ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 277-282

Interpretation and Comparison of Multidimensional Data Partitions

Esa Alhoniemi and Olli Simula

Neural Networks Research Centre
Helsinki University of Technology
P. O. Box 5400
FIN-02015 HUT, Finland
esa.alhoniemi@hut.fi

Abstract. In this paper, a novel visualization method for partitions of multidimensional data is presented. It can be used for characterization of one or comparison of two partitions. The method finds the variables that best describe a partition or difference between partitions. It is especially useful when the data set size is large and there are many variables, i.e., the data dimension is high. The method has been implemented in a software tool prototype which is used in analysis of operational states of a paper machine. For simplicity, however, use of the method is here demonstrated using the well-known Iris data set.

1 Introduction

In several applications, partitions of multidimensional data need to be characterized or compared. Therefore, it is surprising that the topic has received very little attention in the pattern recognition literature.

The partition interpretation problem is commonly faced in inspection of results produced by clustering algorithms. Several methods for validation of the results have been proposed, but the research has mainly concentrated on determination of the number of clusters (see for example [10]), not characterization of the clusters. In [5], some methods for displaying grouping of data can be found, but most of them are useful only when the number of samples is small. Some (interactive) visualizations for hierarchical clustering have also been proposed [2, 6]. However, it should be noted that clustering is by no means the only way to partition data. For example, the partitions may as well be two separate time periods of multidimensional process data which represent normal and test runs of the process.

One possibility to visualize the high-dimensional data partitions is to project them on a two-dimensional display. The most used method is linear Principal Component Analysis (PCA). Also, several non-linear possibilities ex-

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 277-282

ist [3, 8, 9, 12]. Even though all projections provide valuable information on data structure, they cannot be very efficiently used to indicate which variables create the structure. However, among artificial neural networks, the self-organizing map [7] can be used for this purpose.

The method presented in this paper is useful in analysis of partitions when there are many variables and also the number of samples is large. The variables that separate one partition from the others or two partitions from each other are identified based on comparison of variable distributions in different partitions. Then, the variables are ordered according to their ability to discriminate between partitions and shown in a visual display that at one glance roughly shows contents of a partition or partitions.

2 Analysis of Partitions

Let us assume that there are N data vectors of dimension d. The vectors have been divided into C separate partitions $\mathbf{X}_i,\ i = 1, \ldots, C$ so that each vector belongs to exactly one partition. As the contents of the partitions are usually unknown, the following two things are typically of interest.

- How can a single partition be characterized, i.e., which variables best describe each partition and how?
- What are the differences and similarities between partitions?

Our method is based on the observation that the distributions of variables that best describe a partition are maximally different from the distributions of the same variables in the union of all the other partitions. Comparison of two partitions is essentially the same problem except that the first partition is compared with the second one. The problem here is definition of "maximally different", which is by no means unique. The most commonly used measures of difference between variables are the statistics used in χ^2 test for discrete variables and in Kolmogorov-Smirnov (K-S) test for continuous variables [11, 13], but any other similar measures could be used as well. However, from now on we concentrate on continuous variables and the K-S statistic.

In K-S test, there are two series of observations: $u_1, \ldots, u_{N_1}$ and $v_1, \ldots, v_{N_2}$, which are sorted in ascending order; u_0 and v_0 are set to $-\infty$. Comparison of the observation distributions is carried out using cumulative distributions denoted here by $F(x)$ and $G(x)$:

$$F(x) = i/N_1, \quad u_{i-1} < x \leq u_i, \qquad G(x) = i/N_2, \quad v_{i-1} < x \leq v_i. \tag{1}$$

The K-S statistic, i.e., the difference between $F(x)$ and $G(x)$ is computed by

$$D = \max_{-\infty < x < \infty} |F(x) - G(x)|. \tag{2}$$

In the experiment in Section 3, the implementation presented in [11] was used.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 277-282

Characterization of a single partition. Let us denote the partition of interest by $\mathbf{X}_c$. First, the K-S statistic D_k for each variable $k = 1, \ldots, d$ is computed by comparing the variable distributions in the partition $\mathbf{X}_c$ with the corresponding distributions in $\mathbf{X}_{\text{others}} = \bigcup_{i,i \neq c} \mathbf{X}_i$. The K-S statistic D_k describes the importance of variable k: the greater it is, the more important variable k is in characterization of partition $\mathbf{X}_c$.

