BOHR Journal of Biocomputing and Nano Technology
2025, Vol. 3, No. 1, pp. 18–22
DOI: 10.54646/bjbnt.2025.17
www.bohrpub.com

RESEARCH

Principal direction devising partitioning initialization of K-means clustering for discriminating among ovarian cancers while identifying the most salient genes involved

Diego Liberati[*†]

National Research Council of Italy, Politecnico Univeristy, Milan, Italy

[*]**Correspondence:**
Diego Liberati,
diego.liberati@cnr.it

[†]**ORCID:**
Diego Liberati
0000-0002-4294-6705

Received: 08 September 2025; **Accepted:** 15 October 2025; **Published:** 18 November 2025

This paper aims to cluster ovarian cancer patients described by gene expression data in order to discover the most discriminating genes responsible for the clustering. A combined use of Principal Direction Divisive Partitioning (PDDP) and bisecting K-means algorithms is applied to the investigated ovarian cancer dataset. The cascading of PDDP and bisecting K-means does successfully cluster ovarian cancer subjects and efficiently discovers salient genes needed to discriminate such clusters. The combined approach worked well on the automatic clustering of ovarian cancer patients depending merely on the gene expression information, and it has great potential for solving similar problems, like classifying leukemias or pancreatic tumors. The saliently identified genes may thus enhance relevant information for discriminating among ovarian cancers. In conclusion, the approach is shown to be a powerful one even in a complex multifactorial case like the intricate ovarian cancer discrimination.

Keywords: salient genes identification, principal component analysis (PCA), K-means clustering, Principal Direction Divisive Partition (PDDP), ovarian cancer

Introduction

In this paper, we focus on a set of deoxyribonucleic acid (DNA) micro-arrays data kindly made available by the National Institute for Cancer in Milan regarding patients suffering from ovarian cancer. In this study an unsupervised clustering has been performed, with the following aims: (1) to select the most significant gene information and reduce the problem's dimensionality; (2) to create a classification of patients affected by cancer; (3) to extract the genes mostly featuring the obtained clustering. The data analysis procedure we propose consists of four steps. The first two are devoted to preliminary data processing, while the third is devoted to data clustering. Finally, we proceed to the selection of the most relevant genes for pathology classification the four steps can be outlined as follows:

1. A drastic reduction in the number of genes dealt with is obtained thanks to a pruning based on the analysis of inter-subject variance. This will enable passing from the original 8,665 genes to only 227.

2. To define a hierarchy in the remaining set of genes, we resort to a principal component analysis (PCA); in this way it is possible to point out the most significant variables for clustering.

3. The adopted clustering procedure follows the unsupervised point of view, without using a priori information on the patient's pathology in order to highlight the unknown patient casuistry. The clustering algorithm is a cascading of the classical K-means technique (1, 2) with the Principal Direction Divisive Partitioning (PDDP) algorithm

proposed in (3), thus significantly improving their performances (4).

4. By processing the obtained results, the number of genes for the discrimination of the various pathologies is further shrunk, yielding to a classification based on reduced number of genes.

Methodology and data

Description of the dataset

The database so obtained is formed by gene expression data over 48 patients affected by ovarian cancer, each characterized by 8,665 genes, thus forming a matrix, each row corresponding to a patient, while columns, denoting human genes, are their descriptive variables. Each subject is then determined by a sequence of 8,665 real numbers, each measuring the expression of the corresponding gene in its pathological cell. By exploiting the gene expression in such a matrix, the data points can be represented as 48 vectors in an 8,665-dimensions Euclidean space. A simple measure of the genomic difference between two patients can be obtained by computing the distance of two vectors. In order to ease algebraic manipulations of data, the dataset is thus represented by means of a real matrix of dimension 48 $\times$ 8,665, the entry of which measures the expression of the jth gene for the ith patient. In order to reduce dimensionality, only the 227 genes whose variance among patients higher than 0.2 are considered, being the lower-variance genes probabily not involved in discrimination.

