STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Michał KOZIELSKI, Aleksandra GRUCA
Politechnika Śląska, Instytut Informatyki

VISUAL COMPARISON OF CLUSTERING GENE ONTOLOGY WITH DIFFERENT SIMILARITY MEASURES

Summary. The work presents comparison of four Gene Ontology term similarity measures combined with two methods calculating gene similarity on the basis of terms similarity. Visual comparison of clustering results, where different clustering methods were applied, indicates the best combination of similarity methods that can be utilised in a clustering process.

Keywords: Gene Ontology clustering, term, similarity, gene similarity

WIZUALNE PORÓWNANIE GRUPOWANIA ONTOLOGII GENOWYCH PRZY ZASTOSOWANIU RÓŻNYCH MIAR PODOBIEŃSTWA

Streszczenie. W artykule przedstawiono porównanie czterech miar podobieństwa terminów ontologii genowych w połączeniu z dwoma miarami podobieństwa genów, przypisanych do tych terminów. Porównane zostały wizualne wyniki grupowania (takie, jak dendrogram), uzyskane za pomocą dwóch algorytmów różnego typu. Wyniki analizy pokazują, które połączenie miar podobieństwa niesie najwięcej informacji wykorzystywanej w procesie grupowania.

Słowa kluczowe: grupowanie ontologii genowych, podobieństwo terminów ontologii, podobieństwo genów, Gene Ontology

1. Introduction

Bioinformatics is an intensively developed area where a significant position is occupied by genes function analysis. Years of research conducted on genes and gene products resulted in a knowledge represented among others by Gene Ontology (GO) database. GO provides an ontology of defined terms where each term represents gene product properties in three separate domains: biological process, molecular function and cellular component. When a new function of gene or gene product is discovered it may be annotated by the Gene Ontology terms describing that function, which makes GO a valuable source of knowledge. The knowledge represented by Gene Ontology may be also applied as an input to further analysis. It can be used e.g., as an additional expert knowledge gathered throughout the years and supporting an analysis of a new dataset [11,12].

Clustering is one of the data mining methods that can be applied when gene analysis is performed. Clustering is particularly useful for analysis of gene expression values received from a microarray experiment [4, 7]. Such analysis can be supported by an expert knowledge in the form of Gene Ontology. In this case it is needed to combine the clustering of genes in two domains (expression values and Gene Ontology) [5]. Therefore, it is needed to verify how useful this expert knowledge can be, what kind of information it can provide and whether it can improve the gene analysis in expression values domain.

The goal of this work is comparison of the methods determining the similarity of genes when they are described in Gene Ontology domain.

Similarity is a basic notion utilised by clustering algorithms. There are several similarity measures that can be applied to GO clustering [8, 9, 10]. Some attempts to comparison of the clustering results when different similarity measures are applied were performed in previous years [1]. However, in the opinion of the authors additional analysis and research is needed. The work [1] presents the analysis of very small datasets and it also does not clarify all the details how the final gene similarity was calculated. Therefore, another approach to the comparison of clustering results in the context of similarity measures is justified.

The contribution of this work is comparison of the clustering results, where four different GO term similarity measures were applied which were processed further by two methods calculating gene similarity. Two different clustering algorithms were applied to the analysis and their comparison was performed on the basis of visualisation of the algorithms results. A visual approach to the comparison of clustering results enables to easily notice the differences in applied similarity measures, understand these differences and draw valuable conclusions.

The structure of this work is as follows. Section 2 presents the similarity measures that are compared in the analysis. Clustering algorithms enabling visual comparison of their results are presented in section 3. Section 4 presents the datasets, the experiments and their results, and a discussion of the obtained outcome. Conclusions of the work are presented in section 5.

2. Similarity measures

This work is focused on Gene Ontology analysis. However, we presented in the previous section how the clustering of gene expression data can be improved by combination of gene expression and Gene Ontology data analysis. Therefore, clustering of genes in gene expression domain was used in a present work as a reference result. The similarity of genes described by gene expression values is calculated in most cases by application of Pearson correlation coefficient.

The similarity of genes described by Gene Ontology terms can be calculated in several ways. In this work the similarity measures consisting of two steps are taken into consideration. During the first step the similarity of terms creating Gene Ontology is calculated. During the second step the similarity of genes, which is based on the similarity of terms annotating the genes, is calculated.

