Two Extensions to Multi-label Correlation-Based Feature Selection: a case study in bioinformatics

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 7NF, 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 — **This paper proposes two extensions to a Multi-Label Correlation Based Feature Selection Method (ML-CFS): (1) ML-CFS using the absolute value of the correlation coefficient in the equation for evaluating a candidate feature subset, and (2) ML-CFS using Mutual Information for class label weighting. These extensions are evaluated in a bioinformatics case study addressing the multi-label classification of a cancer-related DNA microarray dataset with over 20,000 features. The results show that ML-CFS with absolute value of correlation obtained a significantly better predictive accuracy (smaller hamming loss) than the original ML-CFS. On the other hand, using Mutual Information to assign weights to labels showed some positive effect when using the ML-RBF classifier, but it showed a negative effect when using the ML-kNN classifier.**

Keywords - multi-label feature selection, multi-label classification, microarray data

I. INTRODUCTION

Classification is a data mining task which aims to learn the relationship between the values of the predictor attributes of an instance and its class label(s). This relationship is learned from pre-classified instances in the training set, and then the learned classification model is used to predict the class label of previously unseen instances in the test set. Traditionally, the vast majority of works on the classification task have addressed a single-label classification problem, where each instance in the data set is associated with just one class label.

Another type of classification problem is multi-label classification. In this type of problem each instance can be associated with a set of class labels, rather than just one class label as in single-label classification. Multi-label classification methods have been used in many application domains; such as text classification, music classification, bioinformatics and medical diagnosis [1]. In this work we focus on multi-label classification of a DNA microarray dataset.

The main challenge in microarray data classification is that the number of features (genes) is very large – more than 20,000 in the dataset used in this work – whilst the number of instances is very small – only 24 instances (cell lines) in this work. This high degree of data sparseness makes classification models prone to over-fitting. Hence, feature selection is an important task in microarray data classification, and we address this task in the context of multi-label classification.

In this context, we propose two extensions to a recently proposed multi-label correlation-based feature selection method [2]. The first extension consists of using the absolute value of the correlation coefficient in the equations for evaluating a candidate feature subset. The second extension consists of computing the mutual information between pairs of class labels and using that information to assign, to each label, a weight that depends on its degree of correlation with other labels – again, modifying the equation for evaluating a candidate feature subset. These two extensions are evaluated in a case study with a cancer-related DNA microarray dataset.

The rest of this paper is organized as follows. Section II gives an overview of microarray data classification. Section III presents a brief review of feature selection methods, both single-label and multi-label ones. Section IV introduces the two proposed extensions to a multi-label feature selection method. Section V reports the computational results. Section VI concludes the paper and mentions future work.

II. MICROARRAY DATA CLASSIFICATION

DNA microarray technology was developed for measuring the gene expression levels of tens of thousands of gene simultaneously [3]. DNA microarray datasets are widely used to find out correlations between gene expression values and diseases or different functional statuses of cells.

The vast majority of DNA microarray datasets contain a single column representing the class of each instance, characterizing a conventional single-label classification problem. In this work we focus on a more challenging type of DNA microarray data, where there are 3 columns representing 3 class attributes. This is a multi-label classification problem where each class attribute refers to a drug applied to neuroblastoma (a type of cancer) cell lines, and each cell line (instance) is assigned the label "sensitive" or "resistant" for each of the 3 class attributes (drugs). The drug names are *Cisplatin*, *Carboplatin* and *Oxaliplating*. The dataset used here was obtained from the resistant cancer cell line (RCCL) collection [4].

III. Feature Selection

In the context of microarray data analysis, where the number of features (attributes or genes) is very large while the number of instances (or cell lines) is very small, feature selection is a very important task, and it can significantly decrease the risk of model overfitting [5]. Feature selection is often performed in a data pre-processing step of the knowledge discovery processes, in order to select a relevant or useful feature subset according to an evaluation criterion [6]. Feature selection can improve the predictive performance and eliminate irrelevant and/or redundant features [7].

