [5] and *scene* [2], taken from two characteristic applications for multi-label data: biology and computer vision, respectively. These two datasets are used by both Doquire & Verleysen [4] and Lee & Kim [9] to evaluate their criteria, with which we compare our own in Section 5. Table 2 summarises some characteristics of these datasets. In order to compare the different feature selection techniques we use a nearest neighbor multi-label classifier, ML-kNN with $k = 7$ as suggested

Table 2. Characteristics of the datasets

Name	Application	Examples	Features	Labels	Distinct labelsets
Scene	Computer Vision	2407	294	6	15
Yeast	Bioinformatics	2417	103	14	198

in Zhang & Zhou [14]. We chose a k-nearest neighbor classifier, since it makes few assumptions and it does not perform implicitly any sort of feature selection, as all the features have the same weight. We evaluate our techniques using two different loss functions: hamming loss and ranking loss [15]. Since in multi-label classification the evaluation is a complex task, we chose these two representative measures. We perform 30 random splits of the data into 50% training and 50% testing, reporting averages and 95% confidence intervals. The training data was used for selecting features and training the ML-kNN classifier, while the testing for measuring the performance of the different approaches. To estimate the mutual information we use maximum likelihood estimates, discretising continuous features into 5 bins using an equal width strategy.

In Figure 3 we compare criteria derived from different label space assumptions, making no assumptions in the feature space. The goal is to investigate the effect that the independence assumptions made on the label space have on the quality of the selected feature subset. As we can see the BR version $\left(J^{\text{Y:full}}_{\text{X:none}}\right)$ very marginally outperforms the LP version $\left(J^{\text{Y:none}}_{\text{X:none}}\right)$ for yeast dataset. This reflects that for yeast, a dataset with large number of distinct labelsets, the benefits of the conditional independence assumption regarding the labels (better probability estimates) outweigh its drawbacks (ignoring inter-label interactions). The effect is slightly more pronounced in the case of ranking loss. Naturally, the difference in performance decreases as we increase the number of selected features. We omit the figures for the scene dataset since both approaches have similar performance, and there is no statistically significant difference between the two criteria. Thus in both datasets, *the quality of the selected feature subset is not significantly affected by the different independence assumptions in the label space.* Since by increasing the number of selected features X_θ the estimates of the conditional mutual information in eq. (3) and (4) degrade, it will be interesting to explore how feature space independence assumptions help the situation.



(a) Hamming loss (b) Ranking loss

Fig. 3. Comparing criteria derived from different label space assumptions and making no feature space assumption. Y:none indicates the LP transformation, while Y:full indicates the BR transformation.

4 Criteria under Different Feature Space Assumptions

The previous section investigated independence assumptions in the label space. In this section we will explore assumptions on the feature space. While in Section 2 we reviewed the *CMI*, *JMI* and *MIM* criteria in the context of single label data, in the current section we will present how the approximate criteria *MIM* and *JMI* are converted in the multi-label context.

4.1 *MIM* and *JMI* Criteria under *LP* Transformation

Under the *LP* transformation, and following a similar procedure to that of Section 2, eq. (3) can be approximated by the lower-order criteria

$$J^{\text{Y:none}}_{\text{X:full}}(X_k) = I(X_k; Y_{1:q}), \quad (5) \qquad J^{\text{Y:none}}_{\text{X:partial}}(X_k) = \sum_{j=1}^{|X_\theta|} I\left(X_k X_{\theta_j}; Y_{1:q}\right). \quad (6)$$

The $J^{\text{Y:none}}_{\text{X:full}}$ criterion has been proposed heuristically by Spolaôr et al. [10].

4.2 *MIM* and *JMI* Criteria under *BR* Transformation

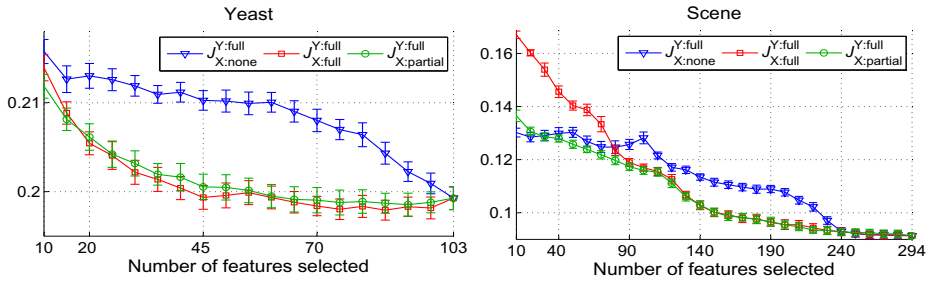
Under the *BR* transformation and following the framework presented in Section 2, eq. (4) can be approximated by the lower-order criteria

$$J^{\text{Y:full}}_{\text{X:full}}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l), \quad (7) \qquad J^{\text{Y:full}}_{\text{X:partial}}(X_k) = \sum_{j=1}^{|X_\theta|} \sum_{l=1}^{q} I\left(X_k X_{\theta_j}; Y_l\right). \quad (8)$$

