An integrated tool for microarray data clustering and cluster validity assessment

Nadia Bolshakova
Department of Computer Science
Trinity College Dublin
Ireland
+353 1 608 3688
Nadia.Bolshakova@cs.tcd.ie

Francisco Azuaje
School of Computing and Mathematics
University of Ulster at Jordanstown
Northern Ireland, UK
+44 28 90368391
fj.azuaje@ulster.ac.uk

Pádraig Cunningham
Department of Computer Science
Trinity College Dublin
Ireland
+353 1 608 1136
Padraig.Cunningham@tcd.ie

ABSTRACT

In this paper we present a data mining system, which allows the application of different clustering and cluster validity algorithms for DNA microarray data. This tool may improve the quality of the data analysis results, and may support the prediction of the number of relevant clusters in the microarray datasets. This systematic evaluation approach may significantly aid genome expression analyses for knowledge discovery applications. The developed software system may be effectively used for clustering and validating not only DNA microarray expression analysis applications but also other biomedical and physical data with no limitations. The program is freely available for non-profit use on request at http://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering – a*lgorithms*.
J.3. [**Life and Medical Sciences**]: - *biology and genetics*.

General Terms

Algorithms, Design, Experimentation,

Keywords

Gene Expression, Data Mining, Clustering, Cluster Evaluation, Validity Indices.

1. INTRODUCTION

The fast growth of data collections in the science and business applications as well as the need to analyse and extract useful knowledge from this data leads to a new generation of tools and techniques grouped under the term data mining. The recent advent of DNA microarray (or gene chips) technologies allows the measuring of the simultaneous gene expression of thousands of genes under multiple experimental conditions [8]. This technology is having a significant impact on genomic and post-genomic studies [17] such as disease diagnosis, drug discovery and toxicological research [5,12]. For instance, the accurate classification of tumours is essential for a successful diagnosis and treatment of cancer. One of the problems associated with cancer tumour classification is the identification of new classes using gene expression profiles. There are two key aspects in this problem: 1) estimation of the number of clusters in the dataset; and 2) classification of unknown tumour samples based on these clusters [6]. In this paper we address the first of these problems. This paper presents a data mining framework to evaluate DNA microarray data clustering results. The Machaon Cluster Validation Environment (*Machaon CVE*) system is intended for 1) performing clustering on microarray data; 2) evaluating the quality of the clusters obtained, which may also be used for estimating the “correct” number of clusters.

2. DEVELOPMENT APPROACH

A principal step in the analysis of gene expression data is the detection of samples or gene groups with similar expression patterns. Several clustering algorithms have been applied to the analysis of gene expression data [13,19]. Also solutions to systematically evaluate the quality of the clusters have been presented [3]. The prediction of the correct number of clusters is a fundamental problem in unsupervised classification. Many clustering algorithms require the definition of the number of clusters beforehand. To overcome this problem, various cluster validity indices have been proposed to assess the quality of a clustering partition [2]. This approach consists of running a clustering algorithm several times and obtaining different partitions, and the clustering partition that optimises the validity index under consideration is selected as the best partition. Thus, the main goal of a cluster validity technique is to identify the partition of clusters for which a measure of quality is optimal.

The recognition of these requirements in analysis of gene expression data led us to the development of *Machaon CVE* system. The major functions of the system can be summarised as follows:

- *Clustering*. In this step we define/extract clusters that correspond to the pre-defined number of clusters for a particular dataset. It offers some of the well-established clustering methods that are available in literature.
- *Evaluation of the clustering scheme or cluster validation*. The clustering methods can find a partition in a dataset, based on certain assumptions. Thus, an algorithm may result in different clustering schemes for

a dataset assuming different parameter values. *Machaon* evaluates the results of clustering algorithms based on quality indices and selects the clustering scheme that best fits the data. The definition of these indices is based on two fundamental criteria of clustering quality: cluster compactness and isolation.

