Scientific Journal of Riga Technical University
Computer Science. Information Technology and Management Science
2010
Volume 44

Problems of Fuzzy Clustering of Microarray Data

Oleg Uzhga-Rebrov, *Rezekne Higher Education Institution*, Galina Kuleshova, *Riga Technical University*

***Abstract* – Microarray technology has been the leading research direction in medicine, pharmacology, genome studies and other related areas over the past years. This technology enables researches to simultaneously study activity expression of tens of thousands of genes. After the experimental data have been processed, arrays of numerical values of gene expressions are obtained that are the basis for receiving relevant information and new knowledge. This paper briefly overviews the basics of microarray technology as well as task classes that could be solved using microarray data. The existing approaches to clustering gene expression sets are discussed. It is shown that the fuzzy c-means clustering method appears the most appropriate for that purpose. Due to that, the problem of choosing an optimal size of fuzziness parameter arises. Three widespread techniques for solving the problem are considered and their comparative analysis is provided.**

***Keywords* – microarray experiments, microarray data, hybridization, gene expression set, fuzzy c-means clustering, fuzziness parameter**

I. Introduction to Microarray Technology

Microarray technology enables researches to obtain expressions (activity representations) of a huge number of cell-specific or tissue-specific genes. A microarray is formed by the basis (a glass plate or a nylon membrane) on which DNA molecules are fixed in an ordered way in certain places, making spots. Each spot is an expression of the corresponding gene. To enable the quantification of expression strength (activity level), all microarray genes are labelled either with fluorescent dyes or with radioactive emanation. For the formation of visual gene expressions, the process of hybridization is particularly important. Each of two microarray samples is marked with a particular colour – red or green. The summation of these paints produces the representation of activity level of the gene under consideration. Then an electronic scanning of the resulting colours of each gene is obtained. As a result, a set of individually coloured spots is obtained. These degrees of colour intensity are then converted into the digital form. The preliminary processing of experimental data is then conducted which includes the normalization of the estimates obtained aimed to eliminate systematic errors. Besides, the evaluation of random errors can be performed, if necessary, which allows researchers to correctly evaluate the quality of the results achieved. Nowadays, multiple publications on the microarray technology and preliminary processing of results can be found. As an example, [4, 5] can be mentioned.

A study of a series of hybridizations (not of a single microarray) performed in sequence over time or by a set of relevant conditions is of great interest. The results of the preliminary processing of such multiple microarrays are represented as a matrix in which every row corresponds to a gene and every column corresponds to a separate experiment. The number in the cell that is the intersection of a certain row and a certain column represents the expression level of the concrete gene in the concrete experiment. This matrix is the initial information that serves as a basis for solving diverse research and applied tasks. In [4], the following classification of this kind of tasks is provided:

1. Finding differences in expression levels among the preliminary determined groups of instances – "the comparison of classes",
2. Identification of instance membership in a class based on the set of gene expressions – "class prediction",
3. Analysis of the given set of gene expressions aimed at recognizing subgroups that possess some common features – "class recognition".

II. Clustering Gene Expression Data Sets

Strictly speaking, the clustering of any set of gene expressions serves for solving the aforementioned task of class discovery, which might also be an independent task. However, in many cases, clustering results can be used for solving wider research and applied tasks, for example, in constructing classifiers aimed at assigning a sample of gene expressions to one of the classes defined a priori. An example of this kind of tasks is patient disease diagnostics based on the analysis of his gene expression sets. Analysing a time series of gene expression patterns, researches can make justified conclusions about the treatment effectiveness for the patient. The solving of such tasks seems to be a very prospective area, though the studies in the area have started only recently. Anyway, clustering of gene expression patterns as the first and basic procedure of their research is of paramount importance. As emphasized in [3], one of the major goals of cluster analysis of gene expression data is to identify functions of new genes by grouping them with the genes having well defined functions. This is based on the well-checkable assumption that genes that show similar activity factors (i.e., genes with similar expression levels) are frequently correlated with each other functionally and their behaviour is conditioned by similar mechanisms. Gene clusters created by clustering are often associated with certain functions. This kind of correlation enables the identification of functions of new genes provided that these genes belong to gene clusters with already identified functions. Such a methodology provides a powerful tool for making scientific and applied researches.

