Knowledge-Based Feature Extraction in Genomics

Miloš Krejník and Jiří Kléma

Gerstner Laboratory
Czech Technical University in Prague
Technická 2, Prague 6, 166 27, Czech Republic
krejnm1@fel.cvut.cz, klema@labe.felk.cvut.cz

Abstract. Dimension reduction is the process of reducing the number of variables under consideration. In microarray classification it is widely applied because the high dimensionality of gene-expression data proved to decrease accuracy and comprehensibility of genomic classifiers. Simultaneously, contemporary genomics offers an opportunity to reach beyond the routine application of purely statistical dimension reduction techniques. Availability of a great variability of knowledge on gene roles, functions and gene-gene interactions allows to benefit from knowledge-based approaches to dimension reduction. This paper introduces and tests a feature-extraction algorithm that employs ontological affinity to define the gene similarity measure. This measure is used to form gene clusters whose centroids serve as new features. The features are of a reasonable number with statistically proven noise robustness and with anticipation of easy interpretability.

1 Introduction

Genomic data have been a great challenge for machine learning in recent years. It is given mainly due to emergence of technologies, like microarray or SAGE, which make possible simultaneous measurement of expression of thousands genes. Genomic data are formed from tissue samples converted into a numeric, and thus computable form. Every sample is described by values of features which represent expression/activity of particular genes. A very important genomic learning task is molecular tissue classification, described e.g. in [13, 10, 5]. The task consists in learning and application of a classifier via classification algorithm on the given data set. The classifier learns on samples with known classes and predicts classes in unknown samples. Class is represented by a type of the tissue, e.g. healthy or cancerous. An extensive overview of recent classification algorithms and tools applied to microarray data is given in [19]. [1]

Genomic classification belongs to the most difficult classification tasks as the data usually show an inconvenient rate of samples (tissues, biological situations) and variables (features, genes). Datasets are often noisy and contain a

[1] Other popular task is class discovery. It deals with unlabelled data and aims to identify new sample classes. Typically, it is carried out by clustering algorithms.

great part of variables irrelevant in the context under consideration. Hence dimension reduction of the feature space makes the crucial classification step. For this purpose, feature selection and feature extraction methods are used. Feature selection consists in selecting of relevant gene subsets that best discriminate different tissue classes. Feature extraction lies in transformation into a new feature space, the new features are calculated from the original set of genes.

High dimensionality of gene-expression data decreases accuracy and comprehensibility of genomic classifiers. Many research papers proved that application of both general and genome-specific dimension reduction techniques helps to increase classification accuracy ([19, 10]). However, contemporary genomics offers an opportunity to reach beyond the routine application of purely statistical dimension reduction techniques. Availability of a great variability of knowledge on gene roles, functions and gene-gene interactions allows to benefit from knowledge-based approaches to dimension reduction. For example, in the field of feature/gene selection, relevance of genes can be co-regulated by their apriori known affiliation to the domain of interest. Having two genes with similar statistical relevance to the target domain, the gene with explicit references to any of the classes in its background knowledge is preferred. In the field of feature/gene extraction, features with well-defined biological function or meaning constructed of functionally related genes can largely be promoted.

This paper deals with the apriori knowledge-based feature extraction. We proposed a feature extraction method exploiting knowledge from general genomic databases. Our method is compared with two common feature extraction methods: principal component analysis and prototype gene. Both of those methods are based purely on gene expression values, thus they are independent of biologic function of particular genes. Contribution of the above mentioned methods was tested and compared on three genomic datasets, the methods were evaluated in terms of accuracy of classifiers applied in transformed feature spaces.

The text is organized as follows. Section 2 describes existing methods of feature selection and extraction. Our proposed knowledge-based feature extraction method is introduced in Section 3. Section 4 describes genomic datasets used in experiments including their preprocessing. Performed experiments and their results are discussed in Section 5. And finally, conclusion and future work are summarized in Section 6.

2 Feature selection and extraction

In this section dimension reduction techniques are introduced. A gene expression sample is represented by the vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, where the p-th component $x_{ip} \in \mathbb{R}$, $p = 1, 2, \ldots, m$ represents a relative expression level of the gene g_p in this sample.

