Fuzzy Clustering and Gene Ontology Based Decision Rules for Identification and Description of Gene Groups

Aleksandra Gruca, Michał Kozielski, and Marek Sikora

Abstract. The paper presents results of the research verifying whether gene clustering that takes under consideration both gene expression values and similarity of GO terms improves a quality of rule-based description of the gene groups. The obtained results show that application of the Conditional Robust Fuzzy C-Medoids algorithm enables to obtain gene groups similar to the groups determined by domain experts. However, the differences observed in clustering influences a description quality of the groups. The rules determined cover more genes retaining their statistical significance. The rules induction and post-processing method presented in the paper takes under consideration, among others, a hierarchy of GO terms and a compound measure that evaluates the generated rules. The approach presented is unique, it makes possible to limit a number of rules determined considerably and to obtain rules that reflect varied biological knowledge even if they cover the same genes.

Keywords: clustering, decision rules, microarray analysis.

1 Introduction

The analysis of the data obtained in the DNA microarray experiment is a complex process involving application of many different methods including statistical analysis and data mining techniques. Such analysis usually consist of identification of differentially expressed genes, application of the algorithms grouping together genes with similar expression patterns and application of the methods for interpretation of the biological functions of the coexpressed genes.

One of the most popular tools used for annotation of the genes and gene products is Gene Ontology (GO) database which provides functional annotation of genes [2].

Aleksandra Gruca · Michał Kozielski · Marek Sikora
Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
e-mail: {Aleksandra.Gruca,Michal.Kozielski,Marek.Sikora}@polsl.pl

K.A. Cyran et al. (Eds.): Man-Machine Interactions, AISC 59, pp. 141–149.
springerlink.com

Gene Ontology database includes hierarchical and structured vocabulary that is used to describe genes and gene products. The information in database is represented in a form of three disjoint directed acyclic graphs (DAGs) describing biological process, molecular function and cellular component. Each node of the DAG is called GO term and it is a single unit describing gene or gene product. The structure of database is hierarchical, GO terms that are close to the root describe general concepts and as a DAG is traversed from the root to the leaves, the concepts represented by GO terms are more specific.

In this paper we present a method that enables to describe gene groups by means of decision rules based on GO terms. Typical analysis that use decision rules to process results of the DNA microarray experiments involve inducing the rules with expression values in a rule conditional part [4, 12]. In our approach we use decision rules to describe the results of DNA microarray experiment. In such case, conditional part of the rule includes GO terms.

For each decision rule we compute its statistical significance and as a result of analysis we present only statistically significant rules. Since we are interested in such rules only, covering whole genes belonging to a group by means of determined rules is usually not possible. Even if the obtained rules do not describe all genes from the group, the result set includes so many rules that it is very difficult for an expert to interpret them.

Therefore, to improve the quality of rule-based gene group descriptions we propose a new method of rules evaluation and filtration. Wishing to increase a covering degree of gene groups by determined rules we propose a method of clustering the results of DNA microarray experiments which combines two sources of information: gene expression values and GO terms of that genes. We apply proposed approach to several well-known fuzzy clustering algorithms and compute decision rules for obtained clusters.

1.1 Gene Ontology

Formally, Gene Ontology is a directed acyclic graph $GO = (A, \leq)$, where A is a set of terms describing genes and gene products and $\leq$ is a binary relation on A such that genes described by a_j are a subset of genes described by a_i, denoted $a_j \leq a_i$, if and only if there exists a path $a_i, a_{i+1}, \ldots, a_{j-1}, a_j$ such that $a_{m-1} \leq a_m$ for $m = i+1, i+2, \ldots, j-1, j$ (we can find here an analogy with the inverse Hasse diagram for ordering relations). Relation $\leq$ is an order relation (reflexive, antisymmetric and transitive).

Each GO term has a level assigned which is defined in the following way: the ith level of the graph is formed by all GO terms $a \in A$ for which there exists a path $(root, a_1, \ldots, a_{i-1}, a_i)$ such that: $a_1 \leq root$, $a_m \leq a_{m-1}$ for $m = 2, 3, \ldots, i-1$ and $a_i \leq a_{i-1}$ (in other words there exists a path of length i from the root to that term).

GO terms can be used as a tool for gene annotations. Each annotation is an association between gene and the GO term describing it, so if we have a set of genes G there are some nodes in GO graph that are annotated with genes from the set G.

