The Variable Neighborhood Search Metaheuristic for Fuzzy Clustering cDNA Microarray Gene Expression Data

Belacel, Nabil; Cuperlovic-Culf, Miroslava; Ouellette, R.

Publisher's version / Version de l'éditeur:

Proceedings of IASTED-AIA-04 Conference, 2004

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

The Variable Neighborhood Search Metaheuristic for Fuzzy Clustering cDNA Microarray Gene Expression Data *

Belacel, N., Cuperlovic-Culf, M., Ouellette, R., and Boulassel M.
February 2004

* published in Proceedings of IASTED-AIA-04 Conference. Innsbruck, Austria. February 16-18, 2004. 6 Pages. NRC 46538.

Canada

THE VARIABLE NEIGHBORHOOD SEARCH METAHEURISTIC FOR FUZZY CLUSERING cDNA MICROARRAY GENE EXPRESSION DATA

Nabil Belacel

National Research Council Canada, Institute for Information Technology-e-Health,

127 Carleton street, Saint-John, NB, E2L 2Z6 Canada

nabil.belacel@nrc.gc.ca

Miroslava Cuperlovic-Culf

and

Rodney Ouellette

Beausejour Medical Research Institute (IRMB)

37 Providence Street, Moncton NB, E1C 8X3, Canada

miroslavac@health.nb.ca

and

rodneyo@health.nb.ca

Mohamed R. Boulassel

Division of Hematology and Immunodeficiency Service, Royal Victoria Hospital, McGill University Health Centre, McGill University, Montreal, Quebec

rachid.boulass@muhc.mcgill.ca

ABSTRACT

Several thousand genes can be monitored simultaneously using cDNA microarray technology. To exploit the huge amount of information contained in gene expression data, adaptation of existing and development of new computational methods are required. Recently, the Fuzzy C-Means (F-CM) method has been applied to cluster cDNA microarray data sets. To overcome some shortcomings of F-CM and to improve its performance, it was embedded into a variable neighborhood search (VNS) metaheuristic. The methodology was used to cluster four cDNA microarray data sets. Results show that VNS+F-CM substantially improves the findings obtained by F-CM. This methodology may yield significant benefit in the improvement of decision support systems used for gene expression classification.

KEY WORDS

Fuzzy Clustering, Gene Expression, Variable Neighborhood Search Metaheuristic, Bioinformatics.

I. INTRODUCTION

With advances in complementary DNA (cDNA) microarray technology, it becomes possible for the first time to monitor the expression levels of thousands of genes simultaneously. To analyze the large amount of data obtained by this technology, researchers usually resort to clustering methods to identify groups of genes that share similar expression profiles or those characterizing various pathological conditions. Clustering methods can be divided into two main groups: hierarchical and non-hierarchical (i.e., partitional) algorithms [1,2]. The hierarchical clustering algorithms are the most common methods used for expression data analyses. They produce dendrograms, in which each branch forms a group of genes that share similar behavior [3]. These types of clustering algorithms have recently been used for cDNA microarray data to identify cancer types [4-5], to discover new subtypes of cancer [6] and to investigate cancer tumorigenesis mechanisms [7].

The non-hierarchical algorithms partition genes into *K* clusters such that genes in the same cluster are more similar to each other (i.e. homogeneity) while genes in different clusters are as dissimilar as possible (i.e. separation). Several partitional algorithms, including K-means, partitioning around medoids and self-organizing maps, have been applied to cDNA microarray data generated from different biological sources [8]. For example, K-means algorithm was used to identify molecular subtypes of brain tumors [9], to cluster transcriptional regulatory cell cycle genes in yeast [10] and to correlate changes in gene expression with major physiological events in potato biology [11].

