Extending Multi-Label Feature Selection with KEGG Pathway Information for Microarray Data Analysis

Suwimol Jungjit
School of Computing
University of Kent, Canterbury, CT2 7NF, UK
sj290@kent.ac.uk

Alex A. Freitas
School of Computing
University of Kent, Canterbury, CT2 7NF, UK
A.A.Freitas@kent.ac.uk

M. Michaelis
School of Biosciences
University of Kent, Canterbury, CT2 7NJ, UK
M.Michaelis@kent.ac.uk

J. Cinatl
Institut fuer Medizinische Virologie,
Klinikum der Goethe-Universitaet, Paul Ehrlich-Str. 40
60596 Frankfurt am Main, Germany
cinatl@em.uni-frankfurt.de

Abstract— We propose three approaches to extend our previous Multi-Label Correlation-based Feature Selection (ML-CFS) method with cancer-related KEGG pathway information, in order to select a better set of genes (features) for cancer microarray data classification. In the approach which produced the best results, ML-CFS was extended with a weighted formula that combines genes' predictive power and occurrence in cancer-related KEGG pathways as criteria for gene selection. We also investigated the effect of different weights for those two criteria. That approach obtained, in general, a statistically significantly smaller hamming loss (i.e. higher predictive accuracy) when compared to the hamming loss obtained by ML-CFS without using KEGG pathway information, in two cancer-related microarray datasets, using two different multi-label classification algorithms – one based on neural networks, the other based on nearest neighbors. In addition to significantly improving predictive performance, the genes selected by that approach were found to be more biologically relevant to the analysis of our datasets than genes selected without using KEGG pathway information**.** To the best of our knowledge, this is the first paper to propose a KEGG pathway-based feature selection method for multi-label classification.

Keywords — multi-label feature selection, multi-label classification, cancer-related microarray data, neuroblastoma, KEGG pathway.

I. INTRODUCTION

DNA microarray is a technology which is widely used to study biomedical samples [1], since it allows us to measure the gene expression levels of a large number of genes simultaneously in a given type of cell or tissue. A common data mining task applied to microarray data analysis is classification, which is the focus of this paper. In this task, a microarray dataset is regarded as a set of instances, where each instance consists of two parts: a set of predictor features (gene expression values) and a special class variable (whose value is to be predicted by the classification algorithm).

More precisely, in the two DNA microarray datasets used in this paper, each instance represents a cancer cell line, there are more than 20,000 features (genes) and the class variables to be predicted are binary variables indicating whether or not each cell line is sensitive or resistant to certain drugs.

A classification algorithm uses a training set where both feature values and the class values are known by the algorithm to build a classification model, which is capable of predicting the class of an instance based on its feature values [2]. The model is then applied to a testing set, consisting of instances whose class value is unknown by the algorithm.

In the context of microarray datasets, the main challenge for a classification algorithm is that the number of features (genes) is very large, whilst the number of instances is very small. A common approach to this problem is to apply a feature selection method in a preprocessing phase [3] – i.e., before applying a classification algorithm to the data – in order to select a small subset of relevant features for microarray data classification. In this paper we follow this general approach in the context of multi-label classification.

Multi-label classification is special (and more difficult) type of classification task in data mining, where each instance can be associated with a set of class labels, rather than just one class label as in conventional single-label classification [4].

In this work we focus on multi-label classification of two DNA microarray datasets, and we propose to extend a recently proposed multi-label correlation-based feature selection method [5]. The proposed extension is based on incorporating biological knowledge in the feature selection process. That is, while conventional feature selection methods select features based on their predictive power, our extended feature selection method selects features (genes) based on both their predictive power and their occurrence in cancer-related KEGG pathways [6]. Hence, the feature selection process is intentionally biased towards the selection of known cancer-related genes, whilst still focusing on selecting genes with a good predictive power.

In essence, we propose three different versions of an extended multi-label feature selection method, incorporating KEGG pathway information in the method in three different approaches. The first approach uses a weighted formula to evaluate the merit (quality) of a candidate gene subset, where the criteria of predictive accuracy and occurrence in cancer-related KEGG pathways are assigned (user-defined) weights indicating their relative importance. In the second approach, information about the occurrence of a gene in cancer-related KEGG pathways is embedded into the merit function, in a way that avoids the need for user-specified numerical weights. In the last approach, the feature selection method selects only a subset of genes that occur in cancer-related KEGG pathways, removing all genes that do not occur in those pathways.

We run experiments on two multi-label microarray datasets, and evaluate the feature (gene) subsets selected by our proposed extended multi-label feature selection method using two well-known multi-label classification algorithms: ML-kNN [7] and ML-RBF [8]. For each dataset, we discuss the predictive accuracy associated with the different versions of the multi-label feature selection method and discuss the biological relevance of the genes most frequently selected by that method.

The rest of this paper is organized as follows. Section II gives an overview of feature selection both in the single-label and the multi-label scenarios. Section III introduces the three proposed extensions to a multi-label feature selection method. Section IV reports the computational results. Section V concludes the paper and mentions future work.

II. Background on Feature selection

A. Conventional, Single-Label Feature Selection

Feature selection is a process which selects a relevant feature subset according to an evaluation criterion(a) [3]. The main objectives of feature selection are to avoid model overfitting, improve the predictive performance of the model, eliminate irrelevant features, and reduce the computational time taken by the classification algorithm [3, 9, 10]. Feature selection methods that are applied in a data preprocessing phase (before applying a classification algorithm to the data) can be classified into two approaches, as follows. The first one is the filter approach, where the evaluation function used to measure the quality of a feature subset is independent of the classifier. This approach is usually fast and scalable to datasets with very large number of features. This approach was used in [11] and [12]. Moreover, Lui et al [13] pointed out that the filter approach is the most widely used in real-world applications, particularly when classifying datasets with a very large number of features, such as microarray datasets. The structure of most filter algorithms is very simple, and it provides a simple way to calculate the relevance of features in large-scale data in a relatively short time.