3. SYSTEM OVERVIEW

The software is implemented as a multi-window Java application, which allows working with different datasets, clustering and validation algorithms, and results simultaneously. The *Machaon* tool is a data mining system based on the framework described in the previous section. The system provides the following services: 1) access to data, 2) implementation of clustering algorithms, 3) evaluation of clustering results, using cluster validity indices. The system supports several modifications of tabular data formats widely used by third-party clustering tools [14,18]. The focus has been made on clustering quality assessment and visualization of data mining results. The following are some highlights of the system:

- *Clustering*: Multiple clustering techniques may be applied to a dataset and the results may be easily compared. The user may select one of the available clustering algorithms in order to define a partitioning for the dataset. Depending on the clustering algorithm, the user defines the values of its input parameters. The results of a hierarchical clustering can also be displayed using dendrograms. Every clustering result may be selected and validated across a number of parameterised validation methods.

- *Cluster Validity*: Selecting the validation task the system searches for the optimal parameters' values for a specific clustering algorithm so as to result in a clustering scheme that best fits our data. The user selects the clustering algorithm and the input parameter based on which the validation task will be performed. Also, the range of input parameters values is defined. Several methods for measuring gene-to-gene (or sample-to-sample), intercluster and intracluster distances can be used in any combination. This is important to research the influence of different distance metrics on both clustering and validation. Both clustering and validation results are represented as a two-level tree in the bottom of the corresponding dataset window. Clustering indices are also displayed in additional columns of a dataset table. Every such column is associated with a single partition. *Machaon CVE* provides data normalization functionality, which may be either selected as an option of clustering/validation or used to produce a normalized dataset.

Apart from the clustering and validation results, the system shows, if known, the natural classification structure of the data, which allows comparisons against clustering results and validation analyses across natural classes.

4. APPLICATION FIELD AND SYSTEM IN USE

The clustering and validation methods included in *Machaon CVE* have been applied to gene expression datasets from recently published microarray studies: the leukemia dataset of Golub et al. [10], which contains 38 samples (27 *acute lymphoblastic leukemia*, ALL, and 11 *acute myeloid leukemia*, AML) represented by the expression values of 50 genes correlated with the AML and ALL cancer types; and *B-cell lymphoma* of Alizadeh et al. [1], which consists of 63 samples (45 *diffuse large B-cell lymphoma* and 18 normal) described by the expression levels of 23 genes.

The software is implemented as multi-window Java application, which allows working with multiple datasets, algorithms and results simultaneously. The *Main Window* (panel) contains the menu and indicates the current working dataset (Figure 1). Multiple *DataSet Windows* provide views on open datasets including expression data table and the *Result Tree,* which displays a list of all clustering and validation results obtained for corresponding dataset. Each row of the table may contain either single sample or single gene data accompanied with cluster indices for each partitioning.

Machaon CVE uses the textual tab-delimited data files described in Table1. The *Machaon* format is very similar to the Stanford tab-delimited format (http://genome-www5.stanford.edu/microarray/help/formats.shtml) and can be created and exported in any standard spreadsheet program, such as Microsoft Excel. The format provides a possibility of saving the clustering results within a dataset.

Table 1. *Machaon* tab-delimited format

Number of rows	**Number of columns**	S_1 or G_1	S_2 or G_2	…	S_i or G_j	
S_1 or G_1	**NC_1**	V_{11}	V_{12}	…	V_{1j}	C_1
S_2 or G_2	**NC_2**	V_{21}	V_{22}	…	V_{2j}	C_2
…	**…**	…	…	…	…	…
S_i or G_j	**NC_k**	V_{i1}	V_{i2}	…	V_{ij}	C_n