2.1 Feature selection

Feature selection (FS) consists in selection of a feature subset that allows or at least simplifies further robust learning. The authors of [18] divide FS methods into three groups. The first group concerns filters evaluating features by a chosen measure/rate. In genomics, the genes with the highest rank are selected. The measure is classifier independent, i.e. it does not take into account the succeeding learning algorithm. The second group are wrappers, which rank particular possible feature subsets by accuracy of the used classifier. The feature subset which results in the highest classifier accuracy is selected. Finally, the last group represents methods that include FS directly into classifier training. Further details on FS can be found e.g. in [21, 32]. In this paper, FS is not our main focus. It only serves to reduce the feature space prior to knowledge-based feature extraction (and any other comparative method of feature extraction). We use a simple filter selection method based on information gain. The rest of this section briefly describes this method.

Information gain Filter selection with evaluation by information gain (IG), described in [21, chapter 3], is a frequent method in genomics mainly for its simplicity and speed. IG measures expected entropy reduction and expresses decrease in class uncertainty with known values of the particular gene with respect to the situation in which the values are unknown.

Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\}$ be a data set, where $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, 2, \ldots, n$ are samples belonging into the class $y \in \{1, 2, \ldots, c\}$ with apriori probability $P(y)$. Then, entropy of the set $\mathcal{D}$ is

$$H(\mathcal{D}) = -\sum_{y=1}^{c} P(y) \log_2 P(y). \tag{1}$$

Next, let the gene g_p, $p = 1, 2, \ldots, m$ split the set $\mathcal{D}$ through its s different values into the disjoint subsets $\mathcal{D}_k^p$, $k = 1, 2, \ldots, s$. Then, IG of the gene g_p is

$$IG(\mathcal{D}, g_p) = H(\mathcal{D}) - \sum_{k=1}^{s} \frac{\mid \mathcal{D}_k \mid}{\mid \mathcal{D} \mid} H(\mathcal{D}_k^p). \tag{2}$$

After evaluation of all genes g_p, the dimension is reduced to the subset of k genes with largest IG. IG supposes nominal features, however gene expression values are numeric. Therefore, gene values must be discretized before IG can be enumerated.

2.2 Feature extraction

Feature extraction (FE), unlike feature selection, constructs new features from original ones. FE transforms the feature space from the original base to a new one. This transformation can be both linear and nonlinear. The most common linear FE method is principal component analysis (PCA). Many other, both linear and nonlinear, FE methods are based just on PCA. These and other methods are described e.g. in [21, 8, 11]. We will use PCA as one of the referential FE methods. Another referential method that we use is called principal gene (PG). In PG new features correspond to cluster centroids of genes showing similar expression profiles. The rest of this section explains both the above-mentioned methods.

Principal component analysis Principal component analysis (PCA), described in [27], consists in linear transformation of the original base to a new one with a reduced dimension and minimal reconstruction error.

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be a gene expression matrix, whose columns are formed by transposed samples $\mathbf{x}_i^T \in \mathbb{R}^m$, $i = 1, 2, \ldots, n$. Then, the covariance matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$ is

$$\mathbf{C} = \frac{1}{m-1}\mathbf{X}\mathbf{X}^T. \tag{3}$$

The eigenvalues $\lambda_j \in \mathbb{R}$, $j = 1, 2, \ldots, m$ of the matrix $\mathbf{C}$ and their corresponding eigenvectors $\mathbf{v}_j \in \mathbb{R}^m$ are computed. The value λ_j and the vector $\mathbf{v}_j$ are the eigenvalue and eigenvector of the matrix $\mathbf{C}$, if

$$(\mathbf{C} - \lambda_j \mathbf{I})\mathbf{v}_j = \mathbf{0}, \tag{4}$$

where $\mathbf{I} \in \mathbb{R}^{m \times m}$ is a unit matrix. From the eigenvectors $\mathbf{v}_j$ is selected k vectors, which correspond to k highest eigenvalues from λ_j. These k eigenvectors form rows of the transformation matrix $\mathbf{T} \in \mathbb{R}^{k \times m}$. The matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{k \times n}$ with a reduced dimension is defined as $\tilde{\mathbf{X}} = \mathbf{T}\mathbf{X}$. Its columns correspond to transposed samples $\tilde{\mathbf{x}}_i^T \in \mathbb{R}^k$ with the reduced dimension.