The second approach, named the wrapper approach, is more complex. In this approach the evaluation function used to measure the quality of a feature subset consists of a measure of the predictive accuracy of a classification algorithm applied to that feature subset. This approach selects feature subsets customized to a given classification algorithm, which tends to improve predictive accuracy. However, when using the wrapper approach there is a risk that the classification model overfits the training data (reducing predictive performance in the testing set) [9, 10]; and the wrapper approach is usually much more computationally expensive than the filter approach – due to the need to run a full classification algorithm for each evaluated candidate feature subset.

In this paper we focus on the filter approach, due to its more natural scalability to datasets with a very large number of features, like the microarray dataset mined in this work.

B. Multi-Label Feature Selection

There are very few papers on multi-label feature selection methods applied to microarray data [14], [5], [15] in contrast to the very large number of papers on traditional single-label feature selection methods applied to microarray data [9]. Note that [15] used a wrapper approach, whilst we focus on the filter approach. Next, we briefly review multi-label feature selection methods following the filter approach in general (not in bioinformatics). The work in [14, 5] will be described in subsection C.

Doquire and Verleysen [16], as well as Spolaor et al. [17], essentially transformed the multi-label dataset into a single-label one and then applied a single-label feature selection method to the transformed data. The disadvantage of their approach is that it cannot deal with the multi-label problem directly, while our approach directly copes with the original multi-label data. Another method proposed by Spolaor [18] selects feature subsets which have a multi-label information gain (IG) value greater than or equal to a pre-defined threshold. This method has the drawback of requiring an *ad-hoc* user-defined threshold value. Lastra et al [19] extended the single-label feature selection method proposed by Yu and Lui [12] to multi-label classification. Their method has a serious drawback in the context of our datasets, namely, it requires all continuous data to be discretized in a preprocessing step. Our microarray datasets have more than 20,000 continuous features, and the discretization of so many features would probably lead to a loss of relevant information.

Note that all aforementioned multi-label feature selection methods evaluate a candidate feature subset based on their features' predictive power; none of them combines predictive power calculations with biological knowledge – in order to try to select genes which are more biologically relevant. By contrast, our proposed extensions to a multi-label feature selection method, described in section III, use information from cancer-related KEGG pathways to favor the selection of biologically-relevant (cancer-related) genes, whilst still considering the predictive power of candidate features (genes). Note also that there are some papers which employ information from KEGG pathways for feature selection, in particular [20, 21], but such works address only the conventional single-label classification task, rather than the more complex multi-label classification task addressed in this paper.

C. Multi-Label Correlation Based Feature Selection (ML-CFS)

The ML-CFS method was first proposed in [14] and later extended in [5], and it is briefly described here, to make this current paper more self-contained – since this current paper extends the original ML-CFS with biological knowledge, as will be discussed later. ML-CFS is essentially an adaptation of the single-label Correlation-based Feature Selection (CFS) method proposed in [11] for the more complex type of multi-label classification problems.

In essence, the CFS method uses a hill-climbing search to try to find a feature (gene) subset *F* that has two properties:

(a) it maximizes the correlations between the features in *F* and the set of class labels *L* ($\overline{r_{FL}}$ in equation (1)), in order to select features with high predictive accuracy; and

(b) it minimizes the correlations between pairs of features ($\overline{r_{FF}}$ in equation (1)) in *F*, in order to avoid the selection of redundant features.

Both the single-label CFS and ML-CFS use equation (1) – where *k* is the number of features in a candidate feature subset *F* and *r* is Pearson's linear correlation coefficient – to measure the quality of a candidate feature subset, but they use a different approach for measuring the average correlation between features and labels ($\overline{r_{FL}}$). More precisely, ML-CFS computes the average correlation coefficient ($\overline{r_{FL}}$) between each feature in feature set F and each of the multiple class labels in label set L, using equation (2); and then averages the result of equation (2) over all features, as shown in equation (3). On the other hand, the average correlation value between features and label in the conventional single-label CFS method is simpler, because there is no need to measure average correlations over multiple class labels.

$$Merit = \frac{k\overline{r_{FL}}}{\sqrt{k+k(k-1)\overline{r_{FF}}}} \quad (1)$$

$$\overline{r_{f\bar{L}}} = \frac{\sum_{i=1}^{|L|} r_{fL_i}}{|L|} \quad (2)$$

$$\overline{r_{FL}} = \frac{\sum_{f=1}^{|F|} r_{f\bar{L}}}{|F|} \quad (3)$$

Two extensions of ML-CFS were proposed in [5]: one based on using the absolute value of the correlation coefficient, and another based on using mutual information between class labels. The former improved the predictive performance of ML-CFS, unlike the later, so we focus here only on the extended version of ML-CFS using absolute correlation function (i.e., without using mutual information), which is the version used in all experiments reported in Section IV.

In that version, the terms in the merit formula were modified to use the absolute (without sign) value of the correlation coefficient, as shown in equations (4) and (5), which compute the average correlation between all feature pairs ($\overline{r_{FF}}$) and the average correlation between features and class labels ($\overline{r_{FL}}$), respectively. In equation (4), *fp* is the number of feature pairs in feature subset *F*. Note that $|r_{f_if_j}|$ and $|r_{f\bar{L}}|$return a value in [0..+1], rather than in [–1..+1].

$$\overline{r_{FF}} = \frac{\sum_{f_if_j=1,i\neq j}^{|F|} |r_{f_if_j}|}{fp} \quad (4)$$

$$\overline{r_{FL}} = \frac{\sum_{f=1}^{|F|} |r_{f\bar{L}}|}{|F|} \quad (5)$$

The motivation for using the absolute value of the correlation coefficient, rather than the original, signed valued of the correlation coefficient, is discussed in detail in [5], which is an issue not very relevant for this current paper's contribution (focusing on extending ML-CFS with biological knowledge). The important point here is that both the original ML-CFS [14] and the ML-CFS using absolute value of the correlation coefficient only measure the merit of a candidate feature subset from a statistical perspective, computing the predictive power and the redundancy in a feature subset. No previous version of ML-CFS uses any form of biological knowledge to evaluate the quality of a candidate feature (gene) subset. Hence, the main contribution of this paper is to extend ML-CFS with biological knowledge, as discussed in the next Section.