Each subject is a priori clinically classified by means of the following variables (types): histology, namely the kind of ovarian cancer; Response to treatment: that is the reduction of the mass's cancer after surgical operation and chemotherapy; Response to treatment with platinum and similar, as above, where platinum and similar are two of the most used substances in chemotherapy; Disease-free survival, that means the relapse of disease; Overall survival, that means the time of survival after operation; p53 status, where p53 is a gene often not activated in cancer disease; Age at onset, which shows the age of the patient, assuming the following values (terms):

histology: 1 = undifferentiated; 2 = serous; 3 = endometrial; 4 = clear cells; 5= ovarian metastasis from other tumor; 6 = cell lines; 7 = normal ovarian cells; 8 = immortalized ovarian cells; 9 = borderline.

Response to treatment (with or without Pt and similar): A = treatment sensitive; B = treatment resistant; NA = not available.

Disease-free survival: C = no relapse; D = relapse; NA = not available.

Overall survival: E = survival $<$ 24 months; F = survival $>$ 60 months; NA = not available or in between.

p53 status: G = active gene; H = inactive gene; NA = not available.

Age at onset: L = age $<$ 45 years old; M = age $>$ 45 years old; NA = not available.

Unfortunately for some types of classification, terms of classification are not always available, as shown by "NA," mainly because it is impossible to define a classification for samples of genes grown and analyzed in vitro.

Principal Component Analysis (PCA) (2, 5) allows reducing the dimensionality of the problem from 227 to 10. The final dimension of 10 refers to the number of linear combinations of genes considered as the most significant ones from the PCA analysis.

Clustering

The clustering of the data obtained at the end of the preliminary data processing described above is performed according to the following basic rationale: maximize the intra-similarity within each cluster and minimize the inter-similarity of two different clusters. In practice we pursue this objective as follows:

(a) The set of data is portioned into two subsets (bisecting procedure) by initially applying an algorithm called PDDP and then the popular K-means procedure.

(b) When advisable, one or both clusters obtained as output of step a) are again sub-partitioned into two clusters by re-applying PDDP followed by K-means.

Obviously, the procedure can be iterated until a suitable number of clusters is obtained.

The PDDP algorithm PDDP is a recently proposed non-iterative technique (3). The idea behind PDDP is that when data are aggregated in two main clusters, then, after performing PCA on available data, the principal component is typically oriented as the direction from one cluster center to the other cluster center. Therefore, the partition of the two clusters can be achieved by computing the first principal direction for the available data (note that for the first bisection such components have been already computed in the pre-processing step; this, however, is not true for subsequent bisections, and PCA has to be performed again in these cases) and then by cutting the data set through a hyper-plane orthogonal to the first principal component and passing through the centroid of the data set.

The K-means algorithm

K-means was first introduced in MacQueen (1), and probably, it is the best known and most widely used iterative

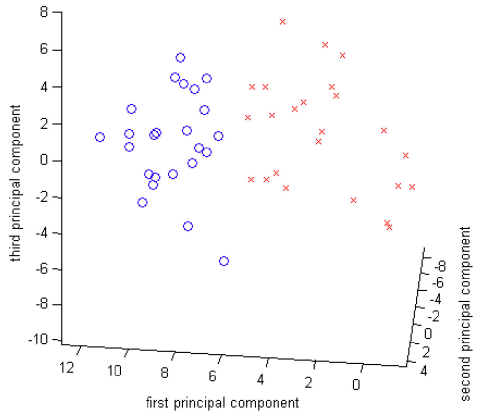

FIGURE 1 | PDDP+K-means data partition ("x" = first cluster R, "o" = second cluster L).

clustering technique. It consists in minimizing the distance among points within the same cluster, while maximising distances across clusters. The number of clusters is fixed at 2 every iteration by the recalled previous PDDP approach.

Results

By applying the above methodology to the ovarian cancer database, the set of 48 subjects has been first subdivided into the two clusters named L and R (happening to be each of the same cardinality of 24 subjects) represented in **Figure 1** (where only the three principal components of higher covariance are shown).