A. Conventional, Single-Label Feature Selection

In general, feature selection methods can be classified into three approaches: (1) the filter approach, (2) the wrapper approach, and (3) the embedded approach [5]. The filter approach is independent of the classifier, and this approach is usually fast and scalable to datasets with very large number of features. This approach has been used to design several feature selection methods, such as Correlation-based Feature Selection [8] and Fast Correlation-based Feature Selection [9].

Additionally, Lui et al [10] highlighted that the filter approach is the feature selection approach most used in real-world applications, especially when the number of features in the dataset is very large, such as in microarray data. The structure of filter algorithms is very simple, and it provides a simple way to calculate the relevance of features in large-scale data in a short time.

On the other hand, the structure of the wrapper approach is more complicated. This approach selects the best feature subset by doing a search in the feature space guided by an evaluation function based on a classifier's predictive accuracy. The wrapper approach tends to be better at maximizing predictive accuracy than the filter approach, because the former directly uses the accuracy of the classifier as the evaluation function of a feature subset. However, when using the wrapper approach there is a risk of model overfitting [5, 7]. Moreover, the wrapper approach is usually much more computationally expensive than the filter approach because a classification algorithm has to be run for each candidate feature subset, which is not the case in filter approach.

The third feature selection approach is the embedded approach. This approach embeds the search for a good feature into the classifier construction process. Hence, this approach is classifier-specific, and can also be computationally expensive.

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

The size of the literature on multi-label feature selection is relatively small, by comparison with the huge size of the literature on traditional single-label selection. Some works on the filter approach for multi-label feature selection (the focus of this paper) are briefly reviewed next.

Doquire and Verleysen [11], as well as Spolaor et al. [12], have essentially transformed the multi-label dataset into a single-label one – which is usually referred to as the binary relevance approach – and then applied a single-label feature selection method to the transformed data. The main drawbacks of this binary relevance approach are that it cannot deal with the multi-label problem directly and it does not consider the relationship between labels – since it considers each label separately. By contrast, we propose two extensions to a multi-label feature selection method that directly copes with the original multi-label data, and one extension considers the relationship between labels during the feature selection process.

Lastra et al [13] extended the single-label feature selection method proposed by Yu and Lui [9] to multi-label classification. Their method assumed all features were discrete. However, microarray data are continuous and the discretization of microarray data can lead to loss of relevant information, especially in microarray datasets with over 20,000 continuous features (like the dataset used in our experiment). By contrast, the extended multi-label feature selection method proposed here does not require data discretization, it can directly cope with continuous attributes.

The method proposed by Spolaor [14], called IG-ML, selects features which have a multi-label information gain (IG) value greater than or equal to a pre-defined threshold. The IG-ML method is a version of the IG where an instance is counted once for each class label's IG calculation, and then the IGs of all labels are added [15]. This method has the drawback of requiring a pre-defined threshold value, which is typically chosen via extensive trial and error experiments or chosen by the user in an ad-hoc fashion.

Multi-label ReliefF and F-statistic feature selection were proposed by Kong et al [16]. In the former, the problem is decomposed into a set of pairwise multi-label two class problems and the multi-class single label ReliffF score is adapted to the multi-label scenario. In multi-label F-statistics, they utilized the class-wise between-class scatter matrix and class-wise within-class matrix associated with a multi-Label linear discriminant analysis algorithm.

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

This ML-CFS method has been recently proposed by Jungjit et al. [2]. In that work the authors extended the single-label Correlation-based Feature Selection (CFS) method – which was proposed by Hall [8] – to the more complex scenario of multi-label classification. The basic idea of the CFS method is to perform a search in the space of candidate feature subsets guided by a *merit* function, which evaluates the merit (quality) of each candidate feature subset. The search method used was a simple hill-climbing algorithm, which tries to find the feature subset with the maximum value of the merit function. This involves maximizing the correlations between features and the class labels for features in a candidate feature subset (to select features with high predictive accuracy) and minimizing the correlations between pairs of features in the feature subset (to avoid the selection of redundant features).