III. Extending Multi-Label Correlation-based Feature Selection (ML-CFS) with Cancer-Related KEGG Pathway Information

Recall that the original ML-CFS method (described in Section II-C) evaluates the quality of a candidate feature subset by using a Merit function, which rewards features that are highly correlated with the class attributes and have a low degree of redundancy. Hence, the Merit function does not incorporate any biological knowledge about cancer-related genes. In this work we propose to extend the ML-CFS method with an evaluation function that uses some biological knowledge about cancer-related pathways.

Intuitively, the use of such biological knowledge would allow the ML-CFS method's search to focus on genes which are already known to be cancer-related, which could help to improve the predictive performance of the ML-CFS method or help to select genes whose role in cancer-related drug resistance/sensitivity is more likely to be meaningful to biologists.

More precisely, we use knowledge about cancer-related KEGG pathways (http://www.gnome.jp/keg/pathway.html) [20-21], which is a well-known pathway database, as part of the function that evaluates a candidate feature subset. KEGG pathways represent the interaction between genes (gene-gene interactions) in a gene map using different components. Moreover, it covers a wide range of organisms and is easy to use because each pathway is stored in well-known formats such as XML format files, text files and so on.

Note that we utilize only 16 cancer-related KEGG pathways, which were selected based on current knowledge about the biology of cancer, because our experiments aim to select genes which are relevant for predicting drug sensitivity/resistance in cancer patients. So, it would not be effective to employ all pathways in the KEGG database.

The selected 16 cancer-related KEGG pathways are: DNA replication, Base excision repair, Nucleotide excision repair, Mismatch repair, Homologous recombination, Non-homologous end-joining, Fanconi anemia pathway, ABC transporters, Wnt signaling pathway, Notch signaling pathway, Hedgehog signaling pathway, Cell cycle, Apoptosis, p53 signaling pathway, Pathways in cancer and Transcriptional misregulation in cancer. Detailed information about these cancer-related pathways is provided on the KEGG website (http://www.genome.jp/kegg/). We assume that if some genes are related with cancer-related drug resistance/sensitivity, they are likely to occur in some of the above cancer-related pathways.

In order to quantify the strength of the relationship between the genes in a candidate feature subset and the aforementioned cancer-related pathways, we propose to compute "the average relative frequency of pathways per gene ($AvgRFP_{FSS_i}$)":

$$AvgRFP_{FSS_i} = \frac{\sum_{f=1}^{k} RFP_f}{k} \quad (6)$$

where the average is computed over all the k features selected in the *i*th candidate feature subset (*FSS$_i$*), as shown in formula (6). For each selected feature *f* in *FSS$_i$*, the relative frequency of pathways for *f*, denoted by *RFP$_f$*, is the number of cancer-related KEGG pathways in which the gene corresponding to *f* occurs divided by the number of user-specified pathways (16 in our case). Each *RFP$_f$* has a value in [0..1], so $AvgRFP_{FSS_i}$ also has a value in [0..1]. Hence, the $AvgRFP$ term rewards feature subsets where most genes in the subset are involved in several cancer-related pathways, and penalizes feature subsets where most genes do not occur in any cancer-related pathway.

This work proposes three different extended versions of the ML-CFS method, which use KEGG pathway information in three different ways, as follows:

(1) Using a weighted formula to combine the Merit function and KEGG pathway information.

(2) Embedding KEGG pathway information into the Merit function.

(3) Selecting only genes that occur in KEGG pathways.

A. Using a Weighted Formula to Combine the Merit Function and KEGG Pathway Information.

In this approach, the evaluation function of the *i*th *FSS* is defined by the following weighted formula:

$$Evaluation\,function\; = \;\alpha * Merit_{FSS_i} + \boldsymbol{\beta} * AvgRFP_{FSS_i} \quad (7)$$

where $\alpha\; and\; \beta$ are weights in [0..1] – α is a user-defined parameter, and $\beta = 1 - \alpha$. $Merit_{FSS_i}$ and $AvgRFP_{FSS_i}$ were discussed earlier.

The advantage of this approach is its simplicity: it computes the value of the merit of a candidate feature subset and its *AvgRFP* value separately (representing two different perspectives, one statistical and another biological, respectively). More precisely, the merit function evaluates candidate feature subsets using the concept of statistical correlation; while *AvgRFP* evaluates candidate feature subsets in terms of how often the genes in a feature subset occur in cancer-related KEGG pathways. An important point of our experiments is that the weight α assigned to the merit function ($Merit_{FSS_i}$) is greater than or equal to the weight $\boldsymbol{\beta}$ assigned to *AvgRFP*. This is because we consider the predictive accuracy (evaluated by the merit function) as the primary evaluation criterion of a feature subset, while *AvgRFP* is a secondary (but still important to users) criterion supporting the discovery of biologically relevant features. There is no point in discovering biologically relevant features with low accuracy.

B. Embedding KEGG Pathway Information into the Merit Function.

We also tried to embed the value of *RFP* into the merit function in order to avoid the need to specify user-defined weights ($\alpha\; and\; \beta$) in our evaluation function. In this approach, the formula to calculate the average value of the correlation between all features in a feature subset *F* and all the labels in class label set *L* is different from the formula in the original ML-CFS. The new formula is as follows:

$$\overline{r_{FL}} = \frac{\sum_{f=1}^{|F|} \left| r_{f\bar{L}} \right| * RFP_f}{\sum_{f=1}^{|F|} RFP_f} \quad (8)$$

The idea behind this formula is that we want to reward the feature-label correlation values in proportion to the strength of the association between the genes in a feature subset and the cancer-related KEGG pathways (as measured by the *RFP* term), while the average correlation between pairs of features in a feature subset (to detect redundancy) is computed in the same way as in the original ML-CFS algorithm.

