Medical & Biological Engineering & Computing

Clustering of pancreatic endocrine tumors via microarray gene expression analysis

--Manuscript Draft--

Manuscript Number:	
Full Title:	Clustering of pancreatic endocrine tumors via microarray gene expression analysis
Article Type:	Original article
Keywords:	Microarray data; PDDP+K-means clustering; Gene selection; Linear Multivariable Classification.
Corresponding Author:	Silvia Carla Strada Milano, ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Silvia Carla Strada
First Author Secondary Information:	
Order of Authors:	Silvia Carla Strada
	Diego Ettore Italo Liberati, PhD Research Director
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	A simple, multivariable and linearly initialized clustering is shown to be able to deal with unsupervised classification of the data originating from pancreatic endocrine tumors (PET). Results are discussed almost only on the data science side, leaving a more biological discussion to future work, even in the quest of possible hidden pathways.
Suggested Reviewers:	Gabriella Pasi Full Professor, Universita della Svizzera Italiana gabriella.pasi@usi.ch Expert in Information Extraction from Data
	Giancarlo Mauri Director Dept. of Information, Systems and Commmunications, Universita degli Studi di Milano-Bicocca mauri@disco.unimib.it Full Professor in Bioinformatics
	Sergio Matteo Savaresi, Full Professor Deputy Director of Department, Politecnico di Milano savaresi@elet.polimi.it Original Developer of the methodology, cited and used

Noname manuscript No.
(will be inserted by the editor)

Clustering of pancreatic endocrine tumors via microarray gene expression analysis

Silvia Strada · Diego Liberati

Received: date / Accepted: date

Abstract A simple, multivariable and linearly initialized clustering is shown to be able to deal with unsupervised classification of the data originating from pancreatic endocrine tumors (PET). Results are discussed almost only on the data science side, leaving a more biological discussion to future work, even in the quest of possible hidden pathways.

Keywords Microarray data · PDDP+K-means clustering · Gene selection · Linear Multivariable Classification.

1 INTRODUCTION

Microarray technology has led to a rapid increase of information about gene expression of subjects in different phato-physiological conditions. The key issue is how to extract useful clinical information from such huge databases, where the number of genes is significantly greater than the number of subjects, in order to either retrieve the patient's case history from gene expression data or to find out which genes are the most significant for subjects discrimination.
The first type of objective can be achieved via automatic clustering of subjects into homogeneous groups (see for example [7], [5] for a detailed coverage of the topic). As it is well known, one can distinguish between supervised and unsupervised procedures [6]. The former,[4] e [9] e [8], uses apriori information on the data, such as subjects' pathological condition, along with subjects' gene expression information, to train a classifier which should be able to distinguish among different pathologies on the basis of gene expression profiles. The obtained classifier can then be used for diagnostic purposes on new subjects. On the opposite side, unsupervised clustering procedures, [11], perform the classification just on the basis of the gene expression dataset itself without apriori knowledge of the subjects pathological condition; this unsupervised clustering is therefore a tool used to discover the gene expression signature of newly discovered pathologies. Our approach is somehow in between: a first step of unsupervised clustering is then followed by a second supervised step, associating and comparing found classes in the previous step to apriori known (if no misclassifications exist in the original data) labeled groups of subjects.
In general, clustering plays a fundamental role in discovering the mechanisms of cellular malfunctioning. It is known, that many multifactorial diseases (e.g. diabetes or cancer) are accompanied by the deregulation of some genes, which are maybe over or under expressed so as to produce abnormal quantities of specific proteins. Needless to say, understanding in detail such deregulation processes would be a most valuable contribution to therapies development.
Concerning the objective of finding which genes are the most significant for a given pathology, classification of subjects on the basis of microarray data is one of the preliminary steps. However, since the number of genes to deal with is normally very large, the cluster discrimination rule returned by standard clustering procedures is based on many genes too. Hence, a further step is normally required to extract the most significant genes among the whole set.

S. Strada
the Department of Electronics, Information Science, and Bioengineering, Politecnico di Milano, Italy, E-mail: silvia.strada@polimi.it
D. Liberati
National Research Council (CNR) of Italy, Institute EIIT @DEIB PoliMI, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, E-mail: diego.liberati@cnr.it

In this paper we consider a dataset of microarray results obtained on Pancreatic Endocrine Tumors (PET), the classification of which is done on the basis of the positivity for insulin by means of immunohistochemstry (IHC).