The component of the single-label CFS method that was extended to derive the multi-label ML-CFS method was the merit function used to measure the quality of a candidate feature subset, as defined by equation (1), where k is the number of features in a candidate feature subset F – other terms are defined below. Both the single-label CFS and ML-CFS use that equation. The difference between these methods is that 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. I.e., for each feature f, it averages the feature-label correlation over all labels in L, using equation (2); and then averages the result of equation (2) over all features, as shown in equation (3). By contrast, in the conventional single-label CFS method the equations are simpler, because there is no need to measure average correlations over multiple class labels. Note that the difference between single-label CFS and ML-CFS refers only to the computation of ($\overline{r_{FL}}$); the term ($\overline{r_{FF}}$) – the average correlation over all pairs of features – is the same in both methods.

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

$$\overline{r_{fL}} = \frac{\sum_{i=1}^{|L|} r_{fL_i}}{|L|} \qquad (2)$$

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

The ML-CFS method has the advantage of being able to cope with multi-label classification problems directly (it does not need to transform a multi-label problem into a set of single-label problems). However, there are some issues in the original ML-CFS that could potentially be improved, in order to try to improve the predictive accuracy of the feature subset selected by this method, as discussed in the next Section.

IV. TWO PROPOSED EXTENSIONS TO MULTI-LABEL CORRELATION-BASED FEATURE SELECTION

This work proposes two different extensions to the multi-label correlation-based feature selection (ML-CFS) method: (1) using the absolute value of the correlation coefficient, and (2) using Mutual Information for class label weighting.

A. Extending ML-CFS with the Absolute Value of the Correlation Coefficient

In the original multi-label ML-CFS method [2], like in the original single-label CFS method [8], Pearson's correlation coefficient was used to estimate the terms $\overline{r_{FF}}$ and $\overline{r_{FL}}$ in equation (1). In general, there are two types of correlation: positive correlation and negative correlation. Both of them can represent redundancy between a pair of features, or represent the relevance of a feature to predict a set of labels, as follows.

For the purpose of measuring redundancy between two features, what matters is the absolute value of the correlation coefficient (r), regardless of its sign. E.g., both r = +0.8 and r = –0.8 represent a strong degree of redundancy. However, in the original single-label and multi-label CFS methods, the value of the merit formulas depend on both the value and the sign of r. If a feature subset contains, say, one pair of features with r = +0.8 and another pair of features with r = –0.8, these two values would cancel each other resulting in an average r over those two feature pairs of 0; a misleading value, since the two r values actually suggest a large degree of redundancy in each of those feature pairs.

Analogously, for the purpose of measuring the relevance of a feature for predicting a set of labels, what matters is the absolute value of the correlation coefficient, not its sign. E.g., both r = +0.8 and r = –0.8 represent strong correlations which can be exploited by a multi-label classification algorithm (to be applied to the features selected in a preprocessing phase).

To mitigate the aforementioned problems, when calculating the value of the average correlation between features in a feature subset F ($\overline{r_{FF}}$) and the average correlation between features and labels ($\overline{r_{FL}}$), we use the absolute (without sign) value of the correlation coefficient in all occurrences of the correlation coefficient r in equation (1). Hence, the average correlation between features in a feature subset F ($\overline{r_{FF}}$) is computed by equation (4), where fp is the number of feature pairs in feature subset F. The average value of the correlation coefficient between features and labels is given by equation (5). Note that $|r_{f_i f_j}|$ and $|r_{f\bar{L}}|$return a value in [0..+1].

$$\overline{r_{FF}} = \frac{\sum_{f_i f_j=1, i\neq j}^{|F|} \left|r_{f_i f_j}\right|}{fp} \qquad (4)$$

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

B. ML-CFS using Mutual Information for Class Label Weighting

In the original ML-CFS method [2], equation (2) computes, for a given feature f, the arithmetic average of the correlation between that feature and a class label over all labels, implicitly assuming that all labels are equally relevant and ignoring dependencies between labels. However, in real-world datasets there might be a significant degree of dependence between some labels, where the occurrence of one label would increase the probability of another label for a given instance. For example, in multi-label classification of emotions in a music dataset, the class label 'Sadness' might be more correlated with the class label 'Depressing' than with the class label 'Cheerful'.