The effect of using this formula with the hill climbing search used by ML-CFS is that the algorithm will select only genes which occur in some KEGG pathway in the first iteration of hill climbing search. This is because in the first iteration of the search each candidate feature subset contains just one feature (gene), and if that gene does not occur in any KEGG pathway the value of $\overline{r_{FL}}$ is equal to zero. In that case the value of the merit function is equal to zero because in the first iteration the average correlation between feature pairs in the feature subset ($\overline{r_{FF}}$) is ignored (there is no feature pair in the feature subset), so that only the correlation between features and labels ($\overline{r_{FL}}$) is considered.

After the first iteration of the hill climbing search, the candidate feature subsets will have at least one gene which occurs in at least one cancer-related KEGG pathway and the correlation between features and labels ($\overline{r_{FL}}$) is taken into account. Therefore, a selected feature subset returned by ML-CFS will have at least one gene occurring in a cancer-related KEGG pathway; and the rest of the genes selected by ML-CFS's hill climbing search is expected not only to be highly correlated with class labels but also to have little redundancy with the gene selected in the first iteration.

C. Selecting Only Genes that Occur in KEGG Pathways

When using the approach of embedding KEGG pathway information into the Merit function, there is a chance that the ML-CFS method selects a feature subset which has only one gene occurring in some cancer-related pathways and the rest of the selected genes are not occurring in any cancer related pathway at all. Note that, in our datasets, only 3.13 % of the genes (690 out of 22,060 genes) occur in some cancer-related KEGG pathway, and most of those genes have an *AvgRFP* value lower than 0.15. Hence, we decided to do experiments with another approach which selects only genes which occur in cancer-related KEGG pathways. The idea behind this approach is to investigate what will happen if we force our feature selection method (ML-CFS) to select a feature subset from a feature space containing only the genes (features) that occur in some cancer-related pathway. Hence, in this approach we remove all genes which do not occur in any cancer-related pathway from the feature space. After that we give all the remaining genes (i.e. all the genes occurring in some cancer-related KEGG pathway) as input to the ML-CFS method.

IV. Computational Results and Discussion

A. Datasets and Experimental Setup

In our experiments, we have analysed two multi-label microarray datasets. The first one (hereafter referred to as dataset 1) consists of 28,536 features (genes), 24 instances (cell lines) and 2 class attributes. These two class attributes stand for two drugs used to treat neuroblastoma, namely: *Nutlin-3* and *Rita*. The second multi-label microarray dataset (hereafter referred to as dataset 2) also has 28,536 features (genes) and 24 instances (cell lines), but it has 3 different class attributes (different drugs used to treat neuroblastoma), namely: *Cisplatin*, *Carboplatin* and *Oxaliplating*. Both these datasets were obtained from the resistant cancer cell line (RCCL) collection [22].

Before running all experiments on those two datasets, all features were normalized according to the zero-mean normalization method. i.e., a feature's mean value is normalized to 0, and the value of a feature for an instance was normalized to the number of standard deviations above or below the feature's mean. Moreover, we remove genes with unknown names because we aimed at selecting genes whose relevance to drug resistance/sensitivity can be interpreted by biologists. After removing unknown genes, the number of features (genes) that remained on dataset 1 is 22060, and 22,058 genes (features) remained on dataset 2 (each dataset had about 22.7% of genes with unknown names).

We ran experiments with 4 types of feature selection approaches applied to those two microarray datasets: (1) running the *multi*-label CFS (ML-CFS) method proposed in [5] – as discussed in Section II-C, (2) running ML-CFS with KEGG pathway information in weighted formula, (3) running ML-CFS with KEGG pathway information embedded into the merit function, and (4) running the ML-CFS for selecting only genes that occur in KEGG pathways.

The feature subset selected by each of those 4 approaches was used as input by two different multi-label classification algorithms, namely ML-KNN (multi-label K-nearest neighbours) [7] and ML-RBF (multi-label radial basis function) neural networks [8]. These algorithms were run with their default parameters, mentioned on the corresponding papers.

In our experiments, we use the leave-one-out cross validation (LOOCV) procedure to estimate predictive accuracy [2]. This well-known procedure runs a classification algorithm n times, where n is the number of instances (cell lines). In each run a different instance is used as the test data and the other n – 1 instances are used as the training data.

The predictive accuracy of each classification algorithm was measured by a well-known multi-label predictive accuracy measure named hamming loss. It takes into account prediction errors (an incorrect label is predicted) and missing errors (a label is not predicted). Note that the smaller the hamming loss value, the better the predictive accuracy. The hamming loss is defined in formula (9):

$$HammingLoss = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \, \Delta \, Z_i|}{|L|} \qquad (9)$$

Where D is a multi-label test data set, consisting of |D| multi-label instances (x_i, Y_i), i = 1..|D|, Y_i is the set of class labels associated with the i-th instance. $Y_i \subseteq L$, L is the set of class labels and $|L|$ is the number of labels in L. Z_i is the set of labels predicted by the multi-label classifier for the i-th instance and Δ is the symmetric difference of two sets and corresponds to the XOR operation in Boolean logic. That is, a class label belongs to the set of labels defined by $Y_i \, \Delta \, Z_i$ if and only if that label occurs in either Y_i or Z_i, but not in both sets.

To measure statistical significance, we use the Wilcoxon's Signed-Rank test, which is a non-parametric significance test here used for comparing the results of two algorithms on a single application domain (dataset) [23]. In our case, we compare the hamming losses obtained by the ML-CFS algorithm without using KEGG pathway information and the ones obtained by the proposed 3 extended versions of ML-CFS using KEGG pathway information, i.e. ML-CFS using a weighted formula, ML-CFS with KEGG pathway information embedded into Merit function and ML-CFS selecting only genes that occur in some KEGG pathway.