Thus, the tumors are classified as:

- functional (F), when the tumor symptoms are caused by hormone islands (insulinomas) i.e. well differentiated endocrine tumors (WDET), which have an indolent clinical course
- non functional (NF), when the tumor does not secrete insulin, among which a further subclassification is usual between well differentiated endocrine carcinomas (WDEC), prone to invasion and metastasis, and poorly differentiated endocrine carcinomas (PDEC), exhibiting the worst prognosis with subjects' survival comparable to adenocarcinomas.

The data set consists of samples obtained from: 11 insulin positive tumors, classified by means of IHC (insulinomas), 25 insulin negative tumors (WDEC; PDEC) and 4 Human islet Preparations from cadaveric donors (HP), which represent the reference healthy tissue to these types of pancreatic tumors. Among the 25 insulin negative tumors, the microarray analysis revealed that a subset expressed insulin mRNA at considerable levels, similar to insulinomas and HP. Therefore three types of tumor cases were defined as -/- (tumors negative both in IHC and mRNA expression studies), -/+ (tumors negative in IHC but positive in mRNA expression studies), +/+ (positive both in IHC and mRNA expression studies).
In this paper we thus consider a dataset of pancreatic cancer subjects (observations) vs. genes (variables), where each patient is associated with one of four different classes based on the positivity for insulin, either by immunohistochemistry (IHC) or by mRNA expression. The aim of the study is to apply an unsupervised clustering procedure to the entire dataset in order to identify specific genes, and possibly pathways, potentially implicated in the control of insulin expression. The adopted methodology is based on two main phases, the first being the unsupervised clustering step and the second being the extraction of the most significant genes for subjects' classification.
Such a method was first developed in [14] and formulated as an improved methodology to cluster a generic dataset. This approach to data analysis was then successfully applied to discriminate between two kinds of leukemia, [3], enabling the classification without any knowledge of the pathology of the subjects. A further application on proprietary data of the Istituto dei Tumori di Milano, has highlighted the fact that the above method, though sufficiently general, doesn't however work on every dataset, being based on the hypothesis that the clusters are linearly separable. Shouldn't this be the case, one has to resort to more sophisticated solutions which take into account the possibility to split each cluster in a non linear way, for instance, for simplicity, either piecewise linear, [2], or even binary, [12], through logical networks identifying both salient genes and binary composition rules, able to correctly partition the clusters. A more sofisticated approach could be to resort to adaptive bayesian networks, whose complexity is objectively determined, as in [1], thanks to the minimum description lenght, [13].
The microarray dataset is described in Subsection 2.1, while Subsection 2.2 addresses the central problem of unsupervised clustering. In Subsection 2.3 the final gene reduction phase is illustrated. The obtained results are subject of Section 3. A concise discussion is the object of Section 4.

2 MATERIALS AND METHODS

2.1 Dataset description

The raw microarray data were obtained using a custom array, which analyzed 72 primary pancreatic tumors, [10]. Of these 72 cases, in the present study there were analyzed, along with 4 human pancreatic islets samples (HP), only: insulinomas (11 cases, insulin positive by immunohistochemstry, WDET); 25 insulin negative by immunohistochemistry, namely those classified as well differentiated endocrine tumors (WDEC) and poorely differentiated endocrine carcinomas (PDEC).
The microarray dataset was first normalized by means of Robust Multiarray average (RMA) and since the 25 cases, classified originally as IHC negative for insulin, showed a subset of subjects with a signifcant level of mRNA expression for insulin, this group was further split into two subgroups, obtaining at the end four groups to be compared. P-values based on a permutation test on the entire dataset and Bonferroni correction, allowed extraction of 542 genes significantly expressed in the three tumor groups: +/+ positive both for insulin protein and mRNA, -/+ negative for the protein, positive for mRNA, -/- negative both for the protein and mRNA and in the human pancreatic islets (HP).
The dataset is constituted by the expression of 542 genes (also denominated variables) over 40 subjects (also denominated observations) of which 36 suffering from pancreatic cancer. Just to clarify the dataset structure, a small part of it is depicted in Figure 1. The rows correspond to subjects while columns are associated with hu-

man genes and are treated as partially independent, descriptive variables. Each patient is associated with 542 real values, each measuring the expression level of the corresponding gene. The smaller the expression value, the less the corresponding gene is activated. Moreover, each patient has been assigned to one of the four different above defined categories so that each observation is associated with a different clinical condition.