The correlation between labels is important in multi-label classification [17]. If the labels were independent from each other, we could simply transform a multi-label problem into a set of single-label problems using the binary relevance method. However, when there are strong dependences among labels in the data, simply using an approach that ignores label correlations, like binary relevance or computing the arithmetic average of correlations across all labels may not be sufficient to cope well with the label-dependence problem.

To take label dependences into account, we used mutual information (MI) to measure the correlation between each pair of labels. We use MI, rather than Pearson's correlation coefficient, because labels are nominal, rather than numerical, and MI is often used to measure dependencies between nominal variables in feature selection. If the MI between two variables is near zero, this would indicate that the variables are close to independent. The mutual information $I(X; Y)$ between the random variables (class attributes) X and Y is shown in equation (6), where $p(x,y)$ denotes the joint probability of class labels x and y, $p(x)$ denotes the marginal probability of x, the *log* is in base 2, and the summation is over all values of variables *X* and *Y* – i.e., over the *sensitive* and *resistant* values of the three class attributes, in our case.

$$MI(X;Y) = \sum\sum p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad (6)$$

To use MI as a measure of label correlation, we first compute the average MI of each label L_i (AvgMI(L_i)) as defined in equation (7). This is simply the mean of the MI between label L_i and each of the other class labels L_j ($j \neq i$).

$$\mathrm{AvgMI(L_i)} = \frac{\sum_{\mathrm{j=1,j\neq i}}^{|\mathrm{L}|}\mathrm{MI(LiLj)}}{|\mathrm{L}|\text{- }1} \qquad (7)$$

The AvgMI(L_i) value for each label $\mathrm{L_i}$ can then be used to modify the Merit function as follows. When computing the correlation between a feature and a set of labels, equation (2) is extended by assigning a different weight to each feature-label correlation term (for each label L_i), where the weights are based on the AvgMI values computed by equation (7). We investigated two opposite approaches to assign such weights, based on two opposite rationales, as follows.

On one hand, it could be argued that a greater weight should be assigned to feature-label correlations involving labels with *greater* AvgMI values. The rationale for this is that, if a given label L_i is highly correlated with the other labels – i.e., AvgMI(L_i) is large – one should reward features which are strong predictors of that label because a multi-label classification algorithm exploiting label correlations could use an accurate prediction of that label to improve the accuracy in the prediction of other labels. Hence, one approach investigated in this work is to extend equation (2) with equation (8).

On the other hand, it could be argued that a greater weight should be assigned to feature-label correlations involving labels with *smaller* AvgMI values. The rationale for this is that, if a given label L_i is weakly correlated with the other labels – i.e., AvgMI(L_i) is small – a multi-label classification algorithm exploiting label correlations would not be able to use an accurate prediction of *other* labels to improve the accuracy in the prediction of label L_i, and therefore features which are strong predictors of that label should be rewarded regardless of their ability to predict other labels.