2.1. Gene Ontology term similarity

Four different GO term similarity measures were considered in the work. Three of these measures are classified as semantic similarity measures and they utilise the concept of Information Content $\tau(a)$ of an ontology term $a \in A$ (where A is a set of all GO terms) given by the following formula:

$$\tau(a) = -\ln\left(P(a)\right), \tag{1}$$

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

The simplest semantic similarity measure was proposed by Resnik [10]. It takes into consideration only the Information Content of the most informative common ancestor $\tau_{ca}(a_i,a_j)$ of the compared terms a_i and a_j:

$$s_A^{(R)}\left(a_i,a_j\right) = \tau_{ca}\left(a_i,a_j\right). \tag{2}$$

More complex approach was proposed by Jiang-Conrath [8], where term similarity is defined as:

$$s_A^{(JC)}\left(a_i,a_j\right) = \left(d_A^{(JC)}\left(a_i,a_j\right)+1\right)^{-1}, \tag{3}$$

where $d_A^{(JC)}(a_i,a_j)$ is a term distance defined as:

$$d_A^{(JC)}\left(a_i,a_j\right) = \tau\left(a_i\right)+\tau\left(a_j\right)-2\tau_{ca}\left(a_i,a_j\right). \tag{4}$$

Another semantic approach was presented by Lin [9]:

$$s_A^{(L)}\left(a_i,a_j\right)=\frac{2\tau_{ca}\left(a_i,a_j\right)}{\tau\left(a_i\right)+\tau\left(a_j\right)}. \tag{5}$$

The last similarity measure, which does not belong to the class of semantic measures, regards gene ontology as a graph, where each term is a vertex of a graph. Therefore, it is possible to define the distance between two terms a_i and a_j as a length $l(a_i,a_j)$ of the shortest path between them. Therefore, the similarity of the two GO terms can be defined as exponential dependency on $l(a_i,a_j)$ [1]:

$$s_A^{(p)}\left(a_i,a_j\right)=e^{-l\left(a_i,a_j\right)}. \tag{6}$$

2.2. Gene similarity

When the term similarity is known it is possible to calculate gene similarity based on the similarity of terms describing the genes. The similarity $s_G(g_k,g_p)$ between genes g_k and g_p can be calculated according to one of the approaches presented in literature.

The first approach [3], which will be further referred to as Avg-max, is defined as:

$$s_G\left(g_k,g_p\right)=\left(m_k+m_p\right)^{-1}\left(\sum_i\max_j\left(s_A\left(a_i,a_j\right)\right)+\sum_j\max_i\left(s_A\left(a_i,a_j\right)\right)\right), \tag{7}$$

where m_k and m_p are the number of annotations of genes g_k and g_p respectively, a_i and a_j belong to the term sets describing genes g_k and g_p respectively.

Another approach, which will be further referred to as Avg-sum, was applied in [13]:

$$s_G\left(g_k,g_p\right)=\left(m_k m_p\right)^{-1}\sum s_A\left(a_i,a_j\right), \tag{8}$$

where m_k and m_p are the number of annotations of genes g_k and g_p respectively, a_i and a_j belong to the term sets describing genes g_k and g_p respectively.

3. Clustering algorithms

The goal of the analysis is to compare how the application of different similarity measures impacts the clustering process. There was also made an assumption that visual comparison of clustering results will be performed, as an intuitive and easily interpretable approach. Therefore, two clustering methods which can produce a visual result and which are based on different principles were applied to the analysis.

OPTICS [2] algorithm does not perform partitioning into clusters in fact, but it orders the data objects enabling the insight into a dataset and selection of proper parameter values for further density based clustering. The result of OPTICS algorithm can be further utilised by

DBSCAN algorithm. The ordering performed by OPTICS reflects the density properties of a dataset analysed. Each data object is ordered with respect to density reachability distance to its predecessor. If this distance is small for several consecutive data objects then there is a cluster of densely placed data. Such situation is visualised on a plot presenting data density reachability distance as a valley. The valleys representing clusters are separated by the hills representing the data objects distant in terms of density reachability from other data and not belonging to any cluster.

The plot produced on the basis of OPTICS results gives a general overview of data that are analysed and shows whether there are any dense clusters that can be recognised in data.

The hierarchical agglomerative algorithm [6] is another method that can produce interesting results in a context of visual results representation. The method starts treating each data object that is analysed as a separate group. Next, the two closest groups are merged iteratively until one group containing all the data objects is created.