	NM_005302	NM_004810	NM_003385	NM_000898	.....
Patient1	7,645	5,079	3,055	5,232	.....
Patient2	7,965	5,887	9,458	9,696	.....
Patient3	6,899	6,342	7,458	8,257	.....
Patient4	6,490	4,969	4,704	3,196	.....
Patient5	6,985	6,117	6,088	7,249	.....
.....	.....	.....	.....	.....	.....

Fig. 1 subjects-Genes dataset

2.2 Clustering

2.2.1 Principal components analysis

Principal components analysis (PCA) is a well known multivariable technique to represent the data in a new space whose variables are linear combinations of the original ones. The new variables are orthogonal to each other and ordered in such a way that the explained variance is decreasing. It is thus easier to select a small set of variables still explaining a significant percentage of the total variability in the observations. The main advantages of such an approach are, on one side the ease of graphical representation in a reduced dimensions space, and, on the other side, the possibility to identify a hierarchy among the variables most useful for the classification. More in detail, PCA returns a new set of coordinates, which are ordered in such a way that: the first one, the first principal component, denotes the direction with the greatest intersubject variance; the second one (the second principal component) has the greatest intersubject variance among the residual ones; and so on.

2.2.2 Principal divisive partitioning and k-means on the original dataset

In order to perform data clustering, a reasonable way to initialize the well known k-means algorithm is to split the entire dataset into two subsets (bisection) so as to minimize the similarity of data belonging to opposite subsets and to maximize the similarity of those pertaining to the same one, [7] and [5]. Such initial separation of data is then optimized, [14], via the classical k-means algorithm. Then, the same procedure is iteratively applied, each time dividing a single cluster, among those obtained in the previous step, until a final partition of the initial dataset is reached, satisfying a suitable optimization index. In the present paper, instead of such an index, the a priori knowledge of the true classification of each subject, is simply used to stop bisecting when most of the subjects pertaining to the same class belong to the same cluster.

2.3 Extraction of minimal gene sets

At each partition, we are interested in focusing on the subset of original variables (genes) really needed to bisect the data. In fact, all variables contribute to every principal component, and thus to the corresponding partition, but with a different weight. Thus, being the variables far too many to be manageable, it is worth pruning them without loosing important information for clustering.
To this aim, we ordered the whole set of variables, according to their contribution to the principal component used for the partition, and then pruned them starting from the less significant ones. Such a procedure is iterated until a further step would modify the obtained partition.

3 RESULTS

The original observations (subjects) can be graphically depicted in the bidimensional space of the first two principal components, see Figure 2. Each sample is repre-

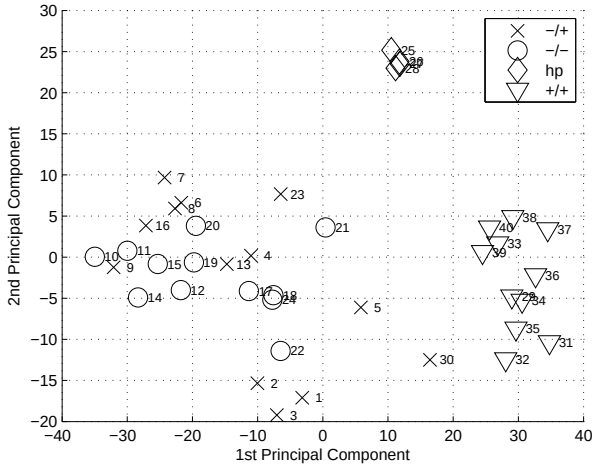

Fig. 2 Subjects-Genes dataset in the first two principal components

sented by one of the four bigger symbols (see the figure legend) each labelling one of the four different original groups. Thanks to the principal component orthonormalization, in Figure 2 it is already possible to visually perceive the distinction of two among the four sets, while the remaining two ones appear like a third mixed set. Figure 3 shows that the first principal component