In equations (8) and (9), the denominators normalize the weight values so that the sum of weights is 1.

$$r_{f\bar{L}} = \frac{\sum_{i=1}^{|L|} r_{fL_i} \times AvgMI(L_i)}{\sum_{i=1}^{|L|} AvgMI(L_i)} \qquad (8)$$

$$r_{f\bar{L}} = \frac{\sum_{i=1}^{|L|} r_{fL_i} \times (1 - AvgMI(L_i))}{\sum_{i=1}^{|L|}(1 - AvgMI(L_i))} \qquad (9)$$

V. Computational Results

We ran experiments with 6 variations of the extended ML-CFS method, i.e., the 6 possible combinations of our two extensions, namely: two ways of using correlation coefficients (with and without absolute correlations) times three MI-based approaches for label weighting (assigning greater weight to labels with larger average MI, assigning greater weight to labels with smaller MI and not using MI-based weights).

For each of those 6 variations, the features selected by ML-CFS were used as input by two different multi-label classification algorithms, namely ML-KNN (multi-label K-nearest neighbours) [18] and ML-RBF (multi-label radial basis function) neural networks [19]. These algorithms were run with their default parameters, mentioned on the corresponding papers.

Before running ML-KNN and ML-RBF, 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.

Genes with unknown names were deleted before running experiments. This is because we aimed at selecting genes whose relevance to drug resistance/sensitivity can be interpreted by biologists. The original dataset had 28,536 genes, 22.7% of which had unknown names, so the reduced dataset used in our experiments has 22,058 genes (features).

Predictive accuracy was measured by hamming loss (equation (10)), a popular measure of multi-label predictive accuracy that takes into account prediction errors (an incorrect label is predicted) and missing errors (a label is not predicted).

$$HammingLoss = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i\,\Delta\,Z_i|}{|L|} \qquad (10)$$

Where D is a multi-label test data set, consisting of |D| multi-label instances *(x*$(x_i,Y_i,)$, 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.

TABLE I. Hamming loss values measured by leave-one-out cross-validation (with standard errors between brackets)

ML-kNN		
Using absolute value of correlation?	**Using mutual information? (Assign greater weight to)**	**Hamming loss**
no	yes (greater AvgMI)	0.388 (0.052)
no	yes (smaller AvgMI)	0.305 (0.072)
no	no	0.291 (0.061)
yes	yes (greater AvgMI)	0.361 (0.072)
yes	yes (smaller AvgMI)	0.250 (0.070)
yes	no	0.153 (0.053)
ML-RBF		
Using absolute value of correlation?	**Using mutual information? (Assign greater weight to)**	**Hamming loss**
no	yes (greater AvgMI)	0.232 (0.062)
no	yes (smaller AvgMI)	0.431 (0.084)
no	no	0.375 (0.076)
yes	yes (greater AvgMI)	0.083 (0.046)
yes	yes (smaller AvgMI)	0.041 (0.035)
yes	no	0.083 (0.041)

TABLE II. GENES MOST FREQUENTLY SELECTED BY DIFFERENT VERSIONS OF THE ML-CFS METHOD

Using absolute correlation?	**Using mutual information?** **(Assign greater weight to)**	**Selected genes**	**Selection Frequency**	**Avg. No. of selected genes**
no	yes (greater AvgMI)	EPRS	19	8.47
		CLMN	18	
no	yes (smaller AvgMI)	C1orf183	23	7.88
		EPRS	22	
		TACC2	21	
		BCL2	19	
		CALN1	15	
no	no	EPRS	24	7.5
		RIMS3	20	
		CYSLTR2	14	
yes	yes (greater AvgMI)	NECAP2	14	4.42
yes	yes (smaller AvgMI)	AURKAIP1	22	6.58
		RASL10B	17	
		KPNA6	16	
yes	no	KIAA2013	22	6.67
		MAD2L2	18	
		CSNK2A1	12	

We used the well-known leave-one-out cross validation (LOOCV) procedure [20] to estimate the hamming loss of the classification models built by ML-KNN and ML-RBF from the features selected by the ML-CFS feature selection method.