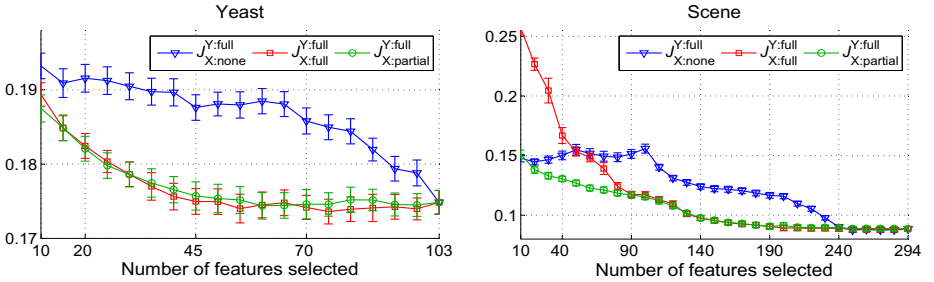
Clearly $J^{\text{Y:full}}_{\text{X:full}}$, makes the most strict assumptions, full independence of both features and labels. As such, this has been suggested heuristically in numerous works [10,11,13]. Here we have shown that this can be derived as an approximate maximizer of the composite likelihood of the model in Figure 2. However this is the first time that *JMI* criteria, such as $J^{\text{Y:none}}_{\text{X:partial}}$ and $J^{\text{Y:full}}_{\text{X:partial}}$, are introduced in the multi-label setting.

4.3 Empirical Comparison of the Assumptions in the Feature Space

Figure 4 compares criteria derived from different feature space assumptions under the same experimental setup we used in Section 3. This comparison was performed under the *BR* transformation but the results under the *LP* transformation are similar. The goal now is to investigate which independence assumption on the feature space gives the best feature selection results. We see that the JMI criterion $\left(J^{\text{Y:full}}_{\text{X:partial}}\right)$ outperforms the other two approaches as it consistently achieves good performance for both datasets. On the yeast dataset the JMI and MIM perform similarly, and we can draw the same conclusion as in Doquire & Verleysen [4], i.e. that the relevant features are non-redundant in this data. On the scene dataset JMI outperforms the other criteria for almost any number of selected features in the cases of hamming loss, and for any number of selected features for ranking loss.



(a) Hamming loss



(b) Ranking loss

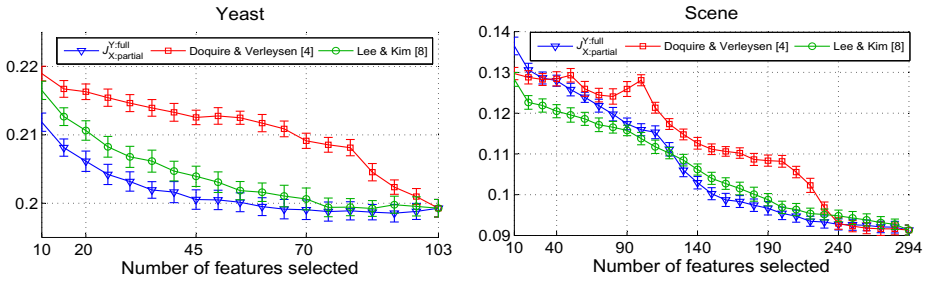
Fig. 4. Comparing criteria derived from different feature space assumptions and assuming full independence in the label space. X:none indicates the CMI criterion, X:partial the JMI and X:full the MIM.

5 Summary and Connections to Literature

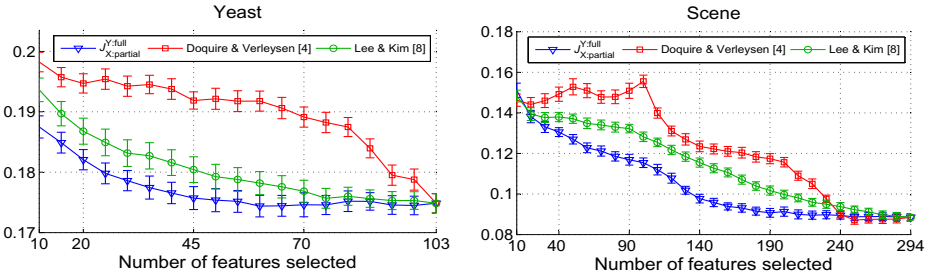
In Section 3.3 we examined the effect of the label space assumptions on the feature selection process and found that *BR* has a marginal advantage over *LP*. In Section 4.3 we investigated the effect of the feature space assumptions and observed an advantage of *JMI* over *CMI* and to a lesser extent over *MIM*. In this section we connect our work with the literature, and we compare the criterion with the best performance under our analysis, with the state-of-the-art in multi-label feature selection.