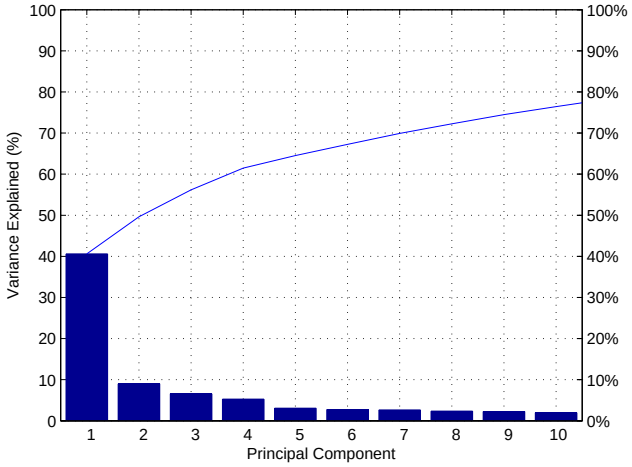

Fig. 3 Weight of the significant principal components of the original dataset

of the whole dataset accounts for about a 40% of the total variability, and is absolutely dominating over the remaining components. In fact, it is possible to observe in Figure 2 that a reasonable bisection of the observations is along the first principal component by means of a cut along the vertical line passing though the origin. In this way, it is already possible to visually appreciate, on the right side, the distinction of the two said subsets, while the mixed set is left to the left. Figure 4 confirms this, even after the k-means optimization of the two clusters. By iterating the principal component orthonormalization, followed by the k-means optimization, on the right side set, a clear separation of the two distinct subsets follows straightforwardly, as depicted in Figure 5. On the contrary, clustering the left side data of Figure 4 is more difficult also because the first principal component of this subset of observations explains only 25% of the variance (Figure 6). In this case, a vertical cut perpendicular to the first principal component is not able to distinguish between the two different conditions. Instead, an horizontal cut perpendicular to the second principal component is able to correctly classify almost all the cases with respect to the original labels (Figure 7). We then refined this partition with the k-means algorithm. The obtained decomposition in the four final clusters is depicted in Figure 8 with re-

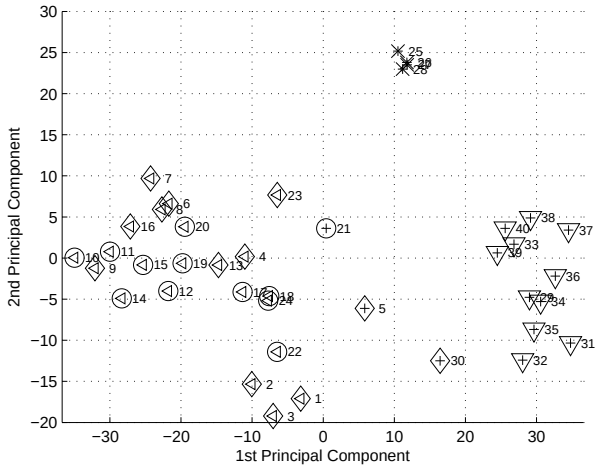

Fig. 4 First level of bisecting clustering of the original dataset - ◁ first cluster + second cluster

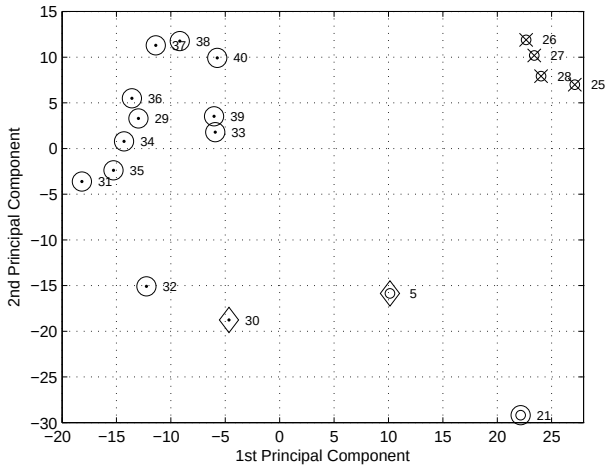

Fig. 5 Second level of bisecting clustering of the right side set of first level - • first cluster ○ second cluster

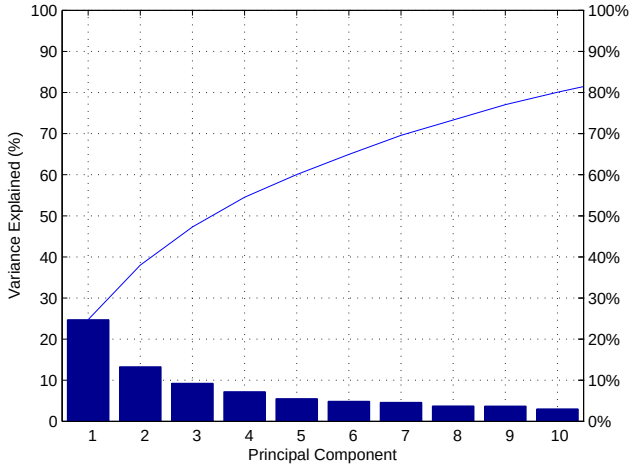

Fig. 6 Weight of the significant principal components of the left side set of first level