The existing clustering methods can be divided into two large groups: non-fuzzy (hard) clustering and fuzzy clustering. For clustering microarray data, these hard clustering methods are used: the k-means method, SOM and hierarchical

clustering. The distinctive feature of these methods is that as a result of clustering a gene will be assigned to exactly one cluster. For specific tasks of gene set clustering, a condition like that turns to be quite restrictive and frequently provides unsatisfactory results. The complicated gene structure is regulated by a set of diverse mechanisms; as a result, certain genes might belong with reason to different clusters. The fuzzy c-means clustering enables researchers to more flexibly model the complicated system of gene interaction.

Another important advantage of the c-means clustering method is its robustness to noise. The noise in microarrays might be caused by different reasons, the main reason being random factors of experiments. The property of robustness enables the c-means method to successfully operate with noisy data.

However, in applying the fuzzy c-means method to clustering microarray data two essential problems occur: the prior determination of the number of clusters and determination of the proper value of fuzziness parameter in each specific task. The following section overviews and analyses some widely used techniques for the determination of the optimal value of fuzziness parameter *m*.

III. TECHNIQUES OF FUZZINESS PARAMETER DETERMINATION

The fuzzy clustering method is based on the concept of fuzzy partitioning of data space. The fuzzy partitioning of the space can be expressed as follows:

$$M_{fc} = \left\{ U_{ij} \in R^{c\times N} \middle| \mu_{ij} \in [0,1], \forall j; \sum_{i=1}^{c} \mu_{ij} = 1, \forall j; 0 < \sum_{j=1}^{N} \mu_{ij} < N, \forall i \right\}, \quad (1)$$

where c – number of clusters;
N – number of objects;
μ_{ij} – value of the function of the j-th object membership in the i-th cluster.

The essence of the fuzzy c-means clustering method [1] for gene expression sets is the minimization of this functional:

$$J_m(G,U,P) = \sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^m \left|\mathbf{g}_j - \mathbf{p}_i\right|_{\mathbf{A}}, \quad (2)$$

where $\mathbf{g}_j$ – vector of expressions of gene j;
$\mathbf{p}_i$ – centre of cluster i (prototype, or centroid of cluster i);
m, m >1, - fuzziness parameter;
$|\mathbf{x}|_{\mathbf{A}}$ – norm distance.

For the k-th cluster and 1st gene, the parameter u_{kl} is calculated as:

$$u_{kl} = \frac{1}{\left(\frac{|g_l - p_k|_A}{\sum_{i=1}^{c} |g_l - p_k|_A} \right)^{\frac{2}{m-1}}}, \forall k, l. \quad (3)$$

The centroid of the k-th cluster is determined using expression (4):

$$p_k = \frac{\sum_{j=1}^{N} \left(\mu_{kj}\right)^m g_j}{\sum_{j=1}^{N} \left(\mu_{kj}\right)^m}, \forall k. \quad (4)$$

At the first step of the algorithm's execution arbitrary values of data points are used as centroids. After that, iterations aimed at calculating new centroids and membership function values for all genes are carried out on the basis of functional minimization (2).

A priori assigned value of fuzziness parameter *m* plays an important role in the c-means algorithm. Many researchers suggest using the value m = 2. For a large number of tasks this value proves quite satisfactory. However, numerous studies have shown that in case of fuzzy clustering of microarray data this value frequently leads to unsatisfactory results. Due to that, a task of the prior determination of the optimal value of parameter *m* in a particular task of fuzzy classification of the given gene set arises. Let us consider three approaches to solving it.

The general idea behind the technique presented in [2] is as follows. The authors have suggested a hypothesis that a correlation between the membership values and variation coefficient (cv) of the set of distances between genes in the initial data set exists:

$$Y_m = \left\{ \left[d^2\left(x_i, x_k\right) \right]^{\frac{1}{m-1}}, k \neq i, i = 1,2,...,N \right\}. \quad (5)$$

Based on the experiments conducted, the following empirical correlation was deduced:

$$cv(Y_m) = \frac{\sigma_{Y_m}}{\overline{Y}_m} \approx 0.03p, \quad (6)$$

where Y_m is determined using expression (5),
σ_{Ym} - standard deviation of set Y_m,
$\overline{Y}_m$ - mean value of set Y_m,
p – dimension of the initial data set.