The *Number of rows* and *Number of columns* indicate the numerical values of rows and columns in the expression table. The terms S_i , $1< i < Ns$ are the names or descriptions of the experimental samples, conditions, strains, or specimens (number of the samples in the dataset equals *Ns*); G_j , $1< j < Ng$, are the names or descriptions of the gene names (number of the genes in the dataset equals *Ng*); NC_k, , $1< k < Nnc$ are the names or descriptions of the natural classes (number of the natural classes in the dataset equals *Nnc*). The terms V_{ij} represent the data values for the *i*th sample/experiment and the *j*th gene. The terms C_n , $1< n < Nc$ are the names of the clusters to which the sample/gene is referred (number of the clusters in the dataset equals *Nc*). Bold entries indicate necessary records. The program can read files, which already contain the number of clusters (datasets, which has already been clustered by other software tools). Thus, the user could apply the validation techniques to the data files, which are provided by other systems. An example of the described format is shown in Table 2.

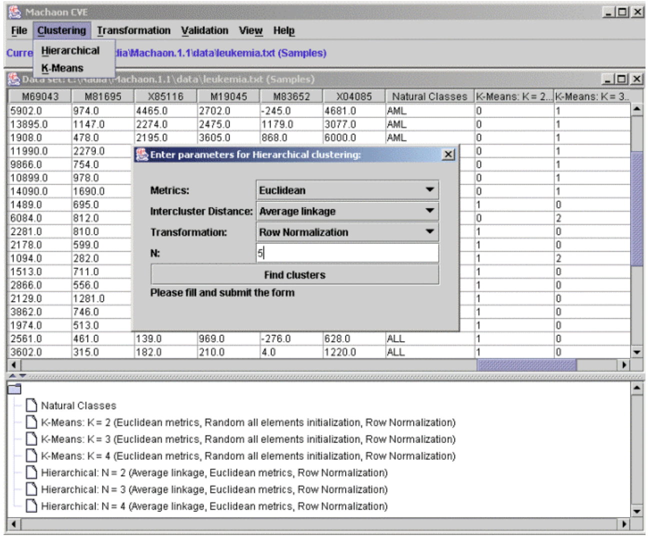
Table 2. An example originated from *leukemia* data. The format described in Table 1 is implemented to the data

5	3	U22376	X59417	U05259	
sample_12	ALL	551	846	2504	0
sample_25	ALL	1872	3878	5070	1
sample_34	AML	1126	782	711	1
sample_35	AML	880	490	654	0
sample_36	AML	473	1648	-14	1

4.1 Clustering in Machaon CVE

Several different types of clustering are implemented in the software. They include: hierarchical clustering (*single, complete, average, centroid, average to centroids* and *Hausdorff* linkages) and non-hierarchical clustering such as the *K-Means* algorithm [9]. Three types of metrics (*Euclidian, Manhattan* and *Chebychev* distances) could be used in clustering algorithms. Optional *Row Normalization* could be also applied to the microarray dataset.

To start the clustering calculation, the user may select a method from the submenu *Clustering* of the main menu. The *Parameter Window* will appear to select the clustering parameters described above.


Machaon CVE
File
Clustering
Transformation
Validation
View
Help
Hierarchical
K-Means
Curre
dia\Machaon.1.1\data\leukemia.txt (Samples)
achaon.1.1\data\leukemia.txt (Samples)

M69043	M81695	X85116	M19045	M83652	X04085	Natural Classes	K-Means: K= 2...	K-Means: K= 3..
5902.0	974.0	4465.0	2702.0	-245.0	4681.0	AML	0	1
13895.0	1147.0	2274.0	2475.0	1179.0	3077.0	AML	0	1
1908.0	478.0	2195.0	3605.0	868.0	6000.0	AML	0	1
11990.0	2279.0						0	1
9866.0	754.0						0	1
10899.0	978.0						0	1
14090.0	1690.0						0	1
1489.0	695.0						1	0
6084.0	812.0						0	2
2281.0	810.0						1	0
2178.0	599.0						1	0
1094.0	282.0						1	2
1513.0	711.0						1	0
2866.0	556.0						1	0
2129.0	1281.0						1	0
3862.0	746.0						1	0
1974.0	513.0						1	0
2561.0	461.0	139.0	969.0	-276.0	628.0	ALL	1	0
3602.0	315.0	182.0	210.0	4.0	1220.0	ALL	1	0