5.1 Connections with the Literature

Yang & Pedersen [13] introduced the first multi-label feature selection criteria which can be classified as the $J^{\text{Y:full}}_{\text{X:full}}$ of our analysis. Trohidis et al. [11] present a comparison between the $J^{\text{Y:full}}_{\text{X:full}}$ and $J^{\text{Y:none}}_{\text{X:full}}$ criteria, but using χ^2-statistic instead of mutual information, while recently these criteria were re-introduced under the problem transformation approach [10]. Doquire & Verleysen [4] proposed $J^{\text{Y:none}}_{\text{X:none}}$. In order to produce better estimates they use a nearest neighbor mutual information estimator and they apply the pruned problem transformation technique, under which the rare label combinations are discarded, and as a consequence this leads to some loss of information. Finally, Lee & Kim [9] propose the use of multivariate mutual information for selecting features in a criterion without applying



(a) Hamming loss



(b) Ranking loss

Fig. 5. Comparing the $J^{\text{Y:full}}_{\text{X:partial}}$ criterion with criteria proposed in the literature

any transformation, but since this method is computationally inefficient they propose an approximate solution which involves only three variables.

5.2 Comparison to the State-of-the-Art

We compare $J^{\text{Y:full}}_{\text{X:partial}}$, the criterion with the best performance under our analysis, with two different criteria proposed recently in the literature: the pruned transformation criterion proposed by Doquire & Verleysen [4] (we prune rare examples using thresholds given in that work) and the multi-variate mutual information criterion proposed by Lee & Kim [9]. As we can see in Figure 5 the proposed criterion $J^{\text{Y:full}}_{\text{X:partial}}$ consistently performs well across the different number of selected features and the different datasets. On the yeast dataset it has the best performance for both loss functions and all numbers of selected features. On the scene dataset it outperforms the other techniques in all areas apart from 10-130 selected features under hamming loss. However, in terms of ranking loss it continuously outperforms the other criteria.

6 Conclusions

We have provided a theoretical justification for multi-label feature selection criteria. Our framework introduces the idea of *maximizing the conditional composite likelihood expression for multi-label decompositions*. Different assumptions lead

naturally to different filters, some of which have been heuristically proposed in the literature, while others are novel. In our experiments we explored how different assumptions of feature/label space compare. The best trade-off appears to be assuming partial independence in feature space, and full independence in label space. Our observation regarding the label space assumptions agrees with recent empirical results in the context of wrapper feature selection [6]. The corresponding filter we propose is shown to outperform the state-of-the-art approaches on real-world datasets. Finally, under this framework we can incorporate assumptions that explicitly encode domain knowledge, leading to filters specialised for particular problems.

Acknowledgments. This work was supported by EPSRC grant [EP/I028099/1]. Sechidis gratefully acknowledges the support of the Propondis Foundation.

References

1. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 5(4), 537–550 (Jul 1994)
2. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757 – 1771 (2004)
3. Brown, G., Pocock, A., Zhao, M., Lujan, M.: Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Journal of Machine Learning Research (JMLR) 13, 27–66 (2012)
4. Doquire, G., Verleysen, M.: Mutual information-based feature selection for multilabel classification. Neurocomputing 122, 148 – 155 (2013)
5. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in Neural Inf. Processing Systems (NIPS) 14. pp. 681–687 (2001)
6. Gharroudi, O., Elghazel, H., Aussem, A.: A comparison of multi-label feature selection methods using the random forest paradigm. In: Adv. in Artificial Intelligence, Lecture Notes in Computer Science, vol. 8436, pp. 95–106. Springer (2014)
7. Guyon, I.M., Gunn, S.R., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction: Foundations and Applications. Springer, 1st edn. (2006)
8. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: ECML/PKDD Workshop on Discovery Challenge (2008)
9. Lee, J., Kim, D.W.: Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters 34(3), 349 – 357 (2013)
10. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science 292, 135 – 151 (2013)
11. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. In: 9th Int. Conf. on Music Inf. Retrieval (ISMIR) (2008)
12. Yang, H.H., Moody, J.: Data visualization and feature selection: New algorithms for nongaussian data. In: Advances in Neural Inf. Processing Systems (NIPS) (1999)
13. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: 14th Int. Conference on Machine Learning (ICML) (1997)
14. Zhang, M.L., Zhou, Z.H.: A k-nearest neighbor based algorithm for multi-label classification. In: IEEE International Conference on Granular Computing (2005)
15. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering (in press) (2013)
16. Zhang, Y., Schneider, J.: A composite likelihood view for multi-label classification. In: 15th Int. Conference on Artificial Intelligence and Statistics (AISTATS) (2012)