B. Hamming Loss Results

Table I shows the hamming loss of two multi-label classification algorithms (ML-kNN and ML-RBF) which were applied to the feature subsets selected by different versions of the ML-CFS feature selection method. The best (smallest) hamming loss obtained for each dataset, for each of the two multi-label classification algorithms, is shown in boldface. Comparing the hamming loss across different parameter settings (α *and* β values), for the ML-kNN classifier (top half of Table I) on dataset 1, the high value of α and low value of β (0.9 and 0.1 respectively) used together obtained the smallest hamming loss; while on dataset 2 the smallest hamming loss was obtained when we set α = 0.7 and β = 0.3. Moreover, those hamming loss values (0.188 and 0.111 from dataset 1 and dataset 2, respectively) are smaller than the hamming loss values which were obtained by the ML-CFS without using KEGG pathway information (0.229 and 0.153 respectively). The corresponding hamming loss differences in each dataset are statistically significant according to the two-tailed Wilcoxon Signed-Rank test at the 5% significance level.

Hence, when the features selected by ML-CFS are used by the ML-kNN algorithm, the best results are obtained using a moderate (0.3) or small (0.1) β value, meaning that the use of KEGG pathway information is beneficial if its relative weight in the evaluation function is moderate or small. A relatively large β value, 0.5, leads to much larger hamming losses (larger than the losses obtained without using KEGG pathway information). This can be explained by the relatively large weight of the KEGG pathway information, and so the relatively low weight of the Merit term in the evaluation function, which means there is so much emphasis on selecting genes that occur in KEGG pathways that there is not enough emphasis on selecting genes with a high Merit value (i.e. genes that are highly correlated with the class and that have little redundancy).

TABLE I. HAMMING LOSS VALUES MEASURED BY LEAVE-ONE-OUT CROSS-VALIDATION (WITH STANDARD ERRORS BETWEEN BRACKETS)

Feature Selection Approach	**Weights**		**Dataset 1**	**Dataset 2**
	α	**β**	**Hamming loss**	**Hamming loss**
		ML-kNN		
Multi-Label CFS	-	-	0.229 (0.060)	0.153 (0.053)
Multi-Label CFS with KEGG pathway information in weighted formula	0.5	0.5	0.354 (0.064)	0.292 (0.073)
	0.7	0.3	0.333 (0.065)	**0.111(0.043)**
	0.9	0.1	**0.188 (0.066)**	0.277 (0.076)
Multi-Label CFS with KEGG pathway information embedded into Merit function	-		0.458(0.073)	0.417(0.078)
Multi-Label CFS selecting only genes that occur in KEGG pathways	-		0.291 (0.067)	0.264(0.063)
		ML-RBF		
Multi-Label CFS	-	-	0.188 (0.066)	0.083 (0.041)
Multi-Label CFS with KEGG pathway information in weighted formula	0.5	0.5	**0.167 (0.049)**	0.236 (0.079)
	0.7	0.3	**0.167 (0.049)**	0.250 (0.073)
	0.9	0.1	0.291 (0.074)	**0.056 (0.033)**
Multi-Label CFS with KEGG pathway information embedded into Merit function	-		0.229(0.052)	0.389(0.081)
Multi-Label CFS selecting only genes that occur in KEGG pathways	-		0.229(0.067)	0.097(0.037)

For the ML-RBF classifier (results in the bottom half of Table I), we obtained the smallest hamming loss when we set (α = 0.5 and β = 0.5) or (α = 0.7 and β = 0.3) on dataset 1, while the smallest hamming loss was obtained when we set α = 0.9 and β = 0.1 on dataset 2. Note that those hamming loss values (0.167 and 0.056 respectively) are smaller than the hamming loss values which were obtained by the ML-CFS without using KEGG pathway information (0.188 and 0.83, respectively). The corresponding differences in each dataset are statistically significant according to the two-tailed Wilcoxon Signed-Rank test at the 5% significance level.

The ML-CFS version where KEGG pathway information is embedded into the merit function and the version selecting only genes that occur in KEGG pathways obtained larger hamming losses, overall, compared with ML-CFS using a weighted formula to combine the value of the merit function and KEGG pathway information.

C. Analysis of Selected Genes

Table II shows the genes most frequently selected by each version of ML-CFS for each dataset, with the corresponding selection frequency shown in brackets. The table shows only genes which were selected in at least 12 out of the 24 iterations of the LOOCV procedure, with the exception of the table entry for genes selected using KEGG pathway information embedded into Merit function, in dataset 1.

We focus our analysis on the genes selected by the ML-CFS version using KEGG pathway information in weighted formula, which was overall the most successful ML-CFS version in terms of predictive accuracy. As shown in Table II, considering the dataset 1, when we use different $\alpha \; and \; \beta$ weight values, our algorithm discovers two different sets of selected genes: genes TP53, PCNA and MDM2 were selected when we set (α = 0.5 and β = 0.5) or (α = 0.7 and β = 0.3); whilst genes ERCC1 and CCDC72 were selected when α = 0.9 and β =0.1. On dataset 2, ML-CFS with KEGG pathway information in weighted formula also selected different genes for different $\alpha \; and \; \beta$ weight values, for example: CCND1 and BCL2 were consistently selected when (α = 0.5 and β = 0.5) or (α = 0.7 and β = 0.3); while MAD2L2, CSNK2A and MTSS1 were selected when α = 0.9 and β =0.1.

Interestingly, the gene 'PCNA' was selected by ML-CFS 24 times (out of 24 runs) for dataset 1 and 21 times for dataset 2, when we set α = 0.5 and β = 0.5. When we set α = 0.7 and β = 0.3, this gene was selected 24 times for dataset 1 and 10 times for dataset 2 (the latter figure is not shown in Table II because in general that Table shows only genes selected in at least 12 out of 24 runs).

Table III shows the genes which were frequently selected by ML-CFS using KEGG pathway information in weighted formula, no matter what parameter setting. The table shows the frequency of selection out of 72 because it considers 3 runs of LOOCV, one run for each parameter setting (α and β values), and each LOOCV run has 24 iterations. As shown in the table, in dataset 1, genes MDM2, TP53 and PCNA are consistently selected in about 50 (out of 72) runs of ML-CFS, considering all variations of parameter settings. In dataset 2, only gene CCND1 achieves this level of selection frequency.