A. *Discussion of Predictive Accuracy Results*

Table I shows the hamming losses (and standard errors between brackets) obtained by ML-kNN and ML-RBF with the features selected by ML-CFS. The effect of using the absolute correlation coefficient values can be seen by comparing pairs of rows that have different values in the column "using absolute value of correlation?" but have the same values in the column "using mutual information?" and refer to the same multi-label classification algorithm. Using the absolute correlation values improved the predictive performance (reduced the hamming loss) in every case: comparing the first and fourth rows, the second row and fifth rows, the third and the sixth rows, for each classifier.

We also evaluated the statistical significance of the difference in hamming loss when comparing the third and the sixth rows of results (where mutual information is not used) for each classifier. Hence, this comparison focuses on the different results associated with using or not the absolute value of correlation, without interference of the mutual information. We used a two-sided Wilcoxon Signed-Rank test [21], where the null hypothesis is that the hamming loss obtained by a classifier (ML-kNN or ML-RBF) is the same regardless of whether or not the ML-CFS method uses the absolute value of correlation. For both classifiers, the smaller hamming loss associated with the use of the absolute value of correlation is statistically significant at the 5% level.

The effect of using the MI weights can be seen by comparing pairs of rows that have different values in the column "using mutual information?" but have the same value in the column and "using absolute value of correlation?" and refer to the same multi-label classification algorithm. When ML-kNN is used as the classifier, unfortunately both MI-based weight strategies led to larger hamming loss than not using MI-based weights at all. However, when using ML-RBF as the classifier, the use of MI weights led to better results. In the scenario where absolute correlation values are not used, assigning greater weight to labels with greater AvgMI led to a substantially smaller hamming loss than not using MI weights (0.232 vs. 0.375). This difference is statistically significant according to the two-tailed Wilcoxon Signed-Rank test at the 5% significance level. In the scenario where absolute correlation values are used, assigning greater weight to labels with smaller AvgMI led to a somewhat smaller hamming loss than not using MI weights (0.041 vs. 0.083). This difference is statistically significant according to the two-tailed Wilcoxon Signed-Rank test at the 5% significance level.

The reason why using MI as label weights is more effective when using ML-RBF than when using ML-kNN seems to be because ML-RBF copes better with correlations between labels. Actually, when classifying a test instance, ML-kNN decides to assign the "yes" or "no" value for each class label separately, based on the maximum a posteriori principle, ignoring label correlations [18].

B. *Discussion on the Most Frequently Selected Genes*

Table II shows the genes most frequently selected by each of the 6 previously defined variants of the ML-CFS feature selection method. The "selection frequency" column shows how many times each gene was selected, out of the 24 iterations of the leave-one-out cross-validation (LOOCV) procedure. Note that this table shows only the genes which were selected in at least 12 LOOCV iterations.

In general, the set of most frequently selected genes (features) varied considerably among the 6 ML-CFS. However, the gene EPRS was consistently very frequently selected (with a frequency between 19 and 24) in the scenario where the absolute value of correlation was not used. On the other hand, this gene is not among the ones most frequently selected in the scenario using absolute value of correlation.

A literature search revealed that EPRS ("glutamyl-prolyl-tRNA synthetase") was detected as a tumor-associated antigen in colon cancer [22]. Moreover, another gene (DUS2L – "dihydrouridine synthase 2") which interacts with EPRS was suggested to be involved in pulmonary carcinogenesis [23]. However, conclusive evidence whether (and if yes which) role EPRS might play in cancer is missing.

When using the absolute correlation values, the most frequently selected genes were AURKAIP1 and KIAA2013. Not much is known about AURKAIP1. It induces degradation of the oncoprotein Aurora A [24]. This suggests that AURKAIP1 may act as a tumour suppresor protein. KIAA2013 is an uncharacterised gene for which no relevant information is available.

In terms of the average number of genes selected by each variant of ML-CFS, the variants using absolute correlation selected fewer (4.4 – 6.7) genes than the variants that do not use absolute correlation (which select 7.5 – 8.5 genes).