In this case we treat the dimension of the data as the number of experiments (conditions), i.e., as the number of columns of the matrix of initial data.

At the first step, *m* is assumed to be equal to 2, and the left-hand side of equation (6) is calculated. If this yields the inequality of type „<", the value m=2 is accepted as the working value of fuzziness parameter. In case if the inequality of type ≥ is obtained, the dichotomic partition of the value interval m [1.0; 2.0] is carried out, i.e., an assumption is made that m=1.5 and the left-hand side of equation (6) is calculated. Depending on the result obtained, further searching is conducted either in the interval [1.0; 1.5] or in the interval [1.5; 2.0]. The process is repeated until the value of *m* is found

that meets equation (6). This value is assumed as the working value for performing clustering. The authors show with examples that their method produces good results.

The key idea of the technique proposed in [3] is as follows. If the clustering of the initial set of objects (genes) is performed at a certain specified value of fuzziness parameter *m*, then the quality of clustering could be evaluated using the appropriate functional f: U $\rightarrow$ R. The paper examines three functionals of that kind.

1. Partition coefficient F:

$$F(U) = \sum_{k,i=1}^{c,N} \frac{\mu_{ik}^2}{N}. \qquad (7)$$

This coefficient assumes the largest value for a hard partition and reaches maximum for U = 1/c, when every object is equally assigned to every class.

2. The normalised partition coefficient $\tilde{F}$:

$$\tilde{F} = F - F_0, \qquad (8)$$

where F_0 – partition coefficient obtained from the randomised set of data.

3. S (Xie – Beni) index:

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2 \left\| g_j - p_i \right\|^2}{N \min_{k\neq l}(\left\| p_k - p_l \right\|^2)}. \qquad (9)$$

This index quantifies the ratio between the total variation and cluster partitioning. The minimum index value gives evidence of a high quality of clustering.

By varying the value *m*, the clustering of the initial set is performed. For each clustering variant, the value of the selected functional is computed. As a working value, the value of *m* that minimizes the functional is accepted.

The shortcoming of this technique is a large number of calculations required. A certain advantage of the method is that it is not necessary to perform clustering at the optimal value of *m*, since it has already been performed.

In [6], it is stated that the strong correlation between the fuzziness parameter and basic properties of the data set can be demonstrated using a simplified model of the system. This kind of system consists of binary objects that are characterised by two values of evaluation parameter - {-1, 1}. It is clear that the probability to have an object x_i = {1, 1, …, 1} is 2^{-D}, where D – the dimension of the data set (see the definition above). The probability to have half of all objects of the data set with that vector of values of the evaluation parameter is

$$\binom{N}{\frac{N}{2}} 2^{-D\frac{N}{2}} \left(1-2^{-D}\right)^{\frac{N}{2}} \approx \sqrt{\frac{2}{N\pi}}\, 2^{N\left(1-\frac{D}{2}\right)} \left(1-2^{-D}\right)^{\frac{N}{2}}. \qquad (10)$$

For $2^{-D} << 1$, the right-hand side of equation (10) can be approximated by this expression:

$$2^{N\left(1-\frac{D}{2}\right)} \sqrt{\frac{2}{\pi N}}.$$

Therefore, the probability for a well defined cluster decreases exponentially with regard to the dimension of the data set, and slightly slower for an increasing number of objects in the set. As a consequence, the values of parameter *m* which is the measure of fuzziness of the system will follow this tendency at least qualitatively.