Table IV shows the average number of selected genes for each ML-CFS version used in the experiments. Most feature selection approaches selected less than 10 genes (out of the about 22,000 genes in the feature space) in both datasets. The exception is ML-CFS with KEGG pathway information embedded into the Merit function, which selected a substantially higher number of genes (about 17 genes in both datasets). This was also the approach that led to the worst hamming loss results, as discussed earlier. In any case, the proportion of selected genes by each approach was always smaller than 0.1% of all genes in the feature space.

It should be noted that, comparing the set of genes selected by ML-CFS without using KEGG pathway information with the set of the genes selected by ML-CFS using KEGG pathway information, the two gene sets are very different in general. Moreover, only one gene which was selected by ML-CFS without using KEGG pathway information occurs in KEGG pathways, namely the gene "RASSF2". On the other hand, most of the genes which were selected by ML-CFS using a weighted formula occur in cancer-related KEGG pathways, except the genes "MTSS1" and "CCDC72". (Note that, here we consider only genes which were selected in more than 12 iterations of LOOCV).

TABLE II. SELECTED GENES AND THEIR SELECTION FREQUENCY FOR DIFFERENT VERSIONS OF THE ML-CFS FEATURE SELECTION METHOD

Feature Selection Approach	**Parameter**		**Dataset 1**	**Dataset 2**
	α	***β***	**Selected Gene and Selection Frequency**	**Selected Gene and Selection Frequency**
Multi-Label CFS	-		RASSF2 (23) KLC4 (16) SLC12A7 (15)	KIAA2013 (22) MAD2L2 (18) CSNK2A1 (12)

Multi-Label CFS with KEGG pathway information in weighted formula	0.5	0.5	TP53 (24)
			PCNA(24)
			MDM2(23)
				PCNA(21) RPA4(18)
				CDK4(16) BCL2(12)
				CCND1(19)

	0.7	0.3	TP53 (24)
			PCNA(24)
			MDM2(24)
				DVL1(24) BCL2 (22)
				CCND1(24)
				CSNK2A1 (16)

	0.9	0.1	ERCC1(23)
			CCDC72(19)
				MAD2L2 (24)
				CSNK2A1 (12)
				MTSS1 (12)

Multi-Label CFS with KEGG pathway information embedded into Merit function	-		ERCC1(9) RPRM (8) (None of them were selected more than 12 times of LOOCV)	MAD2L2 (24)

Multi-Label CFS selecting only genes that occur in KEGG pathways	-		FUT8(15) RPRM(14)
			PRKCA(19)
			RNASEH2B (14)
			NOTCH1 (14)
				MAD2L2 (24)
				CSNK2A1 (23)
				BAI1 (22)

TABLE III. GENES WHICH WERE FOUND BY ML-CFS USING KEGG PATHWAY INFORMATION (NO MATTER WHAT PARAMETER SETTING)

Dataset 1		**Dataset 2**	
Selected Gene	**Frequency of selection (Out of 72)**	**Selected Gene**	**Frequency of selection (Out of 72)**
MDM2	52	CCND1	50
TP53	52	BCL2	38
PCNA	50	DVL1	32
		PCNA	31
		MAD2L2	30
		CSNK2A1	28

Table V shows the list of cancer-related KEGG pathways in which each frequently selected gene occurs. Note that the four genes most frequently selected in Table III (MDM2, TP53, PCNA and CCND1) all occur in the cell cycle pathway, and three of those genes (MDM2, TP53 and CCND1) also occur in the p53 signaling and the melanoma pathways.

D. Discussion on the Biological Relevance of the Most Frequently Selected Genes

In this section we compare the biological relevance of the genes selected by different versions of ML-CFS, namely without using KEGG pathway information and using cancer-related KEGG pathway information in a weighted formula – since this was, overall, the ML-CFS version that obtained best predictive accuracy results, among the 3 versions using KEGG pathway information. ML-CFS *with* KEGG pathway information was superior in selecting genes that are potentially relevant for drug resistance, as discussed next.

The genes selected by ML-CFS *without* KEGG pathway information in dataset 1, namely RASSF2, KLC4, and SLC12A7, are difficult to interpret in the context of cancer cell resistance to the drugs associated with that dataset (Nutlin-3 and RITA). Among the genes selected by that ML-CFS version in dataset 2 (associated with drugs Cisplatin, Carboplatin and Oxaliplatin), MAD2L2 and CSNK2A are involved in cell cycle regulation, a relevant process in the cancer cell response to anti-cancer drugs. However, there is no obvious connection.

TABLE IV. THE AVERAGE NUMBER OF SELECTED GENES FOR EACH MULTI-LABEL CFS APPROACH

Feature Selection Approach	Parameter		Dataset 1	Dataset 2
	α	β	Avg No. of selected genes	Avg No. of selected genes
Multi-Label CFS	-		8.29	6.67
Multi-Label CFS with KEGG pathway information in weighted formula	0.5	0.5	2.96	5.54
	0.7	0.3	3.38	7.54
	0.9	0.1	9.67	5.67
Multi-Label CFS with KEGG pathway information embedded into Merit function	-		17.54	16.46
Multi-Label CFS selecting only genes that occur in KEGG pathways	-		8.63	3.79

