Estimating the Number of Clusters in DNA Microarray Data

N. Bolshakova[1], F. Azuaje[2]

[1]Department of Computer Science, Trinity College Dublin, Ireland

[2]School of Computing and Mathematics, University of Ulster, Jordanstown, Northern Ireland, U.K

Summary

Objectives

The main objective of the research is an application of the clustering and cluster validity methods to estimate the number of clusters in cancer tumor datasets. A weighed voting technique is going to be used to improve the prediction of the number of clusters based on different data mining techniques. These tools may be used for the identification of new tumour classes using DNA microarray datasets. This estimation approach may perform a useful tool to support biological and biomedical knowledge discovery.

Methods

Three clustering and two validations algorithms were applied to two cancer tumour datasets. Recent studies confirm that there is no universal pattern recognition and clustering model to predict molecular profiles across different datasets. Thus, it is useful not to rely on one single clustering or validation method, but to apply a variety of approaches. Therefore, combination of these methods may be successfully used for the estimation of the number of clusters.

Results

The methods implemented in this research may contribute to the validation of clustering results and the estimation of the number of clusters. The results show that this estimation approach may represent an effective tool to support biomedical knowledge discovery and healthcare applications.

Conclusion

The methods implemented in this research may be successfully used for the estimation of the number of clusters. The methods implemented in this research may contribute to the validation of clustering results and the estimation of the number of clusters. These tools may be used for the identification of new tumour classes using gene expression profiles.

Keywords
Gene expression, data mining, clustering, cluster evaluation, validity indices

Introduction

DNA microarray technology is increasingly being applied in biological and biomedical research to address a number of critical problems including the classification of tissue samples, e.g. cancer tumours. Recent advances allow the monitoring of the expression levels of thousands of genes simultaneously under multiple experimental conditions [1]. This technology is having a significant impact on genomic and post-genomic studies. Disease diagnosis, drug discovery and toxicological research benefit from the of microarray technology. A principal step in the analysis of gene expression data is the detection of samples or gene groups with similar expression patterns. 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 [2].

A variety of cluster algorithms have been applied to the analysis of DNA microarray data [3,4]. Moreover, a number of solutions to systematically evaluate the quality of the clusters have been presented [5,6,7]. The estimation of the number of clusters in a dataset is a fundamental problem in unsupervised learning. The applications of several validation techniques such as the *Silhouette method* [8], *Dunn's based index* [9,10] and *Davies-Bouldin index* [11] have been previously studied [5,7,12].

Methods

This section introduces the DNA microarray data and the data mining methods under consideration. Three clustering methods: *K-Means*, *Hierarchical (complete linkage)* and *Kohonen Self-organising Maps* [13,14], and two validation methods: the *C-index* [15] and the *Goodman-Kruskal index* [16] were applied. The data studied in this paper consisted of two expression datasets originating from recently published microarray studies [17,18].

Microarray data

The *central nervous system (CNS)* dataset [17] comprise 42 tumour samples (10 *medulloblastomas,* 5 *CNS atypical teratoid/rhabdoid tumours,* 5 *renal and extrarenal rhabdoid,* 8 *supratentorial primitive neuroectodermal tumours,* 10 *non-embryonal brain tumours* and 4 *normal human cerebellas)* described by the expression levels of 50 genes with suspected roles in these types of cancer. These data were obtained from a study published by Pomeroy and co-workers [17]. They demonstrated that *medulloblastomas* are molecularly distinct from other brain tumours.

The *leukaemia* data include 38 samples (27 *acute lymphoblastic leukaemia*, ALL, and 11 *acute myeloid leukaemia*, AML) described by the expression levels of 50 informative genes, which are correlated with the AML and ALL cancer types. These data were obtained from a study published by Golub and co-workers [18]. They presented a model to distinguish two sub-classes of ALL samples, known as *B-cell ALL* and *T-cell ALL.*

The original data and experimental methods for both datasets are available at http://www.genome.wi.mit.edu/MPR.

Cluster validation methods

In this paper cluster validation is performed using two algorithms: the *C-index* [15] and the *Goodman-Kruskal index* [16]. These methods have been chosen to support the investigation of cluster validation techniques for genome expression data classification. For more information on the implementation and analysis of other validation algorithms the reader is referred to our previous studies [5,7,12].