Comparison of two partitions. Also in comparison of two partitions $\mathbf{X}_{c_1}$ and $\mathbf{X}_{c_2}$, the K-S statistics D_k are computed for each variable $k = 1, \ldots, d$, but now by comparing $\mathbf{X}_{c_1}$ with $\mathbf{X}_{c_2}$. Now the value of D_k reflects the ability of variable k to separate between partitions $\mathbf{X}_{c_1}$ and $\mathbf{X}_{c_2}$.

3 Experiment

The well-known simple Iris data set [1] was used to demonstrate the proposed approach to visualization of partitions[1]. The data set contains 150 data vectors of three different Iris species: Setosa, Versicolor, and Virginica. There are 50 instances of each, and all classes are known. All vectors contain four measurements: sepal length, sepal width, petal length, and petal width.

In Fig. 1, a PCA plot of the data set is shown. Based on the visualization, it seems that the Iris Setosa is clearly separate from the two other species.

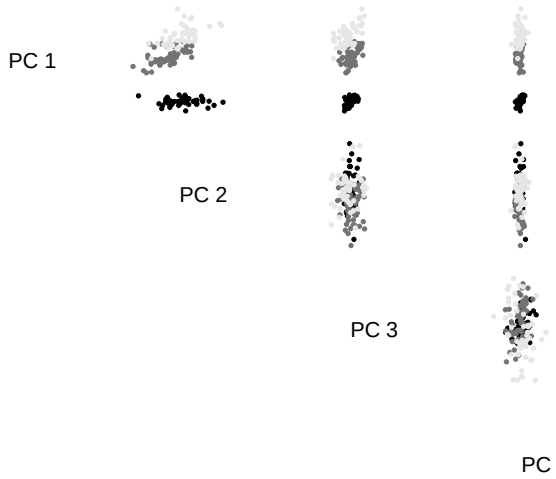


Figure 1: PCA plot of the Iris species. Iris Setosa has been plotted using black, Iris Versicolor using dark gray, and Iris Virginica using light gray color.

In Fig. 2, a partition display of Iris Virginica is shown. The variables have been ordered from top to bottom in such a way that the ones that best describe the partition are at top. The left panel contains the variable names. In the right

[1]The Iris data set is available at UCI Machine Learning Repository, `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/`.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 277-282

panel, data points describing Iris Virginica are plotted using black color, and all the other points (describing Iris Setosa and Iris Versicolor) using gray color. In addition, minimum and maximum values are shown around each of these plots. This illustration immediately shows a general view, a "thumbnail" of the partition and how it is related to the whole data set: the variables that best separate the species from the two others are petal length and width, which are large in this partition. The weakest separation is obtained using sepal width. This can be verified from Fig. 3, where histograms of the variables petal length and sepal width are shown.

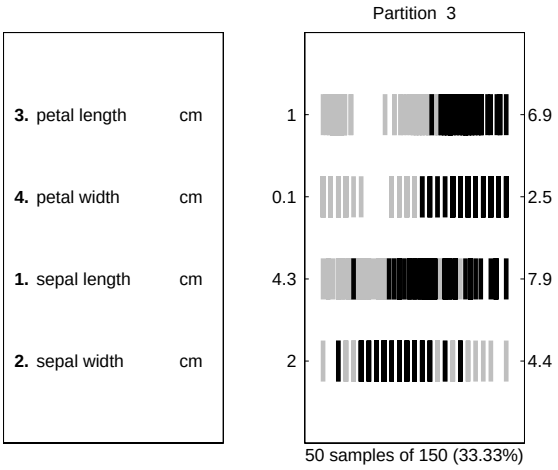
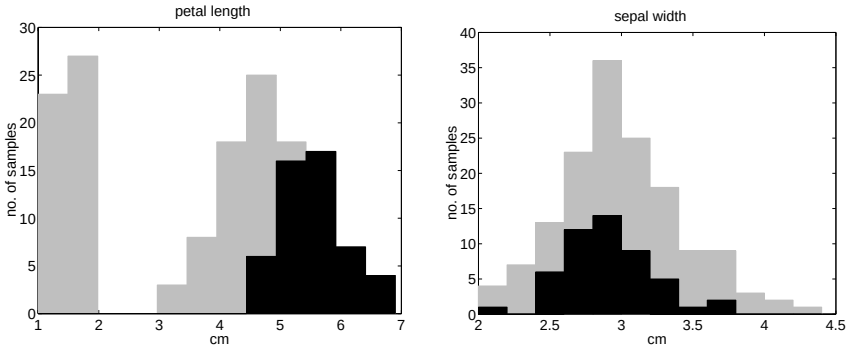


Figure 2: Display of partition 3 (Iris Virginica). The data points that belong to the partition are plotted using black and other points using gray color.



