Medical Informatics Europe 2002
G Surján et al. (eds)
IOS Press 2002

Cluster Analysis of Wisconsin Breast Cancer Dataset Using Self-Organizing Maps

Stefan PANTAZI, Yuri KAGOLOVSKY, Jochen R. MOEHR

School of Health Information Science, University of Victoria,
V8W 3P5 Victoria, BC, Canada

Abstract. This work deals with multidimensional data analysis, precisely cluster analysis applied to a very well known dataset, the Wisconsin Breast Cancer dataset. After the introduction of the topics of the paper the cluster analysis concept is shortly explained and different methods of cluster analysis are compared. Further, the Kohonen model of self-organizing maps is briefly described together with an example and with explanations of how the cluster analysis can be performed using the maps. After describing the data set and the methodology used for the analysis we present the findings using textual as well as visual descriptions and conclude that the approach is a useful complement for assessing multidimensional data and that this dataset has been overused for automated decision benchmarking purposes, without a thorough analysis of the data it contains.

1. Introduction

Multidimensionality has long been an obstacle in data analysis because of our limitations of coping with more than three spatial dimensions. In order to be visualized, a multidimensional data set must be projected into a lower dimensional space where it will invariably lose some of its features. In this paper we show how both, data visualization and cluster analysis of a dataset, can be achieved by using the Kohonen model of self-organizing artificial neural networks and also how the results of this analysis can be used to compare the performances of classification processes. The data used for analysis was taken from the Wisconsin Breast Cancer dataset, which is freely available on the Internet [1, 2].

2. Cluster Analysis

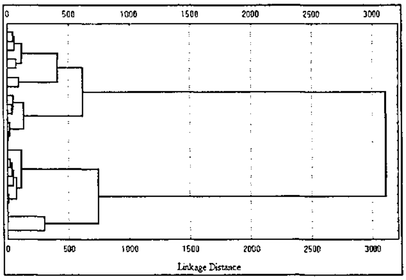

Figure 1: Joining tree (dendrogram) example. There are two very well defined clusters, corresponding to the two branches of the tree spanning over a wide linkage distance range (from approx. 750 to 3000)

Cluster analysis is a method of exploratory data analysis that aims at partitioning a set of data items into groups, based on a measure of distance (or dissimilarity). The groups are called clusters and their number may be pre-assigned (e.g. k-means clustering procedure, which classifies objects into a pre-specified number of clusters by moving objects into different clusters with the goal of minimizing the intra-cluster variability while maximizing the inter-cluster variability) or determined by the algorithms (e.g. joining algorithms which aggregate – amalgamate - increasingly larger clusters of increasingly dissimilar patterns).

As we can see in the Figure 1, clustering is a relative notion. The intra-cluster and inter-cluster distances may vary from data set to data set and because of this, a variable clustering threshold has to be defined for every given problem. In Figure 1, this is equivalent to visually inspecting the amalgamation trees for finding the number of clusters.

Cluster analysis helps to organize observed data into meaningful structures and to develop useful data-driven taxonomies or classifications [3]. The visual depiction of a multidimensional data set provided by self-organizing maps proves useful because it gives a first glimpse of the underlying features of the data and of the distribution of the data points in the multidimensional problem space, helping to identify relations between multidimensional data as well as the number of clusters in the data sets. Unlike joining cluster analysis algorithms, the self-organizing maps do not set any limit on the number of patterns in the data sets.

3. Self-Organizing Maps

Self-organizing maps (feature maps, Kohonen maps) [4] are biologically plausible models of artificial neural networks that provide a very convenient 2-dimensional visual representation of high dimensional input data. They use an unsupervised learning algorithm based on a distance calculation causing each input pattern to be associated with a zone on the resulting map (Figure 2a). The real spatial distances between input patterns are not respected in the sense that two patterns that are close to one another topologically are not necessarily similar from the distance point of view. These patterns will be actually separated by a "valley" (Figure 2a) covering a wide input data space void of input patterns. In contrast, two data points that are close in terms of distance will also be close topologically but there will be a less thick boundary separating them. In this way, the 2-dimensional mapping space becomes a warped projection of the multidimensional input space, with the input patterns squeezed into one another and separated by variable width boundaries.