C-index

For any partition $U \leftrightarrow X$: $X_1 \cup ... X_i \cup ... X_n$, where X_i represents the i^{th} cluster of such partition, the *C-index* [15], *C*, is defined as:

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}}, \qquad (1)$$

where S, S_{min}, S_{max} are calculated as follows. Assume that p is the number of all pairs of samples for which both samples are located in the same cluster. Then S is the sum of distances between samples in those p pairs. Let P be a number of all possible pairs of samples in the dataset. Ordering those P pairs by distances we can select p pairs with smallest and p pairs with largest distances between samples. The sum of the p smallest distances is equal to S_{min}, whilst the sum of the p largest is equal to S_{max}. From this formula it follows that the nominator will be small if pairs of samples with small distances are in the same cluster. Thus, small values of C correspond to good clusters. The number of clusters that minimize *C-index* is taken as the optimal number of clusters, n.

Goodman-Kruskal index

For a given dataset, X_j ($j = 1, \ldots, k$, where k is the number of samples, j, in the dataset), this method assigns all possible *quadruples* [16]. Let d be the distance between any two samples (a and b, or c and d) in X_j. A *quadruple* is called *concordant* if one of the following two conditions is true:

$$d(a,b) < d(c,d) \qquad (2)$$

a and *b* are in the same cluster and *c* and *d* are in different clusters.

$$d(a,b) > d(c,d) \qquad (3)$$

a and *b* are in different clusters and *c* and *d* are in the same cluster.

By contrast, a *quadruple* is called *disconcordant* if one of following two conditions is true:

$$d(a,b) < d(c,d) \qquad (4)$$

a and *b* are in different clusters and *c* and *d* are in the same cluster.

$$d(a,b) > d(c,d) \qquad (5)$$

a and *b* are in the same cluster and *c* and *d* are in different clusters.

A good partition is one with many *concordant* and few *disconcordant quadruples*. Let N_c and N_d denote the number of *concordant* and *disconcordant quadruples*, respectively. Then the *Goodman-Kruskal index*, *GK*, is defined as:

$$GK = \frac{N_c - N_d}{N_c + N_d}, \qquad (6)$$

Large values of *GK* are associated with a good partition. Thus, the number of clusters that maximize the *GK* index is taken as the optimal number of clusters, n.

Results

Three clustering algorithms were implemented to produce different partitions consisting of 2 to 10 clusters. Then, the validity indices were computed for each of the partitioning results. The Euclidean metric was used for all cases to calculate the distances between the samples.

Tables 1 and 2 depict the *C-indices* and *Goodman-Kruskal indices* for each number of clusters, *n*, for *n* = 2 to *n* = 10, using the *CNS* dataset for three clustering algorithms: *K-Means*, *Hierarchical* (*complete linkage*) and *Self-organising Maps*.

The bold entries correspond to the optimal partitions predicted by each validation method. For the *CNS* expression dataset, *n* = 4 is suggested as the best partition. The *CNS* dataset includes the classes: *medulloblastoma*, *CNS rhabdoid* (with *brain* and *renal* subclasses), *PNET*, *malignant glioma* and *normal human cerebella*. Table 3 depicts the clustering results for the best predicted partition for *CNS* data.

An examination of this partition confirms that *normal human cerebella* (*Nc*) is distinguished from other types of cancer in the dataset. Subclasses (*brain* and *renal*) of *CNS rhabdoid* (*Rh*) tend to locate in the same cluster, as well as *medulloblastoma* (*MD*) samples, which are mostly placed in the same cluster. *PNET* and *malignant glioma* (*MG*) are difficult to distinguish in this partition.

Tables 4 and 5 show the *C-indices* and *Goodman-Kruskal indices* for each number of clusters, *n*, for *n* = 2 to *n* = 10, using the *leukaemia* dataset for three clustering algorithms.

An examination of these *leukaemia* data results suggests that the most appropriate partition includes two clusters (for *K-Means* and *SOM* clustering) and three clusters (for *Hierarchical* clustering). Table 6 depicts the clustering results for the partition predicted as optimal for *leukaemia* data. It is shown that, for each of the considered indices, the correct number of clusters corresponds to an optimal index value.

This validation approach may also consist of the implementation of an aggregation method based on a weighed voting strategy, which is implemented in [5,12]. This voting strategy may also be applied to fuse the results originating from different clustering and validation methods. In this study, after computing all validity indices for all obtained clustering techniques, the average weighed vote for each partition has been calculated. Table 7 represents the implementation of the average weighed vote strategy for the *leukaemia* data. This table was obtained from Tables 4 and 5 by replacing the index values by weighed votes, whose values range from 1 to 9 [5,12]. Thus, the average weighed vote for each cluster partition has been calculated, and *n* = 2 is suggested as the optimal partition.