Considering the relation $\leq$ and the definition of GO graph, if there are two GO terms such that: $a_j \leq a_i$, $a_i, a_j \in A$, we can assume that any gene that is annotated with the GO term a_j is also annotated with the GO term a_i. In this paper we construct a GO graph based on the such assumptions. We call this graph *GO-inc*.

2 Rule Induction

Let there be given: a set G of genes, a set A of GO terms that create *GO-inc* and n gene groups $(G(1), G(2), \ldots, G(n))$. It is possible to create a decision table $DT = (G, A \cup \{d\})$, where for all $a \in A$, $a\colon U \to \{0,1\}$, and for all $g \in G$, $d(g) \in \{G(1), G(2), \ldots, G(n)\}$. In this set we search for all statistically significant rules of the following form:

$$\text{IF } a_{i1} = 1 \text{ and } a_{i2} = 1 \text{ and } \ldots \text{ and } a_{ik} = 1 \text{ THEN } d = G(l), \tag{1}$$

where: $\{a_{i1}, a_{i2}, \ldots, a_{ik}\} \subseteq A$, $G(l) \in \{G(1), G(2), \ldots, G(n)\}$.

A rule of the form (1) should be interpreted as follows: if a gene is simultaneously described by terms occurring in a premise of the rule, then it belongs to the gene group $G(l)$.

We denote a set of rules with identical conclusions by $RUL_{G(l)}$ and call the description of the gene group $G(l)$. A set of genes which are described by terms occurring in a premise of a rule r we denote by $match(r)$. By $supp(r)$ we denote these genes belonging to $match(r)$ which also belong to the gene group occurring in the conclusion of r. For each rule $r \in RUL_{G(l)}$ the accuracy of r is given by the formula $acc(r) = supp(r)/match(r)$ and coverage of r is defined as $cov(r) = supp(r)/|G(l)|$. We use hypergeometric test to verify whether a determined rule is statistically significant.

In the field of our interests are all rules with the value p-value less or equal to a value established by a user. In worst case we have to determine $2^{|A|} - 1$ rules, what is impossible in the case of big number of considered ontological terms. Therefore, the EXPLORE [15] algorithm proposed by Stefanowski is more suitable for our aims. For our purposes the algorithm underwent a few modifications. Searching space of potential candidates for rules is made by means of a procedure that is iteratively repeated for each gene group. The main part of the algorithm generates premises with increasing number of terms, beginning from premises containing one GO term. While a rule – candidate achieves desired p-value, it is added to a result rule set and a conjunction is widen on. If, for a given premise, all GO terms were already considered, then a new GO term is selected (not selected yet) and a new rule creation begins. In order to narrow the searching space the following solutions were applied:

- After adding a term a to the rule premise ϕ, no terms lying on any path (from root to leaf on the ontology) that leads to the element a are considered. Let us notice that for any term $b \in A$ for which $b \leq a$ or $a \leq b$, the conjunction $b \wedge a \wedge \phi$ amount

to $a \wedge \phi$ or to $b \wedge \phi$. There is no point to consider conjunction $b \wedge \phi$ because it was considered during induction of a rule including GO term b.
- Assuming that currently created rule has the form $\phi \rightarrow \psi$, the term a will be added to its premise forming the rule $\phi \wedge a \rightarrow \psi$, if $acc(\phi \rightarrow \psi) < acc(a \rightarrow \psi)$. The condition limits a number of GO terms added that do not contribute to improving rule accuracy.

The rule set determined in this way may be large. Therefore, a method of rules evaluation and filtration is required. Apart from statistical significance, the quality of a rule $r \in RUL_{G(l)}$ is also important [1, 14]. To asses the rules quality we modified $\mathrm{WS}^{\mathrm{Yails}}$ measure:

$$mWS^{\mathrm{Yails}}(r) = (0.5 + 0.25acc(r))acc(r) + (0.5 - 0.25cov(r))cov(r). \quad (2)$$

Another quality criterion of the rules determined is the number of GO terms occurring in a rule premise. By $Lenght(r)$ we denote the normalized number of GO terms occurring in the premise of r. We assume that the bigger number of GO terms occurring in the rule premise the more information is included in the rule (we remind that terms occurring in a premise do not lie on common path in ontology graph).