Prototype gene This method, proposed in [16], is based on gene clustering. The clustered objects correspond to genes, their distance is given by the Euclidean distance of their expression profiles. So-called prototype genes (PGs) agree with centroids of the found clusters and represent new features. A great variety of clustering algorithms can be used, in this paper we apply k-means algorithm described in [4]. Since the number of clusters is lower than the number of genes, the feature space represented by PGs has a smaller dimension than the original feature space represented by raw genes.

Let us have samples $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, 2, \ldots, n$. At first, the subset $\mathcal{K}$ is created. The subset $\mathcal{K}$ contains k randomly selected genes from the set of genes $\{g_p,\ p = 1, 2, \ldots, m\}$ of samples $\mathbf{x}_i$. Secondly, any centroid c_q, $q = 1, 2, \ldots, k$ of corresponding cluster $\mathcal{G}_q$ is initiated by gene expression values x_{iq} of gene $g_q \in \mathcal{K}$ in samples $\mathbf{x}_i$. Then in every iteration, the gene g_p is assigned into a cluster $\mathcal{G}_q$, whose centroid c_q has the minimal distance $d(c_q, g_p)$ from the gene g_p, where $d(c_q, g_p)$ is Euclidian distance between two genes. Provided the gene clusters $\mathcal{G}_q$ change between iterations, the centroids c_q are recomputed. The algorithm iterates until gene partitions change. Finally, the samples with reduced dimension $\tilde{\mathbf{x}}_i \in \mathbb{R}^k$ are described by the values of PGs, which correspond to the centroids c_q of clusters $\mathcal{G}_q$.

3 Knowledge-based feature extraction

In this section we introduce a novel apriori knowledge-based FE method called knowledge gene (KG). Apriori knowledge in this domain is represented by gene annotations from public genomic databases. In the first part of this section we describe the process of gene annotation retrieval. The proposed KG method is presented in the second part of the section.

3.1 Gene annotations

In this text, we understand the term genomic background (apriori) knowledge (BK) as any information that is not available in a gene-expression dataset but it is related to the genes contained in this dataset. There is a rich body of BK available for genes – a short textual description of gene function, its cellular location, bibliography, interactions and further links with other genes, memberships in pathways, referential sequences and many other pieces of information.

The first step in gene annotation is to match the gene identifiers used by microarray manufacturers (Probe IDs) and/or the SAGE identifiers (referential sequences) with the identifiers employed by the selected public biological datatbase(s). One of the most widely used interfaces for the retrieval of information from biological databases is the NCBI (National Center for Biotechnology) Entrez system (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`). The main goals and the scope of the system can be found in [22]. The system denotes genes by unique identifiers (Entrez Gene IDs). We used the NCBI Entrez system in our work, to convert from Probe IDs to Entrez Gene IDs we applied the interactive tool PANTHER (`http://www.pantherdb.org/genes/`) exploiting the database which is called PANTHER too, described in [23] and the Gene ID Conversion Tool (`http://david.abcc.ncifcrf.gov/conversion.jsp`) exploiting DAVID database, described in [26].

Having Entrez Gene IDs, we can annotate the gene set with measured expressions. Due to the gene annotation we can perform functional classification or functional clustering using existing common knowledge on the particular genes. These functional approaches can either extend or replace the classification and clustering methods based purely on gene expression values.

In this text we do not apply the whole body of available BK, but focus only on a part of this BK called SP_PIR_KEYWORDS. It provides annotation from two protein databases: Swiss-Prot (SP) and Protein Information Resource (PIR). For each gene the set of keywords/terms relevant to the gene is extracted from SP and PIR databases. The following subsection gives a way the terms can be employed to construct knowledge gene clusters.

3.2 Knowledge gene

The proposed method of gene extraction stems from the prototype gene method. Like PG it creates prototype genes via clustering. Unlike PG it exploits apriori knowledge rather than the gene expression profile distance when creating prototypes. It is based on the assumption that genes that share global functional annotation profiles are functionally related to each other. That is why we called it knowledge gene (KG). The distinction between PG and KG is in the gene similarity measure. PG uses the Euclidean distance computed from gene expression values in particular biological samples. By contrast, KG uses the gene similarity measure proposed in [17]. This method exploits BK for gene similarity measuring. In particular, the gene similarity is represented by terms co-occurrence in given genes. Degree of term co-occurrence in genes is represented by the kappa value. The kappa value is computed from a "flat" binary matrix of annotation terms assigned to the genes (the term can either be present or absent (valid or invalid) for the given gene). How is described above, the terms can be represented by different part of BK. However, in this paper we confine to keywords of particular genes extracted from two protein databases: SP and PIR.