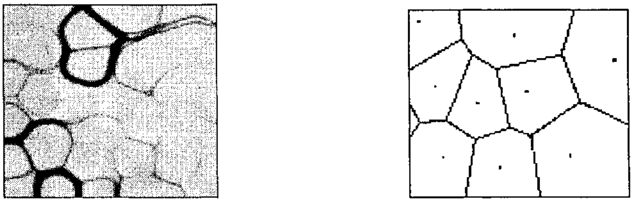

Figure 2: a - Example of self-organizing map. Each data point has a corresponding zone in the map and it is separated form the other points by a "valley" of variable width (e.g. near the top of the picture one cluster made of two patterns is clearly separated from the surrounding patterns by a wide boundary); b - Voronoi diagram. The separating lines do not have width unlike the self-organizing maps where the patterns are separated by boundaries of variable width

Therefore, the boundaries and their thickness will contribute to the description of the homogeneity of a data set helping to perform the cluster analysis. In the terms of classical statistics, the pattern boundaries can be compared to Voronoi diagrams (Figure 2b) [5].

4. Materials

The Wisconsin Breast Cancer dataset was initially created to conduct experiments that were to prove the usefulness of automation of fine needle aspiration cytological diagnosis. It contains 699 instances of cytological analysis of fine needle aspiration from breast tumors. Each case comprises 11 attributes: a case ID, cytology data (normalized, with values in the range 1-10) and a benign/malignant attribute (Table 1). The number of benign instances is 458 (65.52%) and the number of malignant instances is 241 (34.48%). We removed from the data set sixteen instances of cases (14 benign, 2 malignant) with missing values.

Table 1: The data set attributes

#	Attribute	Domain
ID	Sample ID code	integer
A1	Clump Thickness	1 - 10
A2	Uniformity of Cell Size	1 - 10
A3	Uniformity of Cell Shape	1 - 10
A4	Marginal Adhesion	1 - 10
A5	Single Epithelial Cell Size	1 - 10
A6	Bare Nuclei	1 - 10
A7	Bland Chromatin	1 - 10
A8	Normal Nucleoli	1 - 10
A9	Mitoses	1 - 10
A10	Class: 2 for benign, 4 for malignant	2, 4

5. Methodology

Self-organizing maps are computationally intensive applications whose requirements are directly influenced by the input patterns' dimensions as well as by the output map size. The artificial neural network model is a fully connected, 2-layered structure and therefore, for a 9-dimensional input and a 180x180 size output map there will be about $2.9x10^5$ connections. The learning algorithm is a stochastic process that may yield different results for the same problem if the initial conditions are not identical. An 180x180 points two-dimensional self-organizing map was obtained from 9-dimensional patterns taken form the entire data set (i.e. 683 instances), after running the analysis for a few times in order to assure the consistency of the results. The ID and benign/malignant class attributes were not used for the analysis. The information provided by the benign/malignant attribute has been further used for color-coding (Figure 4).

In order to implement a simple classifier based on linear distance calculation, the data set was randomly split into two subsets, one for training and one for testing purposes, each of them containing 341 and 342 patterns, respectively. The self-organizing map used as a classifier which is able to place new patterns in the context of the known, correctly classified patterns by finding the closest match and assuming that the new pattern belongs to the same class as the match.

6. Results

A zone of high homogeneity (i.e. cluster) is shown in the low-right corner of the map and low homogeneity for the rest of the map (Figure 3a). Using a lower threshold, the clustering became more evident (Figure 3b). The information provided by the color-coding (Figure 4), indicate that the benign cases located mostly in the homogeneous cluster, might be differentiated from the malignant instances using only a linear distance based classifier. It also shows that the two classes also exhibit a degree of overlap with each other, as some benign case patterns lie outside the benign case cluster and some malignant case patterns are very close or even inside the benign case cluster.

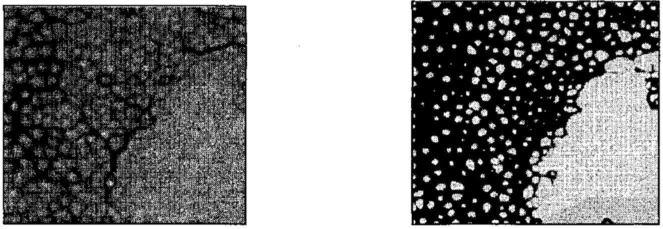

Figure 3: 180x180 points self-organizing maps of the dataset rendered using different clustering thresholds (a – left – higher clustering threshold, b – right – lower clustering threshold)

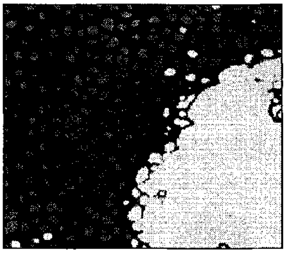

Figure 4: 180x180 points self-organizing maps rendered by color-coding on the benign/malignant class attribute (dark – malignant, light – benign)

For the 342 cases used as a testing subset, the linear distance based classifier accuracy was 95.32%, i.e. 326 patterns were correctly classified.