TABLE V. SELECTED GENES AND THE CANCER-RELATED KEGG PATHWAYS IN WHICH THEY ARE INCLUDED

Gene: Gene Description	Pathways
TP53: tumor protein p53	Cell cycle, p53 signaling pathway, Apoptosis, Wnt signaling pathway, Pathways in cancer, Transcriptional misregulation in cancer, Melanoma
PCNA: proliferating cell nuclear antigen	DNA replication, Base excision repair, Nucleotide excision repair, Mismatch repair, Cell cycle
MDM2: Mdm2 p53 binding protein homolog	Cell cycle, p53 signaling pathway, Transcriptional misregulation in cancer, Non-homologous end-joining, Melanoma
ERCC1: excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence)	Nucleotide excision repair, Fanconi anemia pathway
CCND1: cyclin D1	Cell cycle, p53 signaling pathway, Melanoma, Wnt signaling pathway, Pathways in cancer
RPA4: replication protein A4, 30kDa	DNA replication, Nucleotide excision repair, Mismatch repair, Homologous recombination, Fanconi anemia pathway
CDK4: cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	Cell cycle, p53 signaling pathway, Pathways in cancer, Melanoma
BCL2: B-cell CLL/lymphoma 2	p53 signaling pathway, Apoptosis, Pathways in cancer, Transcriptional misregulation in cancer, Melanoma
DVL1: dishevelled, dsh homolog 1 (Drosophila)	Wnt signaling pathway, Notch signaling pathway, Pathways in cancer
CSNK2A1: casein kinase 2, alpha 1 polypeptide	Wnt signaling pathway
MAD2L2: MAD2 mitotic arrest deficient-like 2 (yeast)	Cell cycle

In contrast, ML-CFS *with* KEGG pathway information selected genes that are known to be affected by resistance development to the investigated compounds. In dataset 1, genes TP53, PCNA, ERCC1, MDM2, and CCDC72 were selected. Nutlin-3 and RITA interfere with the interaction of p53 (encoded by TP53) with its endogenous inhibitor MDM2. Therefore, TP53 and MDM2 are obvious candidates whose expression may be affected in resistant cells. Particularly, biological experimental results demonstrated that TP53 is frequently found mutated in Nutlin-3-resistant cells [22]. PCNA and ERCC1 are both involved in DNA repair, and Nutlin-3 and RITA are both known to induce DNA damage [24-27]. Little information is available on CCDC72 (also known as TMA7). It is a peptide discovered in yeast that binds to the translation machinery. Its depletion results in enhanced cellular resistance to stress.

In dataset 2, ML-CFS *with* KEGG pathway information selected genes PCNA, CDK4, BCL2, MTSS1, CCND1, DVL1, and RPA4 as potentially relevant genes for the neuroblastoma cell resistance to platinum drugs. Platinum drugs are thought to exert their anti-cancer effects primarily through induction of DNA damage [28, 29]. PCNA, RPA4 (both involved in DNA repair pathways), DVL1 (constituent of the Wnt signalling pathway), CDK4, CCND1 (both involved in cell cycle regulation), and BCL2 (which encodes for an anti-apoptotic protein) are all known to be potentially involved in the response and resistance to platinum drugs [29]. A potential role of MTSS1 remains unclear with regard to the current knowledge state.

V. CONCLUSIONS

Although there is previous research on KEGG pathway-based feature selection for conventional, single-label classification of microarray data; to the best of our knowledge, this is the first paper to propose a KEGG pathway-based feature selection method for multi-label classification. We proposed to extend our previous multi-label correlation-based feature selection (ML-CFS) method with cancer-related KEGG pathway information, in order to bias the search towards features (genes) that have not only high predictive accuracy, but also are known to occur in cancer-related KEGG pathways. Genes satisfying the latter criterion are intuitively expected to be more relevant (at least their relevance seems more easily interpretable by biologist) as features predicting whether or not a cancer cell line will be sensitive or resistance to certain drugs. Indeed, in this work genes selected using KEGG pathway information were found more biologically relevant to the analysis of our datasets than genes selected without using KEGG pathway information.

In addition, when we used KEGG pathway information as a part of ML-CFS's evaluation function with the best parameter setting (in a weighted formula), we obtained, in general, a statistically significantly smaller hamming loss when compared to the hamming loss obtained by ML-CFS without using KEGG pathway information (in both our microarray datasets). Hence, the weighted formula approach to use KEGG pathway information in the evaluation function can be considered successful, as long as care is taken to do experiments with different parameter settings.

However, the other two approaches for using KEGG pathway information – i.e., embedding that information into the ML-CFS's Merit function and selecting only genes that occur in cancer-related KEGG pathways – obtained a larger hamming loss compared with ML-CFS without using KEGG pathway information. Hence, it is interesting to note that, when the genes given as input to ML-CFS were only genes occurring in cancer-related KEGG pathways, ML-CFS was not able to select very relevant genes in terms of predictive accuracy, i.e. that approach was not able to improve predictive accuracy and biological relevance simultaneously. In contrast, using a weighted formula to combine the original ML-CFS's Merit function and KEGG pathway information has achieved a good trade-off between improving predictive accuracy and biological relevance simultaneously.

Concerning future research directions, we will develop new multi-label correlation-based feature selection methods based on different types of search methods, such as genetic algorithms. Moreover, we might integrate an adaptive-parameter technique and another type of biological knowledge to improve the use of the weighted formula in ML-CFS.

REFERENCES

[1] D. M. Dziuda, *Data Mining for Genomics and Proteomics: analysis of gene and protein expression data*, Wiley & Sons, New Jersey, 2010.

[2] I. H. Witten, E. Frank, M.A. Hall, *Data Mining: Practical Machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2011.

[3] H. Lui, H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic, Massachusetts, 1998.

[4] G. Tsoumakas, I. Katakis, I. Vlahavas, "Mining Multi-label data," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon, L. Rokach, Eds. Springer, Heidelberg, 2010, pp. 667-685.

[5] S. Jungjit, A.A. Freitas, M. Michaelis and J. Cinatl, "Two Extensions to Multi-Label Correlation-Based Feature Selection: a case study in bioinformatics," in *Proceedings of the 2013 IEEE International Conference on Systems, Man and Cybernetics*, Manchester, UK, 2013, in press.

[6] The Kanehisa Laboratory in the Institute for Chemical Research (ICR), Kyoto University, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Internet: http://www.genome.jp/kegg/, [16 May 2013].

[7] M. L. Zhang and Z. H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40(7), pp. 2038-2048, Jul. 2007.

[8] M. L. Zhang, "ML-RBF: RBF Neural Networks for Multi-Label Learning," *Neural Process. Lett*, vol. 29(2), pp. 61-74, Apr. 2009.