A thorough analysis and calculation of m_t (a threshold value for m) for the randomised data sets at different dimensions and numbers of objects shows a general functional correlation between m_t and properties of the data set:

$$f(D,N) = 1 + \left(\frac{1418}{N} + 22.05\right) D^{-2} + \\ + \left(\frac{12.33}{N} + 0.243\right) D^{-0.0406\ln(N)-0.1134}. \qquad (11)$$

To compute the threshold value m_t, it suffices to calculate the right-hand side of expression (11). Usually, better results are observed for larger N and D, whereas for data sets with small N and D, deviations of m_t from the optimal value *m* might occur.

IV. Conclusions

If we accept the number of calculations required as the criterion for choosing an optimal value of *m* in a specific task, the most suitable is the approach described in [6]. The least preferable with regard to that criterion is the technique suggested in [3]. Actually, this technique does not allow us to a priori determine an optimal value of parameter *m*; this can only be done using the results of clustering quality evaluation performed at different values of parameter *m*. The technique proposed in [2] also requires a considerable number of calculations, but they are of explicit statistical nature and do not need any special software.

When the validity of the results obtained is the criterion for choosing the most suitable technique, the preference has to be given to the method discussed in [2]. This technique appears to be optimal in the sense of the number of calculations.

The suitability of empirical dependence (11) in [6] for any possible task of gene set clustering seems doubtful. Here additional research is necessary. The confidence of different indexes of clustering quality in [3] does not give rise to doubt, but this approach is absolutely unsuitable from the computational point of view.

To summarise the results of the analysis conducted, the technique proposed in [2] has to be considered the most preferable. This technique combines simplicity and visual interpretability with a reasonable amount of calculations.

ACKNOWLEDGEMENTS

This work has been partly supported by Latvia-Belarus Co-operation program in Science and Engineering within the project 'Development of a complex of intelligent methods and medical and biological data processing algorithms for oncology disease diagnostics improvement'.

REFERENCES

[1] **Bezdek, J. C.** *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[2] **Dembèlè, D., Kastner, P.** "Fuzzy c-means method for clustering microarray data," Bioinformatics, Vol. 19, No.8, pp.973-980, May 2003.

[3] **Futschik, M.E., Kasabov, K..** "Fuzzy Clustering of Gene Expression data,"Available: http://itb.biologie.hu-berlin.de/~futschik/publis/futschik_ieee.pdf. [Accessed August 7, 2010].

[4] **Tarca, A. L., Romero, R., Draghici, S.** "Analysis of microarray experiments of gene expression profiling," American Journal of Obstetrics and Gynecology, Vol.195, pp.373-378, 2006.

[5] **Tjaden, B., Cohen, J.** "A Survey of Computational Methods Used in microarray Data Interpretation," in *Applied Mycology and Biotechnology*, Vol. 6, Bioinformatics, D. K. Arora, R. M. Berka and G. B. Sigh, Eds. Elsevier Science, 2006, pp.161-178.

[6] **Schwämmle, V., Jensen, O. N.** "A simple and fast method to determine the parameters for fuzzy c-means cluster validation," Available: http://arxiv.org/abs/1004.1307v1. [Accessed: July 27, 2010].

Oleg Uzhga-Rebrov is Professor in the Faculty of Economics in Rezekne Higher Education Institution (Latvia). He received his doctor's degree in Information Systems from Riga Technical University in 1994. His research interests include different approaches to processing incomplete, uncertain and fuzzy information, in particular, fuzzy sets theory, rough set theory as well as fuzzy classification and fuzzy clustering techniques and their applications in bioinformatics.
Contact information: Rezekne Higher Education Institution, 90 Atbrivoshanas aleja, Rezekne LV-4600, Latvia. E-mail: rebrovs@tvnet.lv.

Galina Kuleshova is Research Scientist in the Faculty of Computer Science and Information Technology at Riga Technical University (Latvia). She received her M.Sc. degree in *Decision Support Systems* from Riga Technical University in 1996. Current research interests include artificial neural networks, classification methods and bioinformatics.
Contact information: Institute of Information Technology, Riga Technical University, 1 Kalku Str., Riga LV-1658, Latvia. E-mail: galina.kulesova@cs.rtu.lv.