7. Discussion

The "visualization" of the clusters provides insight into what it will take to automate this particular classification process and also suggests possible explanations for the overlap of the classes. These explanations might be that some training cases are be misclassified (i.e., outliers), that non-linear relationships not captured by an analysis based on linear distance calculation are present in the data, or that a combination of the two possibilities should be considered.

Given the relatively homogeneous clustering of the benign case patterns, the high accuracy of such a simple classifier is no surprise. Other classifiers, employing more sophisticated decision techniques (e.g. fuzzy rules extraction, hybrid systems rule extraction, Bayesian networks) were reporting comparable accuracies in their classifications [6-8] when analyzing the same dataset. Therefore, we consider that, before applying new algorithms or non-linear classifiers capable of capturing the remaining non-linearity that might characterize the data set, it might be of interest to review original cytological material to ensure data accuracy regarding the problem patterns.

Table 2: Three misclassified patterns and their closest match (training pattern)

Pattern	Class	ID	A1	A2	A3	A4	A5	A6	A7	A8	A9
Test pattern 1	M	837480	7	4	4	3	4	10	6	9	1
Test pattern 2	M	1171710	6	5	4	4	3	9	7	8	3
Test pattern 3	M	63375	9	1	2	6	4	10	7	7	2
Closest match (training pattern)	B	1213375	8	4	4	5	4	7	7	8	2

The sixteen misclassified cases, when plotted on the map, showed that most of the patterns were malignant cases (14) having as the closest match a benign case pattern lying either near the border of the benign case cluster (Figure 5) or far from the cluster. Analysis of the data for three of these misclassifications (indicated by the arrow Figure 5 and summarized in Table 2) showed that the three test patterns indeed have as the closest match a training pattern sitting far from the benign cases cluster and this strongly indicates a problem with the training pattern, which needs to be reviewed. As it can be seen in the Figure 5, this case is not singular and actually many of the misclassifications may be explained as being the results of outliers.

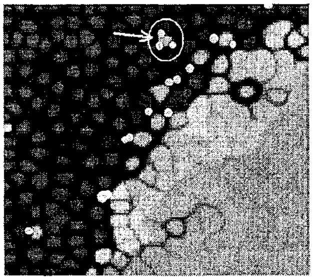

Figure 5: 120x120 points self-organizing maps rendered using the first half of the data set and color-coded (dark – malignant, light – benign); the misclassified patterns are marked by white dots

8. Conclusions

Exploratory data analysis should be the first step that should be taken when analyzing a new data set because beside giving a first impression about the data itself it can also point to the data errors. Although more research about this approach needs to be done (i.e. analyzing other data sets) we are confident that the limitations of these analysis techniques derive only from the computational requirements.

The Wisconsin Breast Cancer dataset should not be used as a benchmark for classification algorithms since any linear distance classifier will probably perform with an accuracy of over 90% and the non-linear classifiers' accuracies will not show notable differences as these differences will derive only from the sparse, potentially erroneous, remaining non-linearity present in the data.

References

[1] Blake CL, Merz CJ. UCI Repository of machine learning databases.
Dept. of Information and Computer Science. Univ. of California, Irvine; 1998. Available:
http://www.ics.uci.edu/~mlearn/MLRepository.html [2002, Jan 15, 2002]

[2] Mangasarian OL, Wolberg WH. Cancer diagnosis via linear programming. SIAM News 1990;23(5):1-18. Available: http://www.cs.wisc.edu/~olvi/uwmp/cancer.html [2002, Jan 15, 2002]

[3] Kaski S. Data Exploration Using Self-Organizing Maps [Doctor of Technology Thesis]. Helsinki: Helsinki University of Technology; 1997.

[4] Kohonen T. Self-organization and associative memory. 2nd ed. New York: Springer-Verlag; 1988.

[5] Haykin SS. Neural networks: a comprehensive foundation. New York: Maxwell Macmillan Intl; 1994.

[6] Sarkar M, Leong T-Y. Application of K-nearest Neighbors Algorithm on Breast Cancer Diagnosis Problem. In: AMIA Symposium 2000. p. 759-763.

[7] Sarkar M, Leong T-Y. Nonparametric Techniques to Extract Fuzzy Rules for Breast Cancer Diagnosis Problem. In: MEDINFO 2001; London: Amsterdam: IOS Press; 2001. p. 1394-1398.

[8] Yen GG, Meesad P. Constructing a Fuzzy Rule-Based System Using the ILFN Network and Genetic Algorithm. International Journal of Neural Systems 2001;11(5):427-443.