Enter parameters for Hierarchical clustering:

Metrics: Euclidean

Intercluster Distance: Average linkage

Transformation: Row Normalization

N: 5

Find clusters

Please fill and submit the form

Natural Classes
K-Means: K= 2 (Euclidean metrics, Random all elements initialization, Row Normalization)
K-Means: K= 3 (Euclidean metrics, Random all elements initialization, Row Normalization)
K-Means: K= 4 (Euclidean metrics, Random all elements initialization, Row Normalization)
Hierarchical: N= 2 (Average linkage, Euclidean metrics, Row Normalization)
Hierarchical: N= 3 (Average linkage, Euclidean metrics, Row Normalization)
Hierarchical: N= 4 (Average linkage, Euclidean metrics, Row Normalization)

Figure 1. Screenshots of the *DataSet Window* with the *Parameter Window* for hierarchical clustering.

For instance, hierarchical clustering may be applied to the *leukemia* dataset (Figure 1). As soon as the calculation is completed, a new entry is being added to the *Results Tree*. The result of clustering is also indicated in the expression table as a new cluster indices column appended to the right of the table.

In the case of hierarchical clustering, a user may view the results as a dendrogram. Selecting a clustering result in the result tree and then choosing the *View/Dendrogram* menu item will achieve that.

4.2 Validation

The *Machaon CVE* tool provides a collection of validity indices: *C-index* [15], *Davis-Bouldin index* [4], *Dunn's index* [7], *Goodman-Kruskal index* [11] and *Silhouette index* [16]. There are six types of intercluster distances (*single, complete, average, centroid, average to centroids* distances and *Hausdorff* metrics), three types of intracluster distances (*complete, average* and *centroid* diameters) and three types of metrics (*Euclidian, Manhattan* and *Chebychev* distances) that can be used with every method in any combination. For further information on the description of the types of metrics the reader is referred to [3]. The tool also provides data normalization functionality, which may be either selected as an option of clustering/validation or used to produce a normalized dataset.

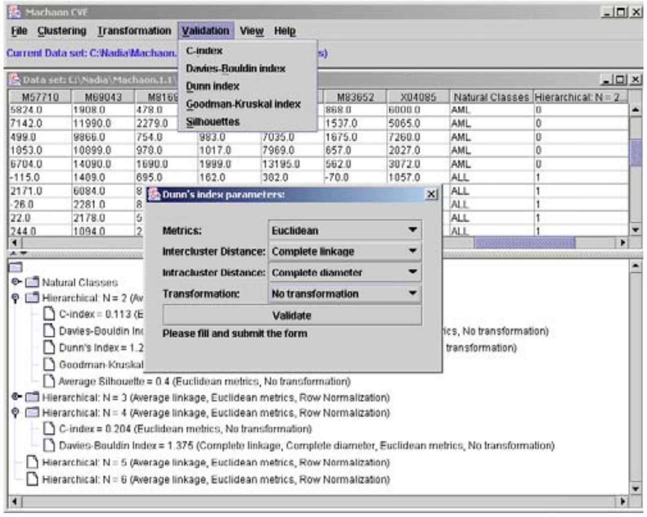
To apply a validation technique, it is necessary to select the *Cluster Set* in the *Result Tree* first and then choose the validation method from the *Validation* submenu of the main menu. Validation parameters may be adjusted using the *Parameters Window* and then the selected method may be executed. As soon as the calculation is completed, a new entry is being added to the *Results Tree*. The result of validation is attached to clustering result node in the tree (Figure 2).

As a way of illustration, different validity indices are applied to the *leukemia* cluster sets to find the optimal partitioning. Let's apply *C-index, Goodman-Kruskal, Silhouette, Dunn's* and *Davis-Bouldin* (with parameters: complete intercluster distance and complete intracluster diameter*) indices* to the partitioning of the *leukemia* dataset (number of clusters from 2 to 6) obtained by average linkage clustering. The results of the validation are shown in Table 3 (low values of the *C-Index* and the *Davis-Bouldin index* are indicative of strong clusters).