Although both hierarchical and non-hierarchical algorithms have yielded very encouraging results in clustering cDNA microarray data, they also have several shortcomings. The major limitation of these algorithms is that they are unable to identify co-expressed genes when analyzing large gene-expression data sets collected under various conditions. This results in inaccurate or wrong clusters leading to incorrect

conclusions regarding genes that share similar behavior. In addition, hierarchical methods suffer from non-uniqueness of dendrograms and higher time as well as space complexities [12], while non-hierarchical methods group gene expression data into a fixed number of predetermined clusters. Moreover, using different data analysis techniques and different clustering algorithms to analyze the same cDNA microarray data set, different conclusions can be drawn [13,14]. The fuzzy clustering algorithms provide a systemic and unbiased way to transform precise values into qualitative descriptors in a process known as *fuzzification*. Thus, these algorithms give more information regarding the relative degree of membership of each gene to each cluster, thereby including the possibility for gene multi-functionality. The main advantage of applying fuzzy clustering algorithms to analysis of cDNA microarray data is that these algorithms inherently account for noise in the data because they extract trends not precise values [15]. More recently, many approaches based on fuzzy logic have been applied to cDNA microarray data giving additional information concerning gene functionality and co-expression [16,17]. In overall applications, the most used method is the Fuzzy C-Means algorithm (F-CM) previously described by Bezdek [18]. F-CM is an extension of the well-known K-means heuristic used for crisp clustering [19]. F-CM searches membership degrees and centroid until there is no more improvement in the objective function value. To overcome the difficulty of being stuck in local minima of poor value, F-CM was embedded into the Variable Neighborhood Search (VNS) metaheuristic [20]. Here, we investigate for the first time the applicability of F-CM embedded into metaheuristic VNS on experimental cDNA microarray data sets.

This paper is organized as follow. Section 2 presents the data set under study, an introduction to the F-CM and VNS methods and the description of the experiments. Section 3 shows the results. Section 4 presents a discussion of the results and their implications to the process of molecular classification of breast cancer and gene expression analysis. This section also discusses possible futures works to be developed.

II. MATERIAL AND METHODS

Data description

Two cDNA microarray data sets from normal human peripheral blood mononuclear cells (PBMC) and from breast cancer cells were used. These data sets were downloaded from the Stanford Microarray Database and are available on the Web site http://genome-ww5.stanford.edu/MicroArray/S

MD/index.shtml. The data set for normal PBMC consisted of 18000 genes generated from 147 experiments using 82 blood samples from healthy donors. Two data sets of 43 (43_147) and 2197 (2197_147) genes were selected for this work. The breast cancer cell data set contained 8102 genes generated from 85 tissue samples. For this purpose we also selected two data sets of 69 (69_85) and 1022 (1022_85) genes. For all data sets, only genes with complete data were kept in order to avoid uncertainties in the clustering results caused by the choice of method for the determination of missing values. The expression values are represented as mean normalized ratios of expression levels: $\log_2 \frac{R}{G}$, where R and G are respectively the expression level in Cy3 and Cy5 channels.

Algorithms

The classical clustering algorithms assign each object (gene) to one cluster only. In the fuzzy clustering method an indicator variable showing whether an object is a member of a given group/cluster is assign to a weighting factor called membership (w), determined for each object and cluster. The membership can have values in the interval [0,1] where membership values of close to 1 indicate strong association to the cluster, and values close to 0 indicate weak or lack of association to the cluster. The basic problem in fuzzy clustering of genes from data obtained by cDNA microarray experiments is to assign each gene to a given number of clusters such that each gene may belong to more than one cluster with different degrees of membership. In the following subsections, we briefly describe the F-CM and VNS methods.

Fuzzy C-Means method

F-CM is a fuzzy logic extension of the classic, crisp, K-means method [18,21-22]. The results of a microarray experiment can be presented in terms of an $n \times N$ matrix where n is the number of genes and N is the number of experiments.

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11}\, x_{12} \dots\dots\dots\, x_{1N} \\ x_{21}\, x_{22} \dots\dots\dots\, x_{2N} \\ \dots\dots\dots \\ x_{n1}\, x_{n2} \dots\dots\dots\, x_{nN} \end{bmatrix} \qquad (1)$$

Here x_{ij} represents the background subtracted, normalized, expression level or $\log_2$ of the expression level (absolute or relative depending on the type of experiment) of a gene i, in experiment j.