There are several approaches how to calculate the distance d_C between the two groups. one of the possibilities is average link method which defines this distance as:

$$d_C\left(C_i,C_j\right)=\frac{1}{\left|C_i\right|\left|C_j\right|}\sum_{x_k\in C_i,x_m\in C_j} d\left(x_k,x_m\right), \tag{9}$$

where C_i and C_j are the compared groups, x_k and x_m are the data objects belonging to these groups respectively, $|\ C_i\ |$ is a cardinality of a group C_i, d is a distance between the data objects.

The result of the method is hierarchy of partitions which can be visualised in a form of dendrogram – a tree structure showing how the groups were merged in the consecutive iterations. A dendrogram can show if the clustering results are balanced in terms of groups cardinality and if we can obtain several well separated groups.

4. Analysis

The following methods were applied to the analysis. Gene similarity was calculated by the methods presented in previous sections – four term similarity measures (2, 3, 5, 6) and two methods of gene similarity calculation (7, 8).

Two clustering algorithms: OPTICS and hierarchical average link algorithm were applied to the data analysis. Both OPTICS and hierarchical algorithm perform analysis on a distance matrix. Therefore, it was needed to calculate distances on the basis of similarities defined by the formulas presented in previous sections. It was assumed that the distance value would belong to the range [0,1].

In case of similarity calculated in gene expression domain and which is a correlation coefficient, the distance d was calculated as:

$$d = \frac{(1-s)}{2}, \tag{10}$$

where s is a similarity value.

In case of similarity based on Resnik term similarity, which can reach the values from a range [0,+∞], the distance d was calculated as:

$$d = (s+1)^{-1}. \tag{11}$$

In case of the other similarity measures distance values were calculated by the following formula:

$$d = 1 - s. \tag{12}$$

4.1. Datasets

Two datasets of different characteristics were used in the experiments performed.

Yeast dataset [4] that consists of 274 genes, 79 expression attributes and 645 GO terms. This dataset contains genes expression profiles from budding yeast *S. cerevisiae* that were measured during several different DNA microarray experiments. For analysis described in this paper we selected only 274 genes that composed 10 well defined functional groups described by the authors of the paper.

Human dataset [7] that consists of 296 genes, 18 expression attributes and 1711 GO terms. This dataset contains expression values of human fibroblasts in response to serum. Similar as in the previous case, we selected only genes from functional groups described by the authors of the paper. However we would like to stress that genes composing these groups were not as functionally uniform as groups described in the case of Yeast dataset.

To annotate genes we used GO terms from Biological Process ontology only. In all cases we included into analysis only genes that were described by at least one GO term.

4.2. Experiments and results

The experiments were implemented and performed in Matlab computing environment. System functions were used in order to calculate shortest paths in ontology graph and perform hierarchical clustering. Other methods were implemented by the authors.

In case of OPTICS method the parameters were set to $\varepsilon=1$ and $m=5$. In case of hierarchical clustering an average approach (9) was used in order to calculate distance between the clusters. Hierarchical method produces a dendrogram which is visualised by Matlab in an aggregated form, so that only 30 data objects are presented on a plot. This approach enables to retain a visualisation readable even if there is a large number of data objects.

Each combination of a data set, term similarity measure, gene similarity measure and clustering algorithm produced a plot which was analysed further. The results did not depend on a dataset, therefore, there are only the results on Human dataset presented in this work (due to the space restrictions). The results for Jiang-Conrath, Lin and shortest path based term similarity measures poses the same characteristic, therefore, (again, due to the space restrictions) only the results for Jiang-Conrath and shortest path based measures are presented in the figures.

This work is intended to reveal which combination of similarity measures enables a clustering process to reveal well separated and balanced groups of genes. The first feature that can be noticed on the plots presented is that analysis of gene expression values (fig. 1A and fig. 1B) produces much more informative results comparing to Gene Ontology clustering.