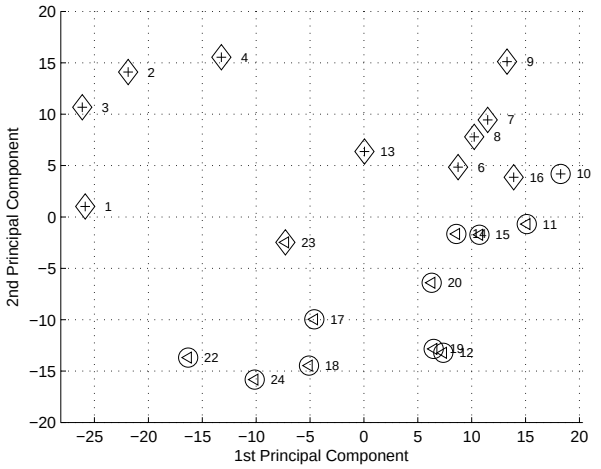

Fig. 7 Second level of bisecting clustering of the left side set of first level - ◁ first cluster + second cluster

spect to the first two principal components of the original dataset. Interestingly enough, in order to obtain the

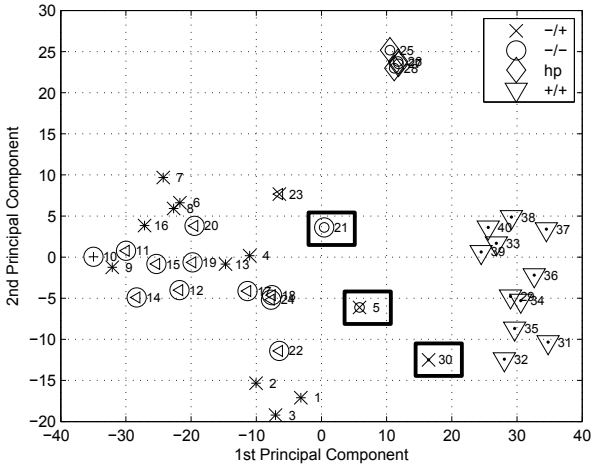

Fig. 8 Data partition in four clusters after the second level of bisecting clustering - ◁ first cluster + second cluster • third cluster ○ fourth cluster - in the bold boxes the three evident outliers

partition of the first bisecting clustering phase depicted in Figure 4, just the eleven genes, listed in Table 1 in decreasing importance order, are needed, among the 542 original ones. To further partition the right side cluster of Figure 4, just the 3 genes listed in Table 3 are needed. Concerning the second level of clustering, the higher degree of similarity between cases in Figure 2 is reflected in a higher number of genes needed to further partition them into the two final clusters (Figure 8). Table 2 lists the 38 needed genes in decreasing significance order out of the 542 original ones.

Gene No.	Gene Name
27	NM000583
30	NM000042
145	NM014576
227	NM000504
36	AF231916
38	NM003963
441	NM006744
25	NM017521
440	NM014255
243	J04422
258	AK024581

Table 1 Most significant genes (in decreasing order) for first level bisecting clustering

Gene No.	Gene Name
536	NM002045
13	AF070524
76	AB028983
411	NM014394
386	NM005573
365	NM006459
185	AL122118
165	AK024475
160	AL050183
140	AK001889
243	J04422
487	AK001109
241	AK022077
127	AL110152
39	AB035130
64	NM000756
137	AK024943
151	NM005654
18	NM000142
217	M26123
23	AF161441
150	AJ297363
41	NM000896
161	NM003878
129	NM007127
219	NM000295
121	NM000790
147	AF130077
191	NM001794
286	NM000390
119	AF273046
154	NM006408
118	NM004063
149	NM005396
201	NM002722
148	BC027895
38	NM003963
288	NM000207

Table 2 Most significant genes (in decreasing order) for second level bisecting clustering for the left side set of first level

Gene No.	Gene Name
149	NM005396
148	BC027895
201	NM002722

Table 3 Most significant genes (in decreasing order) for second level bisecting clustering for the right side set of first level

4 DISCUSSION

Already in Figure 2, as well as in the subsequent Figures 4, 5, 7 and especially 8, as highlightd by the three bold rectangles around subjects 5,21 and 30, some subjects are apparently not belonging to any of the three above mentioned clusters, each grouping either one separate class or the remaining pair of classes together. Such possible outliers should be probably hystologically reconsidered: as already pointed out also in [3], these are typical cases of either errors in getting or transcribing the data - that could always happen even to experienced investigators, due to both the huge amount of data and the demanding exerimental protocol - or peculiar subjects with some additional personal character not allowing to consider them to fully belong to either one of the four classes.