Then, for a chosen number of clusters, c, membership values, w, are defined in a $n \times c$ matrix $W = [w_{ik}]$, where w_{ik} is the membership degree for gene i, $i = 1,\dots, n$ in cluster k, $k = 1,\dots, c$. The F-CM clustering problem can be represented as:

$$(\min_{W,V}) R_m(W,V) = \sum_{i=1}^{n} \sum_{k=1}^{c} w_{ik}^m \left\| x_i - v_j \right\|^2 \qquad (2)$$

where:

- R_m *(W,V)* is the objectivity function defining the quality of the result obtained for centroids *V* and memberships *W*;
- *m* is the fuzziness parameter which regulates the degree of fuzziness in the clustering process; for *m*=1 the problem is the classical minimum sum of squares clustering and the partition is crisp;
- $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c] = \begin{bmatrix} v_{11}\, v_{12} \ldots\ldots v_{1c} \\ v_{21}\, v_{22} \ldots\ldots v_{2c} \\ \ldots\ldots \\ v_{N1}\, v_{N2} \ldots\ldots v_{Nc} \end{bmatrix}$

  gives a set of *c* centroids or prototypes, *i.e.* positions of cluster centres;
- $\|x_i - v_k\|^2 = \langle x_i - v_k | x_i - v_k \rangle = \sum_{j=1}^{N} |x_{ij} - v_{jk}|^2$

  is the Euclidean norm determining distances between expression level vectors and centroids;
- membership degrees w_{ik} are defined such that:

$$0 \le w_{ik} \le 1 \text{ and } \sum_{k=1}^{c} w_{ik} = 1 \ \forall i = 1,\ldots,n.$$

F-CM is a local search heuristic, *i.e.* it searches only for the clustering solution closest to the starting centroid values. Therefore, for applications of this method there is no guarantee that the final solution is the overall optimal clustering solution even when one uses several different starting points. When fuzzy methods are used on large data sets, with large number of clusters, characteristic to the microarray applications, there is a considerable possibility of obtaining only the closest solution instead of the global one, ideally or at least an improved, more distant local one. This problem is alleviated by using the VNS method.

Variable Neighborhood Search metaheuristic

The VNS algorithm is a recently proposed metaheuristic for solving combinatorial and global optimization problems [20]. The basic goal of the method is to proceed to a systematic change of neighborhood within a local search algorithm. This algorithm remains in the same locally optimal solution exploring increasingly far neighborhoods by random generation, until another solution better than the incumbent is found. When so, it jumps to the new solution and proceeds from there. The neighborhood centroid structures are obtained by replacing at random some predetermined number *k* of existing centroids of clusters with *k* randomly chosen patterns, i.e., genes. For a more detailed analysis of VNS metaheuristic, see references [20,21].

III. RESULTS

The F-CM and VNS methods were tested on four cDNA microarray data sets obtained from normal PBMC and breast cancer samples. These methods were coded and run on DELL Latitude c840, Pentium 4 Computer with CPU= 1.60 GHZ and 261.56 KB RAM. The codes were compiled using an optimizing option (C++ -04). The initial step was to determine the optimal fuzziness parameter *m* for further analysis. Then, using the obtained *m* value we compared the values of the objective function using F-CM and VNS methods.

Recently, it has been shown that it is not appropriate for the fuzziness parameter, *m*, to be equal to 2 when the F-CM method is applied to cDNA microarray data analysis [17]. To determine the best value of *m* for cDNA microarray data sets, we follow the approach described by Denbele and Kastner [17]. Using this approach, it was found that the adequate value for *m* is equal to 1.25. The maximum time (t_{max}) of 2 seconds is allowed for each run to VNS heuristic. Note, that the possibility of some further small improvement with much larger (t_{max}) can not be ruled out. Table 1 compares results of F-CM and VNS methods for different numbers of clusters using the same initial solution. The lines of the tables represent the number of clusters and the second column contains the best-known objective function values obtained by the methods. The next two columns show the percentage of deviation calculated as $\left[\left(R - R_{best}\right) \Big/ R_{best} \right] \times 100$, where *R* and R_{best} denote the solution obtained by the methods and the best known solution, respectively. The following observations can be derived from table 1. First, in most data sets tested, VNS outperforms F-CM. Second, F-CM is faster than VNS heuristic, however the solution quality of F-CM is worse than that of VNS. Third, VNS performs much better when the number of clusters is very large.