[9] Y. Saeys, I. Inza, P. larranaga, "A review of Feature Selection Technique in Bioinformatics." *Bioinformatics*, vol. 23(19), pp. 2507-2517, Aug. 2007.

[10] G. V. S. George, V. C. Raj, "Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification Using Gene Expression Profile," *Computer Science & Engineering Survey*. vol.2(3), pp. 16-26, Aug. 2011.

[11] M. A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class machine Learning," in *Proceedings of the 17th International Conference on Machine Learning (ICML-2000)*, Morkan Kaufmann, San Francisco, 2000, pp.359-366.

[12] L. Yu, H. Lui, "Feature Selection for High Dimensional Data: A fast correlation-based feature selection solution," in *Proceeding of the Twenty International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003, pp. 856-863.

[13] H. Lui, H. Motoda, R. Setiono and Z. Zhao, "Feature Selection: An ever Evolving Frontier in Data Mining," in *Proceedings of the Fourth Workshop on Feature Selection in Data Mining*, Hyderabad, India, 2010, pp. 4-13.

[14] S. Jungjit, A.A. Freitas, M. Michaelis and J. Cinatl, "A Multi-Label Correlation Based Feature Selection Method for the Classification of Neuroblastoma microarray data", in *Advances in Data Mining: 12th Industrial Conference (ICDM 2012): Workshop Proceedings – Workshop on Data Mining in Life Sciences (DMLS 2012)*, I. Bichindaritz, P. Perner, G. Rub, and R. Schmidt, Eds, IBAI Publishing, July 2012, pp. 149-157.

[15] M. L. Zhang, J. M. Pena and V. Robles, "Feature selection for multi-label naive Bayes classification," *Information Science*, vol.179(19), pp. 3218-3229, Sep. 2009.

[16] G. Doquire, M. Verleysen, "Feature Selection for Multi-label Classification Problems," in *Lecture Notes in Computer Science*, vol. 6691, Springer, Heidelberg, pp. 9-16, 2011.

[17] N. Spolaor, E.A. Cherman and M.C. Monard, "Using ReliefF for Multi-label feature selection," in *Proceedings of Conferencia Latinoamericana de Informatica*, 2011, pp. 960-975.

[18] N. Spolaor, E.A. Cherman, M.C. Monard and H. D. Lee, "Filter Approach Feature Selection Methods to Support Multi-label Learning Based on ReliefF and Information Gain," in *SBIA 2012, Lecture Notes in Artificial Intelligence*, vol.7589, L. N. Barros et al, Eds, Springer, Heidelberg, 2012, pp.72-81.

[19] G. Lastra, O. Luaces, J. R. Quevedo and A. Bahamonde, "Graphical Feature Selection for Multilabel Classification Tasks." in *Proceedings of the 10th international conference on Advances in Intelligent Data Analysis X. Lecture Notes in Computer Science*, vol. 7014, Springer, Heidelberg, pp. 246-257, 2011.

[20] N. Bandyopadhyay, T. Kahveci, S. Goodison, Y. Sun, and S. Ranka, "Pathway-Based Feature Selection Algorithm for Cancer Microarray Data," in *Advances in Bioinformatics*, 2009.

[21] E. Glaab, J. M. Garibaldi and N. Krasnogor, "Learning pathway-based decision rules to classify microarray cancer samples," in *German Conference on Bioinformatics 2010, Lecture Notes in Informatics (LNI)*, vol.173, Sep. 2010, pp.123-134.

[22] M. Michaelis, F. Rothweiler, S. Barth, J. Cinatl, M. van Rikxoort, N. Löschmann, Y. Voges, R. Breitling, A. von Deimling, F. Rödel, K. Weber, B. Fehse, E. Mack, T. Stiewe, H.W. Do-err, D. Speidel, J Jr. Cinatl, "Adaptation of cancer cells from different entities to the MDM2 inhibitor nutlin-3 results in the emergence of p53-mutated multi-drug resistant cancer cells," in *Cell Death Dis*, vol. 2(e243), Dec. 2011.

[23] N. Japkowicz and M. Shah, *Evaluation Learning Algorithms: a Classification Perspective*, Cambridge University Press, 2011.

[24] R. Verma, M.J. Rigatti, G.S. Belinsky, C.A. Godman and C. Giardina, "DNA damage response to the Mdm2 inhibitor nutlin-3," in *Biochem Pharmacol*, vol. 79(4), pp. 565-74, Feb. 2010.

[25] J. Yang, A. Ahmed, E, Poon, N, Perusinghe, A, de Haven Brandon, G. Box, M, Valenti, S, Eccles, K, Rouschop, B. Wouters and M. Ashcroft, "Small-molecule activation of p53 blocks hypoxia-inducible factor 1alpha and vascular endothelial growth factor expression in vivo and leads to tumor cell apoptosis in normoxia and hypoxia," in *Mol Cell Biol*. vol. 29(8), pp. 2243-53, Feb. 2009.

[26] M.J. Rigatti, R. Verma, G.S. Belinsky, D.W. Rosenberg and C. Giardina, " Pharmacological inhibition of Mdm2 triggers growth arrest and promotes DNA breakage in mouse colon tumors and human colon cancer cells," in *Mol Carcinog*, vol. 51(5), pp.363-78, May 2012.

[27] J.M. Valentine, S, Kumar and A. Moumen, "A p53-independent role for the MDM2 antagonist Nutlin-3 in DNA damage response initiation," in *BMC Cancer*. vol. 11(79), Feb 2011.

[28] L, Galluzzi, L, Senovilla, I. Vitale, J. Michels, I, Martins, O, Kepp, M. Castedo and G, Kroemer, "Molecular mechanisms of cisplatin resistance," *in Oncogene*. vol.31(15), pp.1869-83, Apr 2012.

[29] D. W. Shen, L.M. Pouliot, M.D. Hall and M.M. Gottesman, "Cisplatin resistance: a cellular self-defense mechanism resulting from multiple epigenetic and genetic changes," in *Pharmacol Rev*. vol.64(3), pp.706-21, Jul 2012.