The last quality criterion is a level of GO terms occurring in the rule premise:

$$Depth(r) = \frac{\sum_{i=1}^{NoGOterms(r)} level(a_i)}{\sum_{i=1}^{NoGOterms(r)} max\_path(a_i)}, \quad (3)$$

where: $level(a_i)$ is the level of a GO term a_i that occurs in the rule premise; $max\_path(a_i)$ is the longest path leading from the root to a leaf of GO that passes trough the node described by a_i, and $NoGOterms(r)$ is the number of GO terms occurring in the premise of r. From a description point of view we should prefer rules with premises including terms from as low level of the GO graph as possible.

Finally, a measure that enables to evaluate a rule quality is a product of all component measures:

$$Q(r) = mWS^{Yails}(r) \times Length(r) \times Depth(r)\ . \quad (4)$$

A filtration algorithm that uses rules ranking obtained on basis of the above measure is executed in a loop. Starting from the best rule in the ranking all rules covering the same set of genes (or its subset) are candidates to be removed from the result rule set. However, before any rule is removed its similarity to the reference rule is verified. That similarity is determined by (5). If a rule is similar to the reference rule in more than 50%, it is removed from the set of determined rules, otherwise it remains in an output rule set.

$$Similarity(r_1, r_2) = \frac{UniqGOterms(r_1, r_2) + UniqGOterms(r_2, r_1)}{NoGOterms(r_1) + NoGOterms(r_2)}, \quad (5)$$

where: $UniqGOterms(r_i, r_j)$ is a number of unique GO terms occurring in the rule r_i and not occurring in the rule r_j. The GO term a from the rule r_i is recognized as the unique if it does not occur directly in the rule r_j and there is no path in GO graph that includes both term a and any term b from rule r_j premise.

3 Clustering Methods

In the work presented genes are the multi–represented data objects described by the expression values and by the annotations to Gene Ontology. In order to cluster genes considering these two sources of information a special approach is needed.

Distance of the genes described by means of numeric expression values may be calculated applying Euclidean distance or correlation coefficient [5]. Similarity of the genes described by means of GO terms encoded to the form of binary annotation table may be calculated applying the concept of *Information Content* $I(a)$ of an ontology term $a \in A$ given by the following formula:

$$I(a) = -\ln(P(a)), \tag{6}$$

where $P(a)$ is a ratio of a number of annotations to term a to a number of analysed genes.

In order to calculate the similarity $S_A(a_i, a_j)$ of the ontology terms *Information Content* of the terms and their common ancestor $I_{\mathrm{ca}}(a_i, a_j)$ are applied in the following formula [9]:

$$S_A(a_i, a_j) = \frac{2I_{\mathrm{ca}}(a_i, a_j)}{I(a_i) + I(a_j)}. \tag{7}$$

Next, the similarity $S_G(g_k, g_p)$ between genes g_k and g_p can be calculated according to the following formula [3]:

$$S_G(g_k, g_p) = (m_k + m_p)^{-1}\left(\sum_i \max_j(S_A(a_i, a_j)) + \sum_j \max_i(S_A(a_i, a_j))\right), \tag{8}$$

where m_k is a number of annotations of gene g_k.

When being able to measure similarity or distance between analysed data objects, it is possible to apply one of the clustering algorithms. However, it must be the method suitable for complex multi-represented data. Thus, combinations of different fuzzy clustering algorithms were analysed.

Apart from the basic fuzzy clustering algorithm (FCM – Fuzzy C-Means), in our research we applied several modifications of the algorithm. The Conditional Fuzzy C-Means algorithm [13] is FCM based method which enables setting a condition on data objects modifying their impact on clustering process. Robust Fuzzy C-Medoids algorithm [8] enables clustering relational data (compared by means of similarity matrix) where computation of cluster prototypes is not possible. Conditional Robust Fuzzy C-Medoids algorithm is a modification of RFCMdd method enabling application of condition on data objects. Proximity-based Fuzzy C-Means

algorithm [10] is FCM based method which enables applying an expert knowledge in the form of proximity matrix to the clustering process.

Using the algorithms mentioned it is possible to propose three approaches to clustering genes described by microarray expression values and Gene Ontology annotations:

- cluster expression data by means of FCM algorithm and apply a resulting fuzzy partition matrix as a condition parameter to CRFCMdd algorithm run on ontology annotations, which is referenced further as CRFCMdd,
- cluster ontology annotations by means of RFCMdd algorithm and apply a resulting fuzzy partition matrix as a condition parameter to CFCM algorithm run on expression data, which is referenced further as CFCM,
- cluster expression data by means of PFCM algorithm applying a distance matrix calculated for ontology annotations as the proximity hints.