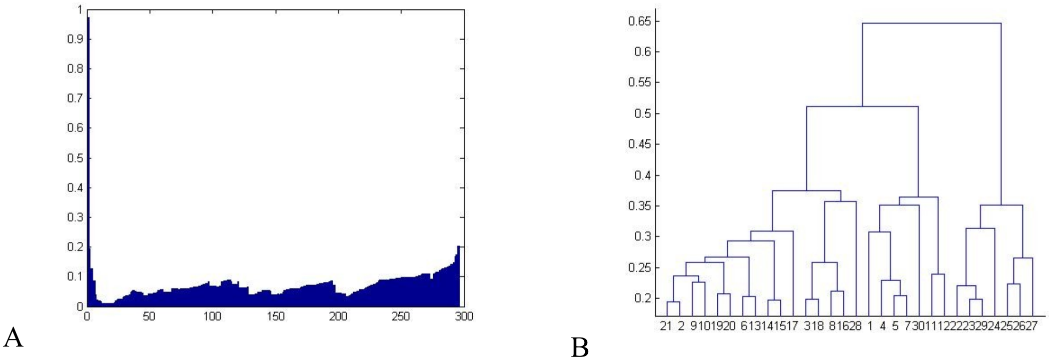

Fig. 1. The results of clustering gene expression data by OPTICS (A) and hierarchical (B) algorithms

Rys. 1. Wynik grupowania genów w dziedzinie ekspresji za pomocą algorytmów OPTICS (A) oraz hierarchicznego (B)

Analysing the results of OPTICS algorithm there is a large number of clearly visible valleys in case of gene expression data (fig. 1A) , which points that there are dens clusters of genes in the datasets. When Avg-sum method was applied together with Jiang-Conrath, Lin and shortest path based similarity measures (fig. 2A, C) the plots are hardly hilly, which is a result of almost no dense data concentrations and very poor discrimination among data. However, the results seem to be much better when Avg-max method was applied together with these measures (fig. 2B, D). Resnik measure (although a semantic one, like Jiang-Conrath and Lin measures) gives opposite results. The dense regions in data are better visible in case of Resnik and Avg-sum measures combination (fig. 2E). It means that the combination of Resnik and Avg-max measures is less discriminative.