Data set	C	Best known solution	% Deviation from the best known-F-CM	% Deviation from the best known-VNS	CPU Times F-CM	CPU Times VNS
1022_85	10	41087.4	0	0	0.27	5.82
	20	36369.4	10.786	0	0.14	8.82
	30	34141.1	11.3006	0	0.22	12.31
	40	31900.3	14.1681	0	0.31	18.53
	50	30222.9	16.6278	0	0.42	71.85
	60	29208.8	17.4818	0	0.58	28.82
	70	28087.9	19.4194	0	0.68	404.26
	80	26990.2	21.8442	0	0.83	47.74
	90	25473.6	26.8643	0	0.97	2253.14
	100	25050.2	27.0069	0	0.8	333.29
Average			16.55	0	0.52	318.46
69_85	3	3429.15	0	0	0.06	0.06
	4	3003.34	1.55056	0	0.03	0.46
	5	2747.83	0	0	0.05	0.27
	6	2582.95	0	0	0.05	0.46
	7	2372.35	0	0	0.08	0.46
	8	2311.15	0	0	0.07	0.99
	9	2105.45	4.33077	0	0.25	4.33
	10	2002.2	14.5776		0.14	7.57
Average			2.55	0	0.091	1.82
2197_147	10	79764.1	0	0	47.83	420.5
	20	59047.7	0	0	72.17	1423.46
	30	50695.9	0	0	91.54	751.18
	40	46187.9	0	0	115.91	833.68
	50	45669.2	0	0	16.68	1268.03
	60	39659.9	0	0	12.14	1479.8
	70	37690.5	13.61	0	8.92	1009.88
	80	36117.5	10.43	0	19.58	539.55
	90	39094	0	0	15.01	2075.37
	100	33725.7	15.60	0	11.98	1792.86
Average			3.01	0	48.09	975.76
43_147	3	2023.35	0	0	0.02	0.13
	4	1512.17	18.38	0	0.01	0.27
	5	1307.82	0	0	0.02	0.31
	6	1212.41	4.15	0	0.02	0.59
	7	1102.54	0	0	0.07	0.51
	8	914.525	8.65	0	0.03	0.66
	9	834.128	7.86	0	0.02	1.12
	10	855.882	0.76	0	0.03	0.76
Average			4.97	0	0.027	0.54

Table 1: Comparison of F-CM and VNS+F-CM methods using the objective function and CPU time for several cluster sizes (with $m = 1.25$)

IV. DISCUSSION

In the context of cDNA microarray, the most significant advantage of fuzzy algorithms is that they allow clustering of genes, taking into account their multifunctionality (i.e., co-expression). This ability to identify multifunctional genes is a great advantage not only to better understand the mechanisms underlying the molecular basis of diseases but also to improve the accuracy of classification for discriminating between different types of tumors.

In a recent publication, Dembele and Kastner demonstrated that F-CM could be a useful tool for the dissection of various regulatory pathways involved in gene co-expression in yeast cell cycles [17]. However, it has been shown that F-CM may be stuck in local minima, possibly of poor value. This may result in inaccurate clusters leading to incorrect conclusions regarding gene multi-functionality. To overcome this shortcoming, the F-CM was embedded into the global VNS metaheuristic (VNS+F-CM). Our aim was to show the viability of VNS+F-CM in terms of clustering and improving robustness of F-CM used for cDNA microarray data analysis. We first determined the optimal fuzziness parameter *m* following the general approach reported by Dembele and Kastner [17]. We then compared the performance of VNS+F-CM against F-CM method. Although the VNS+F-CM is slower than the F-CM method, values of the objectivity factors show that VNS+F-CM gives superior accuracy in clustering cDNA microarray data sets. This improvement in classification is very important in light of the large volume of data generated by cDNA microarray technology. This may lead to highly desirable improvements in diagnosis and treatment of human diseases. In most medical applications, errors in classification could have serious consequences. For example, when using cDNA microarray data sets to cluster breast tumors, errors could correspond to misdiagnosis and assignment to improper treatment protocols. Having in hand robust clustering methods that achieve a high degree of accuracy is crucial for applications in medical fields.