Let us have genes g_u, g_v and a set of observed terms $\mathcal{T}$. Let C_{00} is the number of terms occurring neither in g_u nor g_v, C_{01} is the number of terms, which occur in g_u, but not in g_v, C_{10} is the number of terms, which occur in g_u, but not in g_v and finally, C_{11} is the number of terms observed both for g_u and g_v. Then, the kappa value between genes g_u, g_v is defined as:

$$K_{uv} = \frac{O_{uv} - A_{uv}}{1 - A_{uv}}, \tag{5}$$

$$O_{uv} = \frac{C_{11} + C_{00}}{\mid \mathcal{T} \mid}, \tag{6}$$

$$A_{uv} = \frac{C_{*1}C_{1*} + C_{*0}C_{0*}}{\mid \mathcal{T} \mid^2}, \tag{7}$$

where $C_{*1} = C_{01} + C_{11}$, $C_{1*} = C_{10} + C_{11}$, $C_{*0} = C_{00} + C_{10}$ and $C_{*0} = C_{00} + C_{10}$. Using the kappa value as the gene-similarity measure, gene clusters can be found. The standard k-means algorithm can be applied again (as well as a great variety of other clustering algorithms). However, we follow the default Heuristic Fuzzy Partition (HFP) algorithm, proposed also in [17]. Fuzziness of the algorithm makes possible that a gene can belong into more than one cluster simultaneously. On the contrary, a gene does not need to belong in any cluster. This characteristic complies with the needs of FE as it (1) eliminates irrelevant/weak elements, as orphan genes, to remove noise, (2) its fuzziness suits the biological nature of overlapping gene functions, features can correspond to biological functions and (3) dynamically determines the number of clusters based on the chosen threshold.

Using of the HFP algorithm is following. Let us have samples $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, 2, \ldots, n$. At first, the subset $\mathcal{K}$ is created among genes g_p, $p = 1, 2, \ldots, m$ of samples $\mathbf{x}_i$. The subset $\mathcal{K}$ is created independently on gene expressions, therefore independently on samples $\mathbf{x}_i$. The gene g_p falls into the set $\mathcal{K}$, if at least l genes exist among genes g_q, $q = 1, 2, \ldots, m$, $q \neq p$, so that $K_{pq} \geq \xi$, where l and ξ are chosen constants. Then genes $g_r \in \mathcal{K}$ are medoids of the clusters $\mathcal{G}_r$ (medoid is the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal). Each cluster contains together with gene g_r also genes g_s, for which $K_{rs} \geq \xi$. Clusters $\mathcal{G}_r$ form a set $\mathcal{C}$. Then clusters from the set $\mathcal{C}$ are iteratively merged into larger clusters. The clusters $\mathcal{G}_e, \mathcal{G}_f \in \mathcal{C}$ are merged into one cluster, if

$$\frac{\mid \mathcal{G}_e \cap \mathcal{G}_f \mid}{\max(\mid \mathcal{G}_e \mid, \mid \mathcal{G}_f \mid)} \geq 0.5. \tag{8}$$

Finally, the samples with a reduced dimension $\tilde{\mathbf{x}}_i \in \mathbb{R}^k$ are described by values of knowledge genes, which correspond to centroids c_q of clusters $\mathcal{C}$.

4 Data and preprocessing

In this section, we describe three datasets used in performed experiments. The first and second datasets were created via microarray technology, whose emergence is described in [20]. The third dataset was created via serial analysis of gene expression (SAGE), proposed in [30].

4.1 Motol

This dataset contains gene expression samples of patients with a removed tumor of urinary bladder. The samples are split into two classes depending on whether tumor relapse happened in the given patient or not. This dataset contains 22 samples. 12 patients developed tumor relapse while remaining 10 patients did not. The dataset was prepared on the Applied Biosystems 1700 microarray platform, which is described in [1]. Each sample is characterized through expression measurement of 37 842 genes. The microarray measures several expression characteristics for each gene and each sample. All of these attributes are in detail described in [2]. Especially, SIGNAL, SDEV, S/N and FLAG were important for our work, i.e. served in computation of the final expression matrix.