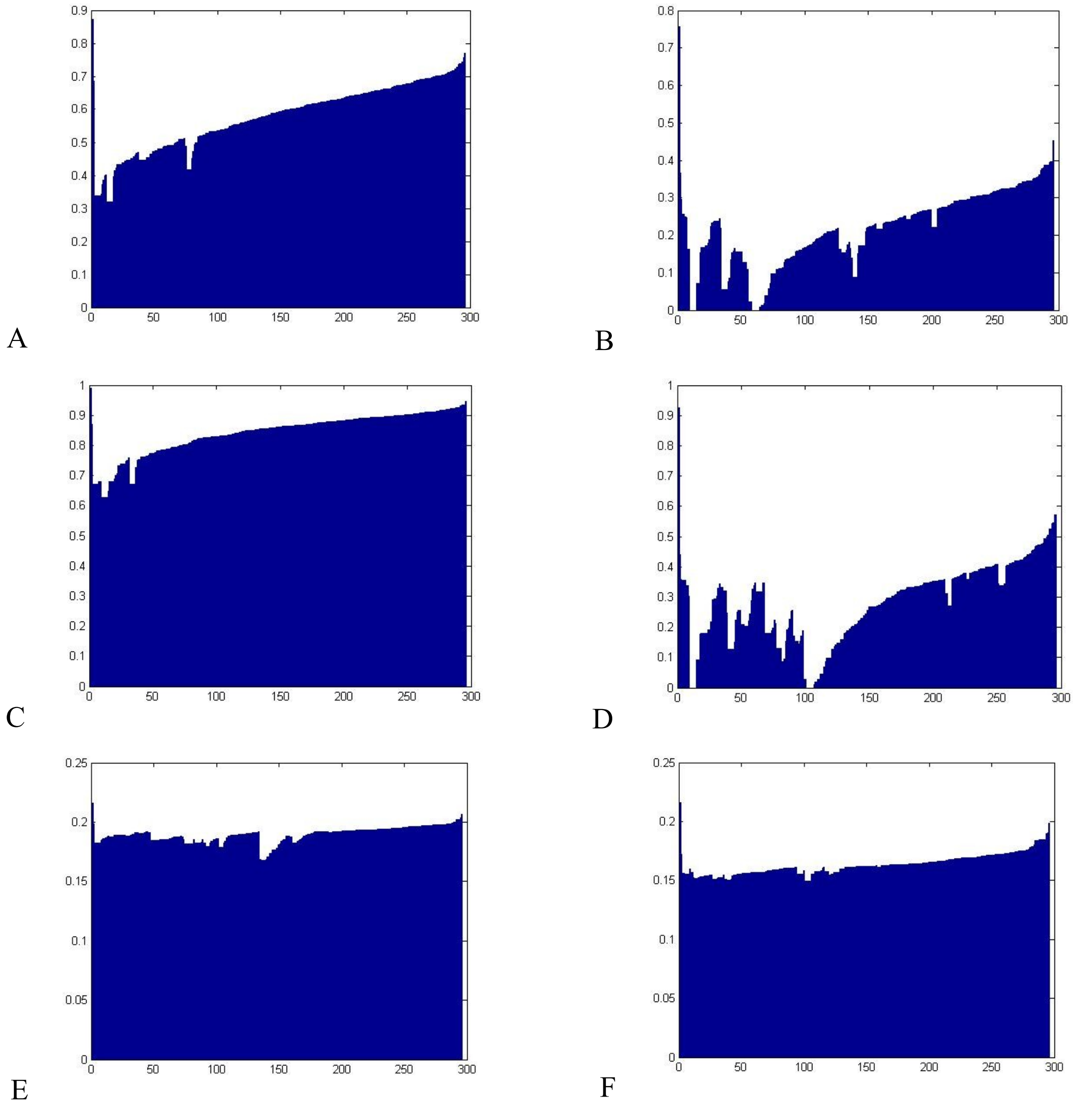

Fig. 2. The results of clustering by OPTICS method when:

- Jiang-Conrath term similarity is combined with Avg-sum (A) and Avg-max (B) gene similarity methods,
- shortest path based term similarity is combined with Avg-sum (C) and Avg-max (D) gene similarity methods,
- Resnik term similarity is combined with Avg-sum (E) and Avg-max (F) gene similarity methods

Rys. 2. Wynik algorytmu OPTICS przy zastosowaniu:

- miary podobieństwa terminów Jiang-Conrath oraz metody wyznaczania podobieństwa genów Avg-sum (A) i Avg-max (B),
- miary podobieństwa terminów bazującej na ścieżkach oraz metody wyznaczania podobieństwa genów Avg-sum (C) i Avg-max (D),
- miary podobieństwa terminów Resnik oraz metody wyznaczania podobieństwa genów Avg-sum (E) i Avg-max (F)