Discussion

Clustering has become a fundamental data mining approach to analysing DNA microarray data [3,4]. It can support the identification of existing primary relationships among a set of variables such as biological conditions or perturbations. Clustering may represent a basic tool not only for the classification of known categories, but also for the discovery of relevant classes. The description and interpretation of its outcomes may also allow the detection of associations between samples or variables, the generation of rules for decision-making support and the evaluation of experimental models [5]. In the genome expression domain it has provided the basis for novel clinical diagnostic and prognostic studies [19], and other applications using different model organisms [20].

Cluster validity indices represent important tools to support unsupervised data mining. They are particularly useful in applications in which the definition of the number of clusters in the dataset is required beforehand.

In this paper three clustering algorithms (*K-Means*, *Hierarchical* and *Self-organising Maps*) and two validation indices (*C-index* and *Goodman-Kruskal*) were applied to two cancer tumour datasets (*CNS* and *leukaemia*). Recent studies confirm that there is no universal pattern recognition and clustering model to predict molecular profiles across different datasets [5]. A number of clustering [4,13] and validation methods [5,7,12] have been previously studied. Each of these methods has their advantages and limitations. For example, it has been shown that the *Silhouette* method [8] is suitable for estimating only the first choice or best partition. Nevertheless, this method has been successfully used in combination with other validation techniques (*Dunn's* and *Davies-Bouldin* indices) for predicting different optimal clustering partitions [5]. *Goodman-Kruskal* index is expected to be robust against outliers because quadruples of patterns are used for its computation. However, its drawback is a high computational complexity in comparison, for example, with the *C-index*. On the other hand, *K-Means* clustering is dependent on the initial seed cases and a disadvantage of the *Hierarchical* clustering is that the identification of categories and associations is left to the user. Furthermore, if a wrong assignment is made early in the process of hierarchal clustering, it cannot be corrected. Thus, it is useful not to rely on one single clustering or validation method, but to apply a variety of approaches. Therefore, combination of these methods may be successfully used for the estimation of the number of clusters. It has been shown that these methods may support the prediction of the optimal partition [5,7]. A weighed voting technique [5,12] was used to improve the prediction of the number of clusters based on different data mining techniques. Current and future work includes the comparison, combination and estimation of results obtained from different clustering algorithms, and the analysis of more complex datasets.

Conclusion

The methods implemented in this research may be successfully used for the estimation of the number of clusters. The methods implemented in this research may contribute to the validation of clustering results and the estimation of the number of clusters. These tools may be used for the identification of new tumour classes using gene expression profiles. The results show that this estimation approach may represent an effective tool to support biomedical knowledge discovery and healthcare applications.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 1998; 14863-8.
2. Dudoit S, Fridlyand J. A prediction-based resampling method for estimation the number of cluster in a dataset. Genome Biology 2002; 1:21.
3. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics 2001; 309-318.

4. Granzow M, Berrar D, Dubitzky W, Schuster A, Azuaje F, Eils R. Tumor identification by gene expression profiles: a comparison of five different clustering methods. ACM-SIGBIO Newsletters 2001; 16-22.
5. Bolshakova N, Azuaje F. Cluster validation techniques for genome expression data. Signal Processing 2003; 825-33.
6. Azuaje F, Bolshakova N. Clustering genome expression data: design and evaluation principles. A Practical Approach to Microarray Data Analysis 2003; 230-45.
7. Azuaje F. A cluster validity framework for genome expression data. Bioinformatics 2002; 319-20.
8. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comp App. Math 1987; 53-65.
9. Dunn J. Well separated clusters and optimal fuzzy partitions. J.Cybernetics 1974; 95-104.
10. Bezdek JC, Pal NR. Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics 1998; 301-15.
11. Davies DL, Bouldin DW. A cluster separation measure. IEEE Transactions on Pattern Recognition and Machine Intelligence 1979; 224-7.
12. Bolshakova N, Azuaje F. Improving expression data mining through cluster validation. Proc. of the 4th Annual IEEE Conf. on Information Technology Applications in Biomedicine 2003; 19-22.
13. Quackenbush J. Computational analysis of microarray data. Nature Reviews Genetics 2001; 418-27.
14. Everitt B. Cluster Analysis 1993.
15. Hubert L, Schultz J. Quadratic assignment as a general data-analysis strategy . British Journal of Mathematical and Statistical Psychologie 1976; 190-241.
16. Goodman L, Kruskal W. Measures of associations for cross-validations. J. Am. Stat. Assoc. 1954;732-64.
17 Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black P, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR. Gene expression-based classification and outcome prediction of central nervous system embryonal tumors. Nature 2002; 436-42.
18. Golub TR, Slonim DK, Tamayo P, Huard C, Gassenbeck M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 531-7.
19. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2002; 536-40.
20. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, En JK, Bumgarner R, Goodlett DR, Aebersol R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbated metabolic network. Science 2001; 929-33.