VI. Conclusion

In this paper we presented two extensions of a multi-label feature selection method (ML-CFS): (1) ML-CFS using the absolute value of the correlation coefficient, and (2) ML-CFS using Mutual Information for class label weighting. Six ML-CFS versions were evaluated on a bioinformatics dataset, giving the genes selected by those ML-CFS versions to two different multi-label classification algorithms (ML-kNN and ML-RBF) and measuring the corresponding predictive accuracy, in terms of hamming loss. The experiments focused on a case study involving a cancer-related DNA microarray dataset with over 20,000 features (genes) and 3 different class attributes (whose class labels indicate whether a cancer cell line is sensitive or resistant to a certain drug). The results reported can be summarized from two different perspectives:

1) ML-CFS with vs. without absolute correlations. Modifying the ML-CFS's merit function to use the absolute correlation values clearly led to a smaller hamming loss when compared with the original ML-CFS proposed in [2]. The use of absolute correlation values also led to some reduction in the average number of genes selected in each run of ML-CFS.

2) ML-CFS with vs. without mutual information (MI) weights: Unlike the use of absolute correlation values, the use of MI weights led to mixed results: broadly speaking, it reduced hamming loss (at least in some cases) when using ML-RBF as the classifier, but it increased hamming loss when using ML-KNN as the classifier.

We also reported a brief analysis of the biological relevance of some genes selected by the ML-CFS method.

Our experiments were run on only one multi-label microarray dataset, as a case study; and so a natural future research direction is to run further experiments evaluating the proposed ML-CFS extensions on other multi-label datasets. In addition, we would like to incorporate biological knowledge as a part of the merit function that evaluates the quality of candidate feature subsets.

References

[1] 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.

[2] 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.

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

[4] 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, 2(e243), 2011.

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

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

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

[8] 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.

[9] 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.

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

[11] G. Doquire, M. Verleysen, "Feature Selection for Multi-label Classification Problems," in LNCS, vol. 6691, Springer, Heidelberg, 2011, pp. 9-16.

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

[13] G. Lastra, O. Luaces, J. R. Quevedo and A. Bahamonde, "Graphical Feature Selection for Multilabel Classification Tasks." in The 10th international conference on Advances in intelligent data analysis X. LNCS, vol. 7014, Springer, Heidelberg, 2011, pp. 246-257.

[14] 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, LNAI 7589, L. N. Barros et al, Eds, Springer, Heidelberg, 2012, pp.72-81.

[15] A. Clare and R. D. King, "Knowledge Discovery in Multi-Label Phenotype data," in Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '01), L. D. Raedt and A. Siebes Eds., Springer, Heidelberg, 2001, pp. 42-53.

[16] D. Kong, C. Dang, H. Huang and H. Zhao, "Multi-Label ReliefF and F-statistics Feature Selection for Image Annotation," in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(CVPR '12), IEEE Computer Society, Washington, DC, USA, 2012, pp. 2352-2359.

[17] M.-L. Zhang, K. Zhang, "Multi-label learning by exploiting label dependency," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), Washington D. C., 2010, pp. 999-1007.

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

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

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

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

[22] A. Line, Z. Slucka, A. Stengrevics , K. Silina, G. Li and RC. Rees, "Characterisation of tumour-associated antigens in colon cancer," in Cancer Immunol Immunother, vol. 51(10), Nov 2002, pp. 574-82.

[23] T. Kato, Y. Daigo, S. Hayama, N. Ishikawa , T. Yamabuki, T. Ito, M. Miyamoto, S. Kondo and Y. Nakamura, "A novel human tRNA-dihydrouridine synthase involved in pulmonary carcinogenesis," in Cancer Res, vol. 65(13), Jul 2005, pp. 5638-46.

[24] S.K. Lim and G. Gopalan. "Aurora-A kinase interaction protein 1 (AURKAIP1) promotes Aurora-A degradation through an alternative ubiquitin-independent pathway", in Biochem. J., vol. 403(1), 2007 Apr 1, pp. 119-127.