This bisection is basically performed orthogonally to the first principal component, thus simplifying the following "gene shrinking." Interestingly enough, similarities arise between our unsupervised classification form data and a priori clinical classification. The terms of classification showing large differences between clusters L and R over a large percentage of subjects are the more interesting ones.

Some aspects of possible biological interest are:

Concerning histology, we notice that terms 1, 3, and 5 (together including about one-third of subjects) mainly belong to cluster L, while term 2 (even if it would have notable weight) does not provide (like term 4) useful information because of the almost equal distribution of subjects in both clusters L and R. Finally, it has to be mentioned that subjects classified according to terms 6, 7, 8, and 9 fully belong to cluster R: all these terms are linked to tumor cells grown in vitro; thus, it could be assumed that cluster R performs a quite effective characterization of in vitro (as opposed to in vivo) cells. This could imply that, through the different expression of some genes, it would become possible to verify how the in vitro growth involves different biological gene expression and changes the complex partially unknown gene coregulations/pathways with respect to what happens in vivo, thus possibly implying that in vitro studies are not always fully representative of tumor behavior in vivo.

Concerning response to treatment (without or with platinum and taxolo, the results are comparable, as expected), there is a significant predominance of cluster L, mainly tending to include subjects sensible to treatment, while many "NA" mainly fall in cluster R.

Concerning disease-free survival, overall survival, p53 status, and age at onset, we notice that L mainly clusters both opposite terms, while R is again mainly representative of "NA."

The substantive presence of the "NA" in cluster R is congruent with the fact that such NA labeling for classification types other than histology is quite common for in vitro cells.

The final gene-shrinking step leads to a very small number of significant genes needed for such classification, namely the 31 genes listed below:

Gene Accession Number (GAN)

A405891
AA424824
AA429422
AA434342
AA448478
AA459632
AA481944
AA482328
AA486092
AA486471
AA487797
AA490694
AA634006
AA634109
AA664389
AA683520
AA700832
AA968896
AI088704
AI089989
AI206407
AI689992
H08564
H12338
H50623
R14230
R35283
T67053
T70503
T71551
T98612

The corresponding value through which the classification is performed is equal to 2.98.

The existence of a complex bio-molecular network of gene co-regulation and of pathways of gene control may explain the need for just a few core genes to identify the edge of classification, being the further information provided by the other genes not needed any more. After the first bisection, the PDDP + K-means algorithm has been re-applied one more time to both the clusters obtained before in order to obtain the subclusters named LL and LR (each with 12 subjects) for the cluster L, and RL (with 11 subjects) and RR (with 13 subjects) for the cluster R.

Results suggest the following interpretation:

Concerning histology, terms 1 and 3, being together more than half of the considered population, account for almost all the subjects clustered in LL, while term 2, accounting for almost all the residual subjects, well belongs to cluster LR, thus appearing to mainly identify type 2, namely carcinoma ovarico sieroso, the most frequent malignant form in the analyzed context.

Incidentally, no subjects belong to types 6, 7, 8, or 9, since they have already all been classified in cluster R in the previous step.

Concerning response to treatment (without or with platino e taxolo) and disease-free survival, the results are quite interesting from a prognostic point of view; in fact, term A mainly clusters in LR, while term B mainly clusters in LL, helping to blindly forecast prognosis, and overall survival.

p53 status and age at onset do not exhibit, instead, any interesting correlation with clustering.

Gene shrinking extracted a number of 98 genes needed for the sub-clustering of L in LL and LR, much bigger than the only 31 genes needed at previous step classifying in L and R, probably because a larger quantity of information is required to refine classification.

The sub-clustering R in RL and RR shows that:

Concerning histology, terms 1 and 9 cluster in RL, together with most of the subjects belonging to term 2. Types 4–8 belong to cluster RR (types 6–8 feature in vitro cells);

Concerning response to treatment (without or with platinum and similar), disease-free survival, and overall survival (even if an important presence of the terms "NA" has to be taken into account) and p53.