Table 1. *C-indices* for expression clusters originating from the *CNS* data. Bold entries highlight the optimal number of clusters, *n*, predicted by this method.

Clustering	*K-Means*	*Hierarchical*	*SOM*
$n = 2$	0.249	0.132	0.334
$n = 3$	0.04	0.016	0.21
$n = 4$	**0.036**	**0.014**	**0.135**
$n = 5$	0.082	0.086	0.157
$n = 6$	0.075	0.052	0.151
$n = 7$	0.082	0.044	0.139
$n = 8$	0.084	0.031	0.141
$n = 9$	0.047	0.028	0.145
$n = 10$	0.054	0.028	0.143

Table 2. *Goodman-Kruskal indices* for expression clusters originating from the *CNS* data. Bold entries highlight the optimal number of clusters, *n*, predicted by this method.

Clustering	*K-Means*	*Hierarchical*	*SOM*
$n = 2$	0.543	0.781	0.325
$n = 3$	0.901	0.968	0.512
$n = 4$	**0.908**	**0.971**	**0.679**
$n = 5$	0.788	0.777	0.622
$n = 6$	0.806	0.852	0.632
$n = 7$	0.787	0.889	0.65
$n = 8$	0.782	0.927	0.66
$n = 9$	0.885	0.935	0.632
$n = 10$	0.87	0.938	0.634

Table 3. Clustering for *CNS* data. Partition predicted as the optimal choice in Tables 1 and 2.

Cluster	*K-Means*	*Hierarchical*	*SOM*
1	9MD, 10 Rh, 8 PNET, 1 MG	9 MD, 10 Rh, 8 PNET, 5 MG	1 MD, 10 Rh, 4 PNET, 1 MG
2	9 MG	5 MG	9 MD, 1 PNET
3	1 MD	1 MD	3 PNET, 9 MG
4	4 Nc	4 Nc	4 Nc

MD – *medulloblastoma*; Rh - *CNS atypical teratoid/rhabdoid tumours;* PNET – *primitive neuroectodermal tumours*; MG - *malignant glioma*; Nc - *normal human cerebella*.

Table 4. *C-indices* for expression clusters originating from *leukaemia* data. Bold entries highlight the optimal number of clusters, *n*, predicted by this method

Clustering	*K-Means*	*Hierarchical*	*SOM*
n = 2	**0.042**	0.038	**0.113**
n = 3	0.05	**0.023**	0.165
n = 4	0.094	0.096	0.149
n = 5	0.044	0.09	0.233
n = 6	0.064	0.058	0.173
n = 7	0.056	0.054	0.195
n = 8	0.054	0.049	0.158
n = 9	0.051	0.047	0.186
n = 10	0.048	0.042	0.183

Table 5. *Goodman-Kruskal indices* for expression clusters originating from *leukaemia* data. Bold entries highlight the optimal number of clusters, *n*, predicted by this method

Clustering	*K-Means*	*Hierarchical*	*SOM*
n = 2	**0.932**	0.942	**0.762**
n = 3	0.886	**0.96**	0.587
n = 4	0.76	0.727	0.594
n = 5	0.884	0.743	0.355
n = 6	0.812	0.825	0.491
n = 7	0.827	0.836	0.419
n = 8	0.832	0.85	0.525
n = 9	0.843	0.856	0.436
n = 10	0.851	0.861	0.45

Table 6. Clustering for *leukaemia* data. Partition predicted as the optimal choice in Tables 4 and 5.

Cluster	*K-Means*	*Hierarchical*	*SOM*
1	1 AML, 27 ALL	27 ALL, 2 AML	25 ALL
2	10 AML	4 AML	11 AML, 2ALL
3	-	5 AML	-

AML – *acute myeloid leukaemia*; ALL - *acute lymphoblastic leukaemia*

Table 7. Predicting the correct number of clusters for *leukaemia* data by aggregation of clustering and validation methods. Bold entries highlight the optimal number of clusters, *n*, predicted by the methods.

Clustering	*Validation*	*n*=2	*n*=3	*n*=4	*n*=5	*n*=6	*n*=7	*n*=8	*n*=9	*n*=10
K-Means	*C-index*	**9**	6	1	8	2	3	4	5	7
	GK index	**9**	8	1	7	2	3	4	5	6
Hierarchical	*C-index*	8	**9**	1	2	3	4	5	6	7
	GK index	8	**9**	1	2	3	4	5	6	7
SOM	*C-index*	**9**	6	8	1	5	2	7	3	4
	GK index	**9**	7	8	1	5	2	6	3	4
Average		**8.7**	7.5	3.3	3.5	3.3	3.0	5.2	4.7	5.8