Concerning the tables, it is far beyond the scope of this paper, as well as of the competencies of the two present authors, to discuss the biological meaning of the genes needed, at the two levels of clustering, to reconstruct, thus discriminating, the original four classes of subjects. We will thus just limit ourself to general data science considerations, without even try to infer a possible biological meaning, that would hopefully be the object of a more complex and richer following paper - together with our biological partners, that identified the problem we believe we helped to solve, and kindly provided their data.

As a matter of fact, some of the variables (genes) present in more than one table, needed either for the same second level of clustering or in the subsequent levels, are the same, thus underlying the importance of such genes in discriminating the four investigated subjects' condition. It should thus be of even more interest to have them specifically discussed by biologically competent colleagues.

As a matter of general discussion, care should probably be put to those subsets of important genes recurring in the tables, whose over-expression or/and under-expression is congruent in one or more class(es) of the subjects with respect to one or more other class(es).

In particular, such reduced sets of discriminating genes, should ease the always complex task to determinate not just the pool of the synergic genes, but even, if possible, the pathway linking them, as cared in the paper [10], where details on the experimental protocol can be also found.

Interestingly enough, one of the genes had been, blindly to both of us, identically triplicated, besides a not influent multiplicative factor, in order to investigate the rejection capability of our algorithm to linear combinations: as one could easily understand even from the sole consideration of the algorithm, such robustness is strong, as also experimentally evidenced by the fact that only two of such replications have been considered important by the algorithm, needing at least two entries, no matter of possible different names of the same gene, in order to discriminate the two levels.

Acknowledgements The authors are warmly indebted to Maria Luisa Malosio, providing a subset of data from her colleagues in [10], as well as for valuable discussions. We would welcome her, and also them, to deepen together with us the proposed analysis with their complementary knowledge, also toward the quest of gene pathways not yet completely clarified.

References

1. A. Bosin, N. Dessì, D. Liberati, and B. Pes. *Learning Bayesian Classifiers from Gene-Expression MicroArray Data*, volume 3849 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006.
2. G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205 – 217, 2003.
3. S. Garatti, S. Bittanti, D. Liberati, and A. Maffezzoli. An unsupervised clustering approach for leukaemia classification based on dna micro-arrays data. *Intelligent Data Analysis*, 11(2):175–188, 2007.
4. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.
5. D. Hand, H. Mannila, and P. Smyth. *Principles of Data-Mining*. The MIT press, Cambridge, Massachusetts, USA, 2001.
6. A. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.
7. M. Kantardzic. *DATA MINING: Concept, Models, Methods and Algorithms*. IEEE Press, Wiley, 2001.
8. J. Liu, H. Iba, and M. Ishizuka. Selecting informative genes with parallel genetic algorithms in tissue classification. *Genome Informatics*, 12:14–23, 2001.
9. L. Lu and J. Han. Cancer classification using gene expression data. *Information Systems*, 28:243–268, 2003.
10. E. Missaglia et al. Pancreatic endocrine tumors: expression profiling evidences a role for akt-mtor pathway. *J. Clin. Oncol.*, 28:245 – 255, 2010.
11. B. De Moor, K. Marchal, J. Mathys, and Y. Moreau. Bioinformatics: Organisms from venus, technology from jupiter, algorithms from mars. *European Journal of Control*, vol. 9, no. 2-3:237–278, 2003.
12. M. Muselli and D. Liberati. Binary rule generation via hamming clustering. *IEEE Transactions on Knowledge and Data Engineering*, 14(6):1258 – 1268, 2002.

13. J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465 – 471, 1978.
14. S. Savaresi, M. Boley, and L. Daniel. A comparative analysis on the bisecting k-means and the pddp clustering algorithms. *Intelligent Data Analysis*, 8(4):345–362, 2004.