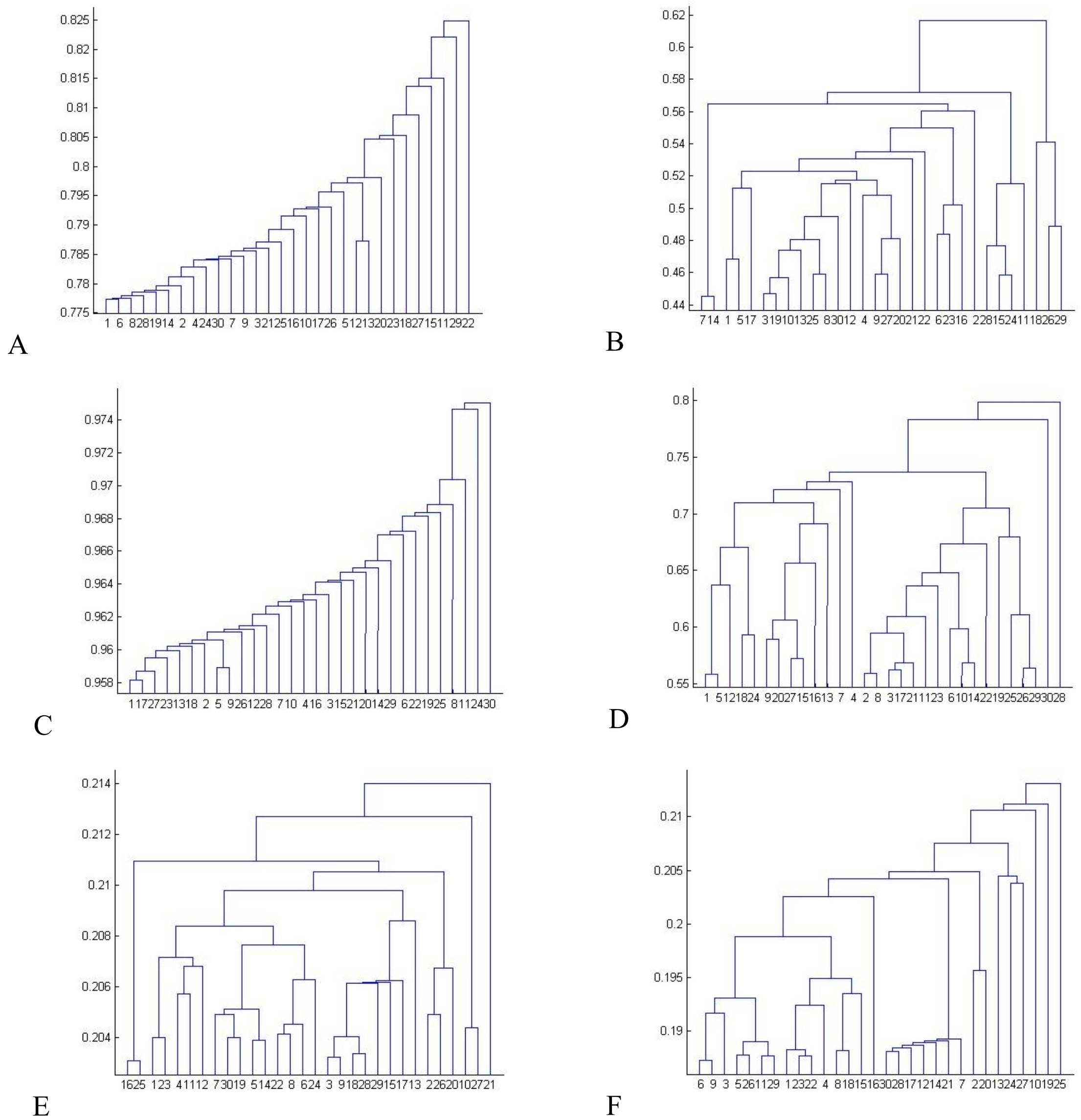

Fig. 3. The results of clustering by hierarchical method when:

- Jiang-Conrath term similarity is combined with Avg-sum (A) and Avg-max (B) gene similarity methods,
- shortest path based term similarity is combined with Avg-sum (C) and Avg-max (D) gene similarity methods,
- Resnik term similarity is combined with Avg-sum (E) and Avg-max (F) gene similarity methods

Rys. 3. Wynik algorytmu hierarchicznego przy zastosowaniu:

- miary podobieństwa terminów Jiang-Conrath oraz metody wyznaczania podobieństwa genów Avg-sum (A) i Avg-max (B),
- miary podobieństwa terminów bazującej na ścieżkach oraz metody wyznaczania podobieństwa genów Avg-sum (C) i Avg-max (D),
- miary podobieństwa terminów Resnik oraz metody wyznaczania podobieństwa genów Avg-sum (E) i Avg-max (F)

The analysis of the results produced by hierarchical clustering algorithm leads to the analogical observations as in case of OPTICS method. The dendrogram based on the analysis of gene expression values (fig. 1B) seems to present the best shape. It is possible to point at this plot several well separated groups consisting of approximately similar number of genes. None of the other plots presents as well separated groups. The results received for Jiang-Conrathm, Lin and shortest path based similarity measures are strongly unbalanced when Avg-sum method was applied (fig. 2A, C). The characteristic of the plot is improved when these term similarity measures are combined with Avg-max method (fig. 3B, D). Again, Resnik measure enables to receive better clustering results when it is combined with Avg-sum method (fig. 3E). The combination of Resnik and Avg-max measures is more unbalanced (fig. 3F).

5. Conclusions

This work presents the analysis of the methods calculating gene similarity. These methods consist of two steps: GO term similarity calculation and gene similarity calculation. Several different methods were compared in the analysis and two datasets of different characteristics were clustered.

The results, that were presented, show that not a single method (either term or gene similarity measure) but a combination of these two methods has to be taken into consideration. The plots produced by density based OPTICS method and hierarchical clustering algorithm show that the most discriminative and balanced results were delivered by a combination of Jiang-Conrath, Lin and shortest path based measures with Avg-max method, whereas in case of Resnik measure its combination with Avg-sum method.

The results show also that more informative for clustering algorithms (more discriminative) are genes described in gene expression domain.

The proposed approach to visual analysis of clustering results proved to present interesting characteristics of the methods analysed and enabled valuable conclusions concerning these methods. However, the analysis of numerical clustering quality indices, e.g. Dunn’s index, should provide additional valuable information and will be taken into consideration in future works.