4 Experiments

Experiments were conducted on two freely available data sets: YEAST and HUMAN. The data set YEAST contains values of expression levels for budding yeast *Saccharomyces cerevisiae* measured in several DNA microarray experiments [5]. Our analysis was performed on 274 genes from 10 top clusters presented in the paper [5]. The data set HUMAN contains values of expression levels of human fibroblast in response to serum [7]. In the paper [7], 517 YEAST sequences were reported and divided into 10 clusters. After translation of the sequences for unique gene names and removal sequences that are duplicated or that are currently considered to be invalid, we obtained set of 368 genes. Then, each gene from YEAST and HUMAN sets were described by GO terms from Biological Process (BP) ontology. There were some genes in the HUMAN data set that had no GO term from BP ontology assigned, so we removed them from further analysis. After that step we obtained set consisting of 296 objects. To induce decision rules we created decision tables on the basis of the *GO-inc* graph for BP ontology. We used GO terms from at least second (for HUMAN set) and third (for YEAST set) ontology level and describing at least five genes from our data sets. After removing from genes description GO terms that did not fulfill this condition we had to remove two more genes from HUMAN data set. Finally we obtained two decision tables: decision table for YEAST data set consisting of 274 objects (genes) described by 244 attributes (GO terms) and decision table for HUMAN data set consisting of 294 objects described by 358 attributes.

The following parameter values were applied to the clustering algorithms analysed. All the clustering algorithms use a number of clusters to be created which was set to $c = 10$. The value of the parameter m impacting the fuzziness of the partition was set to $m = 1.1$. The quality of resulting fuzzy partition determined by each clustering algorithm was assessed on the basis of quality index presented in [11]. PFCM algorithm performs gradient optimization using additional parameter α

Table 1 YEAST results

Algorithm	[%] Average coverage	Before filtration		After filtration	
		Rules	Average p-val	Rules	Average p-val
Eisen	89.5	105306	0.00122	100	0.00080
FCM	93.0	111699	0.00182	108	0.00100
CRFCMdd	90.6	106750	0.00146	106	0.00107
CFCM	94.8	81210	0.00181	102	0.00102
PFCM	98.4	62184	0.00153	90	0.00103

Table 2 HUMAN results

Algorithm	[%] Average coverage	Before filtration		After filtration	
		Rules	Average p-val	Rules	Average p-val
Iyer	54.7	65582	0.00438	149	0.00416
FCM	58.2	74106	0.00487	107	0.00398
CRFCMdd	61.9	49780	0.00466	119	0.00463
CFCM	57.5	46076	0.00491	96	0.00495
PFCM	58.6	72720	0.00426	101	0.00366

which was set to $\alpha = 0.0001$. Clustering Gene Ontology data was performed only on Biological Process part of the ontology.

In case of YEAST data for genes similarity calculation during clusterization the correlation coefficient [5] was applied, and in case of HUMAN data the Euclidean distance was used. In the rule induction algorithm a number of terms occurring in a rule premise was limited to five, and the significance level was established on 0.01. Results of experiments on YEAST and HUMAN datasets are presented in Tables 1 and 2, respectively. Clustering results combined with expert (biological) analysis of the problem are presented in the first row and the results of clustering expression values only by means of FCM algorithm are presented in the second row of the tables.

Clustering results presented in the tables are quantitative results. Groups created by means of PFCM algorithm enable to generate the rules of the best coverage of the groups created. Average statistical significance of the rules does not differ significantly from results obtained by means of other clustering methods. A number of generated rules (after filtration) is also the smallest. However, to evaluate the quality of groups obtained exhaustively, quality analysis consisting of a full biological interpretation of the groups and the rules determined from them is needed. Such analysis is beyond the subject area of the paper. Assuming that a reference partition into groups is the partition presented by Eisen and Iyer, we may compare mutual covering of generated groups. The algorithm CRFCMdd that covers model groups to superlative degree, seems to be the best in the case of such comparison. It is also important that genes migration observed among groups is bigger for HUMAN set than for YEAST set.

The above observation confirms the fact that partition into groups of YEAST set proposed in [5] is probably the best possible partition. In case of HUMAN set and algorithm CRFCMdd, genes migrations among groups are meaningful. We also obtained significantly greater coverage. As a matter of fact, we could not determine a statistically significant rules for one of the groups from the obtained partition, however, this group contained very few objects.