To the best of our knowledge, it is the first time that the relocation heuristic VNS has been applied in the context of cDNA microarray, to solve certain problems related to fuzzy algorithms. We believe that VNS+F-CM will allow simultaneous determination of genes, which are strongly associated to one group or those correlated only under some conditions as well as those, which are always correlated. This information will help in determining cellular pathways, which are largely independent of the environment of the particular cell type, for example pathways, which are important for all breast cancer cells regardless of the type or development. Future work will be focused on: (i) the investigation and exploration of the membership degrees of VNS+F-CM (ii) the extending of the applications to other cDNA microarray data sets generated from leukemia, lymphomas brain and colon cancers.

Acknowledgments

This work is supported by the National Research Council (NRC), Institute for Information Technology-e-Health (IIT-e-Health) Saint John, Canada and Atlantic Innovation Fund.

References

1. Stanford, DC., Clarkson DB. and Hoering, A. Clustering or Automatic Class Discovery: Hierarchical Methods, in Berrar DP, Dubitzky W and Granzow M (Ed.) *A practical Approach to Microarray Data Analysis*, (Kluwer Academic Publishers, 2003) 246-260.
2. Yeung, KY. Clustering or automatic class discovery: non-hierarchical, non-SOM. Clustering Algorithms and Assessment of Clustering Results, in Berrar DP, Dubitzky W and Granzow M (Ed.), *A practical Approach to Microarray Data Analysis*, (Kluwer Academic Publishers, 2003) 274-288.
3. Eisen, MB., Spellman, PT., Brown PO and Botstein, D. Cluster analysis and display of genome-wide expression patterns, Proc. *Natl Acad Sci*, USA. 1998; 95, 14863-14868.
4. Nielsen, TO *et al*. Molecular characterisation of soft tissue tumours: a gene expression study, *Lancet*; 359, 2002, 1301-1307.
5. Ramaswamy, S. *et al.* A molecular signature of metastasis in primary solid tumors, *Nat Genet*, 33, 2003, 49-54.
6. Alizadeh, A. *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403, 2000, 503-511.
7. Welcsh, PL. *et al.* BRCA1 transcriptionally regulates genes involved in breast tumorigenesis. Proc. *Natl Acad Sci* USA, 2002, 99, 7560-7565.
8. Sherlock, G. Analysis of large-scale gene expression data. *Curr. Opin. Immunol*, 12, 2000, 201-205.
9. Shai, R. et al. Gene expression profiling identifies molecular subtypes of gliomas, *Oncogene*, 22, 2003, 4918-4923.
10. Tavazoie, S. *et al.* Systematic determination of genetic network architecture, *Nat Genet*, 22, 1999, 281-285.
11. Ronning, CM. *et al.* Comparative analyses of potato expressed sequence tag libraries, *Plant Physiol*, 131, 2003, 419-429.
12. Morgan, BJT and Ray, APG. Non-uniqueness and inversions in clusters analysis, *Apll. Statist.*, 44, 1995, 117-134.
13. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast, *Science*, 282, 1998, 699-705.

14. Raychaudhui, S, Stuart, JM. and Altman RM. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing, Hawaii*, 2000, 452-463.
15. Woolf, PJ. and Wang, Y. A fuzzy logic approach to analyzing gene expression data, *Physiol. Genomics*, 3, 2000, 9-15.
16. Gasch, AP. and Eisen, MB. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, *Genome Biol, 3 2002, RESEARCH0059.*
17. Dembele, D. and Kastner, P. Fuzzy C-means method for clustering microarray data, *Bioinformatics*, 19, 2003, 973-980.
18. Bezdek, JC. *Pattern recognition with fuzzy objective function algorithms* (New York: Plenum Press, 1981).
19. McQueen, JB. Some methods for classification and analysis of multivariate observations, Proc. *15th Berkeley Symposium on Mathematical Statistics and Probability*, Vol 2, 1967, 281-297.
20. Mladenovic, N. and Hansen, P. Variable Neighborhood Search: principals and Applications. *Eur. J. Oper. Res.*, 130, 1999, 449-467.
21. Belacel, N., Mladenovic, N. and Hansen, P. Fuzzy J-Means: a new heuristic for fuzzy clustering, *Pattern Recognition*, 35, 2002, 2193-2200.
22. Dunn, JC. A fuzzy relative ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.*, 3, 1974, 32-57.
23. Ruspini, EH. A new Approach to clustering. *Inf. Control*, 15, 1969, 22-32.