Acknowledgements

This paper was partially supported by the European Community through the European Social Fund.

BIBLIOGRAPHY

1. Al Mubaid H., Nagar A.: Comparison of four similarity measures based on GO annotations for Gene Clustering, IEEE Symposium on Computers and Communications, ISCC 2008, 2008, p. 531÷536.
2. Ankerst M., Breunig M., Kriegel H. P., Sander J.: OPTICS: ordering points to identify the clustering structure, SIGMOD Rec., 1999, Vol. 28, No 2, p. 49÷60.
3. Azuaje F., Wang H., Bodenreider O.: Ontology-driven similarity approaches to supporting gene functional assessment In Proc. Of The Eighth Annual Bio-Ontologies Meeting, 2005.
4. Eisen M. B., Spellman P. T., Brown P. O., Botstein D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, Vol. 95, 1998, p. 14863÷14868.
5. Gruca A., Kozielski M., Sikora M.: Fuzzy Clustering and Gene Ontology Based Decision Rules for Identification and Description of Gene Groups, AISC, Vol. 59, 2009, p. 141÷149.
6. Han J., Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Academic Press, San Francisco 2001.
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, Vol. 283, 1999, p. 83÷87.
8. Jiang J. J., Conrath D. W.: Semantic similarity based on corpus statistics and lexical ontology In Proc. on Int. Conference on Research in Computational Linguistics, 1997, p. 19÷33.
9. Lin D.: An information-theoretic definition of similarity In Proc. of the 15th Int'l Conference on Machine Learning, 1998, p. 296÷304.
10. Resnik P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language J. Artif. Intell. Res. (JAIR), Vol. 11, 1999, p. 95÷130.
11. Sikora M., Gruca A.: Induction and selection of the most interesting Gene Ontology based multiattribute rules for descriptions of gene groups. Pattern Recogn. Letters, Vol. 32, 2011, p. 258÷269.
12. Sikora M., Gruca A.: Quality improvement of rules based gene groups descriptions using information about GO terms importance occurring in premises of determined rules. Int. Journal of Applied Mathematics & Computer Science. Vol. 20, No.3, 2010, p. 555÷570.
13. Wang H., Azuaje F., Bodenreider O., Dopazo J.: Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In Proc. of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology CIBCB'04, 2004, p. 25÷31.

Recenzenci: Dr hab. inż. Krzysztof Goczyła, prof. Pol. Gdańskiej
Dr inż. Sławomir Niedbała

Wpłynęło do Redakcji 31 stycznia 2011 r.

Omówienie

Celem niniejszego artykułu jest porównanie miar podobieństwa genów w dziedzinie ontologii genowych, które są reprezentowane przez bazę Gene Ontology, pod względem jakości wyników grupowania, uzyskiwanych przy zastosowaniu analizowanych miar. Porównanie polega na wizualnej analizie wykresów, otrzymanych w procesie grupowania genów.

W artykule przedstawiono porównanie czterech miar podobieństwa terminów ontologii genowych w połączeniu z dwoma metodami wyznaczania podobieństwa genów, na podstawie podobieństwa terminów, do których geny są przypisane. Zastosowane miary podobieństwa terminów obejmują miary semantyczne, takie jak: Jiang-Conrath [3], Lin [5], Resnik [2] oraz miara bazująca na odległości najkrótszej ścieżki [6]. Analizowane metody wyznaczania podobieństwa genów obejmują: wartość średniego sumarycznego podobieństwa (Avg-sum) [8] oraz wartość średniego maksymalnego podobieństwa (Avg-max) [7] terminów opisujących porównywane geny. W analizie zastosowano dwa algorytmy grupowania: OPTICS oraz hierarchiczny algorytm aglomeracyjny, które pozwalają na wizualizację uzyskanego podziału.

Wyniki analizy wskazują, że bardziej wyraziste dla grupowania wyniki podobieństwa daje połączenie miar Jiang-Conrath, Lin oraz bazującej na ścieżkach z metodą Avg-sum, a także połączenie miary Resnik z metodą Avg-max. Porównanie pokazuje również odmienny i mniej wyrazisty charakter grup istniejących w dziedzinie ontologii genowych w porównaniu do opisu genów za pomocą wartości ekspresji.

Addresses

Michał KOZIELSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, michal.kozielski@polsl.pl.
Aleksandra GRUCA: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, aleksandra.gruca@polsl.pl.