Status, the quite sharp opposition between the high percentage of subjects in cluster RL with response A, as opposed to the high percentage in cluster RR of subjects with response B, is useful in prognosis, while age at onset (whose results are not so well defined) is not.

Gene shrinking reduces to 45 genes the minimal need for clustering R in RL and RR.

A possible further partition of (some of the) four described clusters, which could be of help if more than one type were almost fully clustered together in only one partition, would not help in the present case, the data being already mostly sparse across clusters.

Discussion

A technique of unsupervised learning has been proposed in order to mine ovarian cancer micro-arrays data with the aim to help to develop a tool able to learn only from gene expression data how to classify the observed variability in ovarian cancer.

Such a technique basically consists of re-arranging the data in such a way as to more easily and subsequently prune the redundant information before clustering and finally coming back to the original frame.

Such an approach has already proved successful (6) in a simpler problem in this same context, namely classifying leukemia (7). In that case, better classification with a need of fewer genes was obtained with respect to the original work of Golub and co-workers. The results in the present paper are not equally sharp, due to the higher complexity of the problem, together with the present availability of only a few homogeneous samples for each histological type.

Nevertheless, the proposed approach was able to suggest some partitioning of interest. In fact, taking into account the two subsequent partition steps described in the previous result section, one can notice that in vitro tumors cluster together in RR, while on the other side a fraction of clear cells and secondary forms other tumors also fall, while all borderline alone do constitute cluster RL. For the other types of histology, unfortunately, such a sharp partitioning is not yet available (only a tendency to LL for both not differentiated and endometrial is apparent), nor does the minimal set of genes needed for the partitioning seem to be small enough yet to allow direct biological experimentation. Also for prognostic classification, LR is preferred but is not really as sensitive nor specific as one would like when treatment is useful and survival is improved both in quality and quantity. p53 is slightly more active in LR. Age at the onset does not really appear to matter.

Conclusion

Further investigation with a more extensive and maybe multi-centric data set is thus needed to assess if the proposed approach, already successful in other problems in some way quite similar to the one under analysis, could really be of help, maybe together with other methods, in discriminating ovarian cancer histology with a lower need of biological analysis, or in suggesting prognosis, or in extracting a real core of a few genes whose systems biology could assist in explaining key issues in cancer. In a quite recent paper (8), they investigated a rare (thus few subjects available) form of leukemia. The very same approach helped us to confirm a suspected pathway thanks to the enhancement in statistical power provided to the data by our approach.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The Institute for Cancer Research kindly made the data available. MS students co-supervised at Politecnico di Milano processed them in fulfillment of their master thesis. Participants at conferences, discussing preliminary results partially presented there, were also instrumental in better focusing the present analysis.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J editors. *Proceeedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: Univeristy of California Press (1967). p. 291–7.
2. Hand D, Mannila H, Smyth P. *Principles of Data-Mining*. Cambridge, Massachusetts, USA: The MIT Press (2001).
3. Boley DL. Principal direction divisive partitioning. *Data Min Knowl Discov.* (1998) 2(4):325–44.
4. Savaresi SM, Boley DL. On the performance of bisecting K-means and PDDP. *1st SIAM Conference on Data Mining. Chicago, IL, USA, paper n.* (2001) 5:1–14.
5. O'Connel MJ. Search program for significant variables. *Comp Phys Comm.* (1974) 8:49.
6. Garatti S, Bittanti S, Liberati D, Maffezzoli A. Unsupervised mining of genes classifying leukemia. *Intell Data Anal.* (2007) 11(2): 175–88.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression. *Science.* (1999) 286:531–7.
8. Grassi S, Palumbo S, Mariotti V, Liberati D, Guerrini F, Ciabatti E, et al. The WNT pathway is relevant for the BCR-ABL1-independent resistance in chronic myeloid leukemia. *Front Oncol.* (2019) 9: 532.