Oļegs Užga-Rebrovs, Gaļina Kuļešova. Mikromasīvu datu izplūdušās klasterizācijas problēmas
Gēnu mikromasīvu datu tehnoloģijas attīstība deva spēcīgu impulsu zinātniskiem un praktiskiem pētījumiem medicīnā, ģenētikā, farmakoloģijā un citās nozarēs. Šī tehnoloģija ļauj vienlaicīgi iegūt tūkstošus un desmitiem tūkstošus gēnu izpausmes. Tik milzīgs sākotnējās informācijas daudzums prasa jaunu speciālo metožu izstrādi šīs informācijas apstrādei un analīzei. Iepriekšējās datu apstrādes rezultāti tiek atspoguļoti matrices formā, kuras rindas atbilst gēniem un kolonnas atbilst atsevišķiem eksperimentiem. Eksperimenti var būt saistīti ar gēnu aktivitātes atklāšanu dažādos laika punktos, gēnu izpausmju salīdzināšanu veselīgos un slimīgos organismos, gēnu izpausmju izmaiņām terapeitiskas iejaukšanās rezultātā un ar daudzām citām problēmām. Skaitlis matrices ailē atspoguļo dotā gēna aktivitātes pakāpi konkrētajā eksperimentā.
Liela nozīme ir izdalītās gēnu kopas klasterizācijai. Pašlaik ir vispāratzīts, ka tādas klasterizācijas rezultāti var tikt veiksmīgi izmantoti gēnu mijiedarbības atklāšanai un dažādu iekššūnu procesu izpratnei. Tiek izstrādāts liels algoritmu daudzums gēnu izpausmju datu klasterizācijai. Plaši tiek pielietots izplūdušās klasterizācijas c-vidējo algoritms. Šim algoritmam ir daudzas priekšrocības salīdzinājumā ar precīziem algoritmiem, piemēram, ar k-vidējo algoritmu. Tomēr praktiskā šīs metodes pielietošanā mikromasīvu datu klasterizācijai rodas izplūšanas parametra m optimālā lieluma izvēles problēma. Plaši izmantojamais standarta lielums m = 2 gēnu izpausmju klasterizācijas uzdevumu kontekstā neļauj iegūt korektus rezultātus. Šajā darbā tiek veikta šīs problēmas risināšanas trīs pieeju salīdzinošā analīze. Analīzes izpildes rezultātā tiek piedāvāta konkrēta rekomendācija izmantot vienu no apskatītajām pieejām.

Олег Ужга-Ребров, Галина Кулешова. Проблемы нечёткой кластеризации данных микромассивов
Развитие технологии данных микромассивов генов дало мощный импульс новым научным и прикладным исследованиям в медицине, генетике, фармакологии и других областях. Эта технология позволяет одновременно получить выражения тысяч и десятков тысяч генов. Такое огромное количество исходной информации требует разработки специальных методов для её обработки и анализа. Результаты предварительной обработки данных отображаются в форме матрицы, строки которой соответствуют генам, а столбцы – отдельным экспериментам. Эксперименты могут относиться к выявлению активности генов в различных временных точках, сравнению выражений генов для здоровых и больных организмов, изменению выражений генов в результате терапевтического вмешательства и многим другим проблемам. Число в ячейке матрицы отображает степень активности данного гена в конкретном эксперименте.
Большое значение имеет кластеризация выделенного множества генов. В настоящее время считается общепризнанным, что результаты такой кластеризации могут быть успешно использованы для выявления взаимодействий генов и понимания различных внутриклеточных процессов. Разработано большое число алгоритмов для кластеризации данных выражений генов. Среди таких алгоритмов широко используется алгоритм нечёткой кластеризации c-средних. Этот алгоритм имеет многие преимущества по сравнению с чёткими алгоритмами, например, алгоритмом k-средних. Однако, при практическом применении этого метода для кластеризации данных микромассивов возникает проблема априорного выбора оптимального значения параметра нечёткости m. Широко используемое стандартное значение m = 2 в контексте задач кластеризации выражений генов не позволяет получить корректные результаты. В настоящей работе представлен сравнительный анализ трёх подходов к решению данной проблемы. В результате выполненного анализа представлены конкретные рекомендации по использованию одного из рассмотренных подходов.