Preprocessing of this dataset was carried out several steps. The initial steps were recommended in [2]: (1) remove genes labeled as NULL-GENE as these do not correspond to any gene and serve as control probes and (2) find the slots where S/N <3 or FLAGS $>8\,191$ and mark the corresponding SIGNAL values as missing. The further steps followed the general recommendations proposed in [28, 24, 12] and focused purely on the SIGNAL values: (3) remove genes with more than 5 % missing values, (4) carry out logarithmic transformation with base equals 2, (5) replace missing values through $k-$NN impute ($k = 10$). Finally: (6) apply the quantile normalization proposed in [6]. By application of these steps we reduced the size of feature space to 10 961 genes.

4.2 ALL/AML

The ALL/AML dataset, described in [13], contains gene expression samples of tumor tissues of acute leukemia patients. These samples belongs to one of two

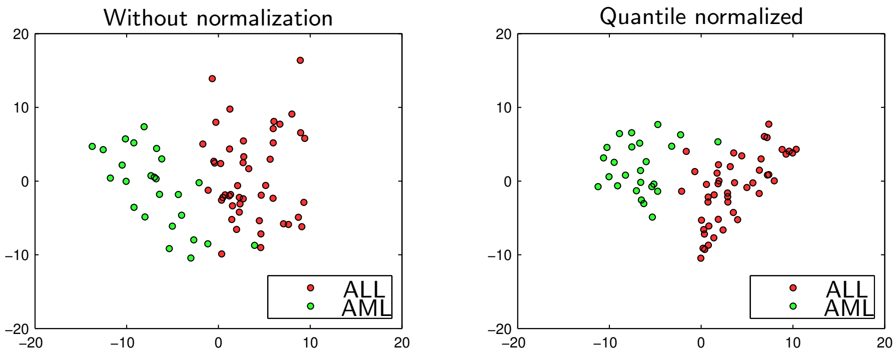

Fig. 1. Projection of ALL/AML samples into a two dimensional space by multidimensional scaling

classes in agreement with leukemia type of given patient: acute lymphoblastic leukemia (ALL) or acute myeloid leucemia (AML). The dataset contains samples of 72 patients, 47 patients of which belong into ALL class and 25 patients of which belong into AML class. The dataset was created on Affymetrix microarray platform. Every sample is described by 7 129 genes. The dataset is well-known and frequently referenced and serves as a benchmark in our study. For illustration we carried out projection of ALL/AML samples into a two dimensional space by multidimensional scaling, described in [29]. This projection you can see on figure 1.

The dataset was preprocessed in the following steps. The first and second step were recommended in [10]: (1) threshold with floor 100 and ceiling 16 000 (the values smaller than floor or higher than ceiling are replaced by floor or ceiling), (2) remove genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min represent maximal and minimal value for the given gene over every sample, (3) carry out the base 10 logarithmic transformation, (4) normalize by quantile normalization. In this manner we reduced the size of feature space to 3 571 genes.

4.3 SAGE

The SAGE dataset contains non-cancerous and cancerous gene expression samples. The cancerous class consists of samples created from tissues with various types of cancer. The non-cancerous class contains samples created from both healthy tissues and tissues with other diseases than cancer, e.g. HIV. This dataset includes 207 samples, 130 of which were created from cancerous tissues and remaining 77 from non-cancerous tissues. The dataset was created by SAGE technology. Every sample is characterized by expression of 11 082 genes.

The data was preprocessed in a similar manner as described in [14]. Preprocessing was done in the following steps: (1) remove the genes with zero expression value in more than 90 % of all samples (in [14] the authors removed the genes with zero expression in 10 samples at least), (2) carry out the scale normalization (the expression sum in a sample was scaled to 10 000), (3) do the natural logarithmic transformation without transformation of null values. In this way we reduced the size of feature space to 5 318 genes.

5 Experiments and results

5.1 Experiments design

We observed effect of three different FE methods on accuracy of classifiers built in the reduced feature space. The hypothesis is that the knowledge-based FE generates more robust and informed features than its expression-based counterparts which results in increase of classification accuracy. The datasets were preprocessed in the steps described in Section 4. The extraction methods PCA and PG described in 2.2 and the novel FE method KG described in 3 were combined with the FS by IG described in 2.1.