(a) (b)

Figure 3: Histograms of variables (a) petal length and (b) sepal width in partition 3 (Iris Virginica). The data points that belong to the partition are plotted using black and other points using gray color.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 277-282

If Fig. 4, partitions corresponding to Iris Versicolor and Iris Virginica are compared. Now the variables have been ordered from top to bottom according to their importance in partition separation. In Fig. 1 it can be noted that these partitions are not completely separate, but merely seem to overlap[2]. Also in this case petal width and length are the variables that best separate the species, but as it can be observed in Fig. 4, the separation is not perfect.

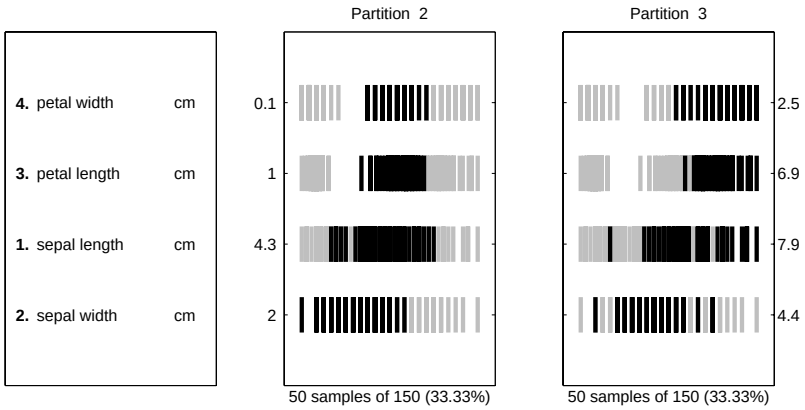


Figure 4: Display of comparison between partitions 2 and 3 (Iris Versicolor and Iris Virginica). The data points that belong to partition 2 (middle panel) and 3 (right panel) are plotted using black and other data points using gray color.

4 Conclusions

In this paper, a novel visualization for partitions of multidimensional data is presented. It can be used for two different purposes: to characterize a single partition and to compare two partitions. Statistics adopted from the K-S test or χ^2 test are used to order the variables in such a way that the most important ones are shown first in the visualization. The method is useful especially when the set of data points is large and the data dimension is high.

The proposed method has been successfully used in analysis and optimization of paper machine control procedures [4] where the data dimension may even be dozens and the number of samples is typically several thousands. The method has been used in interpretation of clustering results as well as comparison of data from two different time periods. In both cases, it has proven to be a valuable tool that improves characterization and comparison of partitions.

[2]It is a known fact that these two species are not linearly separable.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 277-282

References

[1] Edgar Anderson. The Irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59(2–5), 1935.

[2] Eric Boudaillier and Georges Hebrail. Interactive Interpretation of Hierarchical Clustering. *Intelligent Data Analysis*, 2(3), 1998.

[3] P. Demartines and J. Hérault. Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets. *IEEE Trans. on Neural Networks*, 8(1):148–154, January 1997.

[4] Maija Federley, Esa Alhoniemi, Mika Laitila, Mika Suojärvi, and Risto Ritala. State management for process monitoring, diagnostics and optimization. In *Control Systems 2000 Preprint*, pages 295–298, 2000.

[5] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[6] Sung-Soo Kim, Sunhee Kwon, and Dianne Cook. Interactive Visualization of Hierarchical Clusters using MDS and MST. *Metrika*, 51(1):39–51, 2000.

[7] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995.

[8] J. B. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, 29(1):1–27, March 1964.

[9] R. C. T. Lee, J. R. Slagle, and H. Blum. A Triangulation Method for the Sequential Mapping of Points from N-Space to Two-Space. *IEEE Transactions on Computers*, C-26(3):288–292, March 1977.

[10] Glenn W. Milligan and Martha C. Cooper. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2):159–179, June 1985.

[11] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C*. Cambridgbe University Press, 2nd edition, 1997.

[12] John W. Sammon, Jr. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18(5):401–409, May 1969.

[13] Alan Stuart and J. Keith Ord. *Kendall's Advanced Theory of Statistics*, volume 2. Edward Arnold, 5th edition, 1991.