Table 3. Validity indices for expression clusters originating from *leukemia* data. Bold entries highlight the optimal number of clusters, *n*, predicted by each method

Validity index	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
C-index	**0.113**	0.199	0.204	0.192	0.171
Goodman-Kruskal index	**0.762**	0.494	0.477	0.506	0.548
Silhouette index	**0.4**	0.18	0.072	0.043	0.083
Dunn's index	**1.214**	0.566	0.361	0.361	0.201
Davis-Bouldin index	**1.29**	1.488	1.375	1.343	1.297

A user may now browse through a *Result Tree* and compare different partitioning validity indices to determine, for example, optimal clustering parameters. In our case, as it is seen on Table 3, it may be concluded that the most appropriate partitioning for the *leukemia* dataset consists of two clusters, which is supported by all validation methods.



Current Data set: C:\Nadia\Machaon. s)

Data set: L:\Nadia\Machaon.1.1

M57710	M69043	M8169			M83652	X04085	Natural Classes	Hierarchical: N = 2
5824.0	1908.0	478.0			868.0	6000.0	AML	0
7142.0	11990.0	2279.0			1537.0	5065.0	AML	0
499.0	9866.0	754.0	983.0	7035.0	1675.0	7260.0	AML	0
1053.0	10899.0	978.0	1017.0	7969.0	657.0	2027.0	AML	0
6704.0	14090.0	1690.0	1999.0	13195.0	562.0	3072.0	AML	0
-115.0	1409.0	695.0	162.0	302.0	-70.0	1057.0	ALL	1
2171.0	6084.0	8					ALL	1
-26.0	2281.0	8					ALL	1
22.0	2178.0	5					ALL	1
244.0	1094.0	2					ALL	1

Dunn's index parameters:

Metrics: Euclidean

Intercluster Distance: Complete linkage

Intracluster Distance: Complete diameter

Transformation: No transformation

Please fill and submit the form

Natural Classes
Hierarchical: N = 2 (Av
C-index = 0.113 (E
Davies-Bouldin In ics, No transformation)
Dunn's Index = 1.2 transformation)
Goodman-Kruskal
Average Silhouette = 0.4 (Euclidean metrics, No transformation)
Hierarchical: N = 3 (Average linkage, Euclidean metrics, Row Normalization)
Hierarchical: N = 4 (Average linkage, Euclidean metrics, Row Normalization)
C-index = 0.204 (Euclidean metrics, No transformation)
Davies-Bouldin Index = 1.375 (Complete linkage, Complete diameter, Euclidean metrics, No transformation)
Hierarchical: N = 5 (Average linkage, Euclidean metrics, Row Normalization)
Hierarchical: N = 6 (Average linkage, Euclidean metrics, Row Normalization)

Figure 2. Screenshots of the *DataSet Window* with the *Parameter Window* for *Dunn's index* validation.

The software may be also used to research the influence of different metrics and linkage on clustering and/or validation methods. For instance, Table 4 contains results of validation by *Dunn's index* of the same *B-cell lymphoma* dataset partitioning by hierarchical method with different number of clusters and different linkage calculation algorithms used. Hence, the researcher may conclude that *Hausdorff* linkage used in hierarchical method produces noticeably different partitioning.