Classifiers were created on Motol, ALL/AML and SAGE datasets, 4 different classification algorithms were used: random forests (RF) (500 trees from $\sqrt{n}$ random features, where n is the size of original dimension), described in [7]; k-nearest neighbor (KNN) ($k = 5$), described in [9]; support vector machines (SVM), described in [15] and C4.5, described in [25]. Classification accuracy was estimated by leave-one-out cross-validation (LOOCV) on Motol dataset and via 10 cross-validation (10CV) on ALL/AML and SAGE datasets.

The data was separated into train and test set in any fold. The learning and evaluation procedure in every particular fold was following. At first, IG selection was applied on the training set. The genes with zero information gain were removed. Then, feature extraction via PCA, PG or KG (the terms are represented by keywords of particular genes extracted from two protein databases: SP and PIR) were applied on the training set with selected genes. For the PCA and PG methods we had to set a number of extracted features explicitly. We extracted $1, 2, 5, 10, \ldots, s$ features, where s is the minimal number of genes with nonzero information gain of all folds. For the KG method, the number of extracted features is given implicitly by the clustering algorithm. Classifiers were learned on training set with extracted features. Accuracy of these classifiers was estimated on test set. Eventually, the final accuracy is the mean accuracy of particular folds.

For implementation of these experiments we exploited several software tools. In any experiments we combined MATLAB [3] and WEKA, described in [31]. In the experiments exploiting KG, we used DAVID functional annotation clustering (`http://david.abcc.ncifcrf.gov/summary.jsp`), in addition.

5.2 Results

In Table 1 we can see a comparison of common FE methods PCA and PG with our proposed apriori knowledge-based method KG. The PCA and PG accuracies

Dataset	Reduction	RF	KNN	SVM	C4.5	Mean
Motol	IG + PCA	33.52	28.98	34.09	50.00	36.65
	IG + PG	51.14	44.32	40.34	52.84	47.16
	IG + KG	63.64	59.09	72.73	54.55	62.50
ALL/AML	IG + PCA	90.69	94.16	77.08	94.44	89.09
	IG + PG	90.83	91.80	91.80	87.64	90.52
	IG + KG	94.44	94.44	97.22	95.83	95.48
SAGE	IG + PCA	73.48	71.23	68.12	71.66	71.12
	IG + PG	72.14	70.91	72.89	67.79	70.93
	IG + KG	74.88	76.33	71.01	72.95	73.79

Table 1. Accuracy of classifiers on Motol, ALL/AML and SAGE datasets with the reduced dimension. The classifiers were created by RF, KNN, SVM and C4.5 algorithms. The classifier accuracy was estimated via LOOCV on Motol dataset and via 10CV on ALL/AML and SAGE dataset. The datasets were preprocessed in the steps described in Section 4. The IG selection was applied first, thereby minimally n genes with a non-zero information gain was selected in each fold. Next, feature extraction via PCA, PG or KG was carried out on the subset of selected genes. The column Mean aggregates the results reached by different classifier types.

are means over accuracies on data with various numbers of extracted features $(1, 2, 5, 10, \ldots, s)$. In KG, the number of extracted features is given implicitly by the clustering algorithm HFP on particular data.[2] We can see, that classifiers based on KG features get in most cases better accuracy than classifiers based on PCA or PG features. Although the datasets do not cover the full scale of gene-expression data and statistical significance of the difference in accuracies cannot be proven, the results suggest that exploiting apriori knowledge in FE can get more accurate and interpretable classifiers than common FE methods based only on gene expression values.

6 Conclusion

In this paper, we have presented the apriori knowledge-based feature extraction method for genomic data. Our method is inspired by prototype genes and consists in construction of so-called knowledge genes. Compared to PG and other methods based only on gene expression values, the method makes use of co-occurrence of the keywords/terms in gene annotations extracted from two protein databases: SW and PIR.

We compared the proposed method with common purely gene expression-based FE methods. In particular, we compared our method with PCA and PG. We observed how these ways of data preprocessing influence accuracy of classifiers. Classifiers were created by RF, KNN, SVM and C4.5 learning algorithms on Motol, ALL/AML and SAGE datasets. It was experimentally demonstrated

[2] It is good to mention, that this discrepancy can have some influence on PCA, PG and KG comparability.

that the KG method can achieve better accuracy of classifiers than common FE methods. The presented results showed potential of apriori knowledge-based FE in genomics.