Regardless of a clustering algorithm, the proposed rule induction, evaluation and filtration method that uses GO terms for groups description always enables to obtain not large set of rules (on average 10 rules per group) describing of gene groups. Such description is composed of the rules with a given (or better) statistical significance. The chosen rules induction algorithm guarantees that all possible combinations of ontological terms are considered. The applied method of rules evaluation and filtration guarantees that the rules filtered cover the same genes as input (unfiltered) rules and that induced rules are specific in the sense that they use GO terms from the lowest possible level in the ontology graph. It is important that each pair of rules has to differ in two ways: either cover different genes or differ from one another by at least 50% of terms (considering similarity of terms lying on the same path) occurring in premises. Such approach allows to obtain a description which includes various aspects of biological functions of described genes.

5 Conclusions

In this paper the proposal of considering both gene expression levels and their position in the ontology graph was presented. Several algorithms that enable to combine clustering such data representing different types of information were analyzed. Considering the similarity to the reference groups, the algorithm CRFCMdd combined with FCM turned out to be the best. In that way we obtained better gene groups coverage by statistically significant rules. In relation to the coverage and rules number better results were obtained using the PFCM algorithm, but genes migration among groups was then considerable and evaluation of partition quality requires deeper qualitative (biological) analysis.

The presented method of rules induction, evaluation and filtration appeared to be very effective. After filtration we obtained small rule sets having average statistical significance better than for the unfiltered rule set.

It is worth highlighting that the rules are generated only for description purposes. Due to the structure of the rules (only the descriptors corresponding to the terms having gene annotations were considered) a part of the rules determined though statistically significant is approximate and therefore genes classification by means of the rules determined could be incorrect in many cases.

Future research will concentrate on determining the rules including descriptors referencing GO terms which does not describe the analysed gene. In that case we will be interested in occurrence of this type of descriptors on the highest level of the ontology. The appropriately modified version of LEM algorithm [6] will be implemented in order to accelerate the calculations. The rules obtained by LEM

and EXPLORE algorithms after filtration will be compared. Considering clustering methods our research will focus on defining other than *Information Content* (6) measure of gene similarity.

References

1. An, A., Cercone, N.: Rule quality measures for rule induction systems description and evaluation. Computational Intelligence 17, 409–424 (2001)
2. Ashburner, M., Ball, C.A., Blake, J.A., et al.: Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
3. Azuaje, F., Wang, H., Bodenreider, O.: Ontology-driven similarity approaches to supporting gene functional assessment. In: Proceedings of the 18th Annual Bio-Ontologies Meeting, Michigan, US (2005)
4. Carmona-Sayez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M., Pascual-Montano, A.: Integerated analysis of gene expression by association rules discovery. BMC Bioinformatics 7(1), 54 (2006)
5. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Science 95, 14, 863–14, 868 (1998)
6. Grzymala-Busse, J.W., Ziarko, W.: Data mining based on rough sets. In: Wang, J. (ed.) Data Mining Opportunities and Challenges, pp. 142–173. IGI Publishing, Hershey (2003)
7. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., Brown, P.O.: The transcriptional program in the response of human fibroblasts to serum. Science 283, 83–87 (1999)
8. Krishnapuram, R., Joshi, A., Nasraoui, O., Yi, L.: Low-complexity fuzzy relational clustering algorithms for web mining. IEEE Transactions on Fuzzy Systems 9(4), 595–607 (2001)
9. Kustra, R., Zagdański, A.: Incorporating gene ontology in clustering gene expression data. In: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (2006)
10. Loia, V., Pedrycz, W., Senatore, S.: P-FCM: a proximity-based fuzzy clustering for user-centered web applications. International Journal of Approximate Reasoning 34, 121–144 (2003)
11. Łęski, J., Czogała, E.: A new artificial neural network based fuzzy inference system with moving consequents in if-then rules and its applications. Fuzzy Sets and Systems 108(3), 289–297 (1999)
12. Midelfart, H.: Supervised learning in gene ontology Part I: A rough set framework. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets IV. LNCS, vol. 3700, pp. 69–97. Springer, Heidelberg (2005)
13. Pedrycz, W.: Conditional fuzzy c-means. Pattern Recognition Letters 17, 625–631 (1996)
14. Sikora, M.: Rule quality measures in creation and reduction of data role models. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS, vol. 4259, pp. 716–725. Springer, Heidelberg (2006)
15. Stefanowski, J., Vanderpooten, D.: Induction of decision rules in classification and discovery-oriented perspectives. International Journal on Intelligent Systems 16(1), 13–27 (2001)