Table 4. *Dunn's* validity indices for expression clusters originating from *B-cell lymphoma* data. Bold entries highlight the optimal number of clusters, *n*, predicted by each hierarchical clustering method

Clustering	*n* = 2	*n* = 3	*n* = 4	*n* = 5	*n* = 6
Complete linkage	**1.25**	0.859	1.08	0.813	0.813
Average linkage	**1.096**	0.641	0.715	0.881	0.743
Single linkage	**1.096**	0.567	0.567	0.673	0.673
Centroid linkage	**0.855**	0.597	0.686	0.567	0.528
Average to centroids	**0.855**	0.597	0.718	0.541	0.534
Hausdorff linkage	0.818	0.853	**1.11**	0.977	0.776

5. CONCLUSION

This paper describes a software tool (*Machaon CVE*) that offers multiple clustering and cluster validity methods for DNA microarray data analysis. There are different commercial and non-commercial software packages and web applications available with implementations of different clustering methods, but they lack facilities for estimating the optimal number of clusters, as well as components for evaluating the quality of the clusters obtained. The *Machaon CVE* allows the application of various validation methods to multiple datasets, which may be clustered by third-party tools. Five validation and two clustering techniques (with various combination of gene-to-gene (or sample-to-sample), intercluster and intracluster distances) have been implemented in this system.

The tool described in this paper will contribute to the evaluation of clustering outcome and the identification of optimal cluster partitions. The estimation approach described represents an effective tool to support Even though *Machaon CVE* was developed for DNA microarray expression analysis applications, it may be effectively used for clustering and validating other biomedical and physical data with no limitations.

6. ACKNOWLEDGMENTS

This research has been partly supported by grants from *Science Foundation Ireland* and *Enterprise Ireland Research Innovation* Fund 2001.

7. REFERENCES

[1] Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L. Marti G.E., Moore T., Hudson J., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Bird J.C., Botstein D., Brown P.O. and Staudt M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403 (2000), 503-511.

[2] Azuaje F. A cluster validity framework for genome expression data. Bioinformatics, 18 (2002), 319-320.

[3] Bolshakova, N. and Azuaje F. Cluster validation techniques for genome expression data. Signal Processing, 83 (2003), 825-833.

[4] Davies D.L. and Bouldin D.W. A cluster separation measure. IEEE Transactions on Pattern Recognition and Machine Intelligence, 1 (1979), 224-227.

[5] Debouck C. and Goodfellow P.N. DNA microarrays in drug discovery and development. Nature Genet., 21 (1999), 48-50.

[6] Dudoit S. and Fridlyand J. A prediction-based resampling method for estimation the number of cluster in a dataset. Genome Biology, 3 (2002), 1-21.

[7] Dunn J. Well separated clusters and optimal fuzzy partitions. J.Cybernetics, 4 (1974), 95-104.

[8] Eisen M.B., Spellman P.T., Brown P.O., and Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95 (1998), 14863-8.

[9] Everitt B. Cluster Analysis, Edward Arnold, London, 1993.

[10] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeck M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D. and Lander E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286 (1999), 531-537.

[11] Goodman L. and Kruskal W. Measures of associations for cross-validations. J. Am. Stat. Assoc., 49 (1954), 732-764.

[12] Grey N.S., Wodicka L., Thunnissen A.M., Norman T.C., Kwon S., Espinoza F.H., Morgan D.O., Barnes G., LeClerc S., Meijer L., Kim S.H., Lockhart D.J. and Schultz P.G. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. Science, 281 (1998), 533-538.

[13] Granzow M., Berrar D., Dubitzky W., Schuster A., Azuaje F. and Eils R. Tumor identification by gene expression profiles: a comparison of five different clustering methods. ACM-SIGBIO Newsletters, 21 (2001), 16-22.

[14] Herrero J., Valencia A. and Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics, 17 (2001), 126-136.

[15] Hubert L. and Schultz J. Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie, 29 (1976), 190-241.

[16] Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comp App. Math, 20 (1987), 53-65.

[17] Schena M., Shalon D., Davis R.W. and Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270 (1995), 467-470.

[18] Sturn A., Quackenbush J. and Trajanoski Z. Genesis: cluster analysis of microarray data. Bioinformatics, 18 (2002), 207-208.

[19] Yeung K.Y., Haynor D.R. and Ruzzo W.L. Validating clustering for gene expression data. Bioinformatics, 17 (2001), 309-318.