In future work, we would like to extend the performed experiments with a broader parameter setting. The parameters to be set are mainly term representation and clustering algorithm thresholds. To standardize and fix conditions of different FE methods, a uniform clustering algorithm shall be used across the methods. The next possibility to improve prototype gene extraction is to combine background knowledge with expression measurements. In other words, the gene distance measure can be based both on co-occurrence of terms in particular genes and co-expression of these genes. In fact it consists in a combination of prototype and knowledge gene methods. We could also use other terms in place of keywords. From larger perspective, one can proceed towards a real information fusion and make the best of a wider scale of heterogeneous knowledge types available for genes.

Acknowledgements

The work of Miloš Krejník has been supported by the grant 1ET101210513 “Relational Machine Learning for Analysis of Biomedical Data” funded by the Czech Academy of Sciences. The work of Jiří Kléma was funded by the Czech Ministry of Education in terms of the research programme “Transdisciplinary Research in the Area of Biomedical Engineering II”, MSM 6840770012.

References

1. Applied Biosystems expression array system. System profile, Applied Biosystems, Foster City, CA, USA, 2004.
2. Applied Biosystems 1700 chemiluminescent microarray analyzer. User bulletin, Applied Biosystems, Foster City, CA, USA, 2005.
3. MATLAB R2006a (version 7.2.0.232). Software, The MathWorks, Natick, MA, USA, 2006.
4. K. Alsabti, S. Ranka, and V. Singh. An efficient k-means clustering algorithm. *First Workshop on High-Performance Data Mining*, 10(1), 1998.
5. A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. In *Proc. of the fourth annual int. conf. on computational molecular biology*, pages 54–64, 2000.
6. B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.
7. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
8. M. A. Carreira-Perpiñán. A review of dimension reduction techniques. Technical Report CS–96–09, Department of Computer Science, University of Sheffield, Sheffield, South Yorkshire, UK, 1997.
9. T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

10. S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.
11. I. K. Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Laboratory, Livermore, CA, USA, 2002.
12. X. Gan, A. W.-C. Liew, and H. Yan. Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Research*, 34(5):1608–1619, 2006.
13. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
14. O. L. Griffith, E. D. Pleasance, D. L. Fulton, M. Oveisi, M. Ester, A. S. Siddiqui, and S. J. Jones. Assessment and integration of publicly available sage, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics*, 86:476–488, 2005.
15. S. R. Gunn. Support vector machines for classification and regression. Technical report, University of Southampton, 1998.
16. B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clément, and J.-D. Zucker. Improving classification of microarray data using prototype-based feature selection. *SIGKDD Explorations Newsletter*, 5(2):23–30, 2003.
17. D. W. Huang, B. Sherman, Q. Tan, J. Collins, W. Alvord, J. Roayaei, R. Stephens, M. Baseler, H. Lane, and R. Lempicki. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8:R183–R198, 2007.
18. E. B. Huerta, B. Duval, and J. K. Hao. *A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data*, volume 3907/2006, pages 34–44. Springer Berlin/Heidelberg, 2006.
19. J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.
20. T. Lenoir and E. Giannella. The emergence and diffusion of DNA microarray technology. *Journal of Biomedical Discovery and Collaboration*, 1:11–49, 2006.
21. H. Liu and H. Motoda. *Feature extraction, construction and selection: a data mining perspective*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.
22. D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res*, 33(Database issue), 2005.
23. H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, and P. D. Thomas. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*, 33(Database issue), 2005.
24. D. V. Nguyen, N. Wang, and R. J. Carroll. Evaluation of missing value estimation for microarray data. *Journal of Data Science*, 2(4):347–370, 2004.
25. J. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.
26. B. Sherman, D. Huang, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M. Baseler, H. Lane, and R. Lempicki. DAVID knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, 8:426, 2007.
27. J. Shlens. A tutorial on principal component analysis, 2005.

28. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
29. J. Tzeng, H. H. Lu, and W.-H. Li. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, 9:179–195, 2008.
30. V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.
31. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, 2005.
32. E. P. Xing, M. I. Jordan, and R. M. Karp. Feature selection for high-dimensional genomic microarray data. In *Proc. of the Eighteenth Int. Conf. in Machine Learning*, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001.