Feature Extraction For Agglomerative Clustering

V. Menaka, J. Mary Dallfin Bruxella
Assistant Professors, Department of Computer Science, KSR college of arts and science, Tiruchengode

Abstract

Clustering is the division of data into groups of similar objects. The main objective of this unsupervised learning technique is to find a natural grouping or meaningful partition by using a distance or similarity function. Clustering techniques are applied in pattern classification schemes, bioinformatics, data mining, web mining, biometrics, document processing, remote sensed data analysis, biomedical data analysis, etc., in which the data size is very large. The medical data statistical analysis often requires the using of some special techniques, because of the particularities of these data. The principal components analysis and the data clustering are two statistical methods for data mining which are very useful in the medical field, the first one as a method to decrease the number of studied parameters, and the second one as a method to analyze the connections between diagnosis and the data about the patient's condition. This work investigates the implications obtained from a specific data analysis technique that is the data clustering preceded by a selection of the most relevant parameters using the principal components analysis.

1. Introduction

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [5].

A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [4]. The users can show this with a simple graphical example:

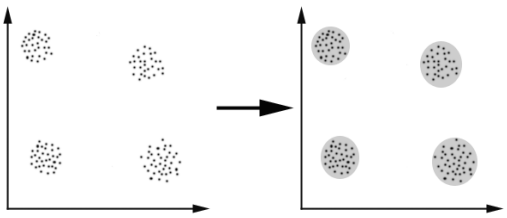

Figure 1: Clusters

In this case the user can easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called *distance-based clustering*. Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Distance measure

An important step in any clustering is to select a distance measure, which will determine how the *similarity* of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another. For example, in a 2-dimensional space, the distance between the point (x=1, y=0) and the origin (x=0, y=0) is always 1 according to the usual norms, but the distance between the point (x=1, y=1) and the origin can be 2, or 1 if the users take respectively the 1-norm, 2-norm or infinity-norm distance.

2. Types of Clustering

There are two main approaches to clustering – hierarchical clustering and partitioning clustering. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters[4].

Another important distinction is whether the clustering uses symmetric or asymmetric distances. A property of Euclidean space is that distances are symmetric (the distance from object *A* to *B* is the same as the distance from *B* to *A*). Besides, clustering algorithms differ among themselves in their ability to handle different types of attributes, numeric and categorical, in accuracy of clustering, and in their ability to handle disk-resident data.

Agglomerative hierarchical clustering

In Agglomerative Clustering, each object is initially placed into its own group. Therefore, if the users have N objects to cluster, they start with N groups. Each of these groups contains only a single object, and is known as a *singleton*. Before start the clustering, the user needs to decide on a threshold distance. Once this is done, the procedure is as follows:

1. Compare all pairs of groups and mark the pair that is closest.
2. The distance between this closest pair of groups is compared to the threshold value.
   - If the distance between this closest pair is less than the threshold distance, these groups become *linked* and are merged into a single group. Return to Step 1 to continue the clustering.
   - If the distance between the closest pair is greater than the threshold, the clustering stops.

If the threshold value is too small, there will still be many groups present at the end, and many of them will be singletons. Conversely, if the threshold is too large, objects that are not very similar may end up in the same cluster.

To run an agglomerative clustering, the users need to decide upon a method of measuring the distance between two objects. In addition, they need a measure to determine which groups should be linked. Some options are simple linkage, average linkage, complete linkage, and Wards method [8].

Dendogram

The dendrogram is a graphical representation of the results of hierarchical cluster analysis. This is a tree-like plot where each step of hierarchical clustering is represented as a fusion of two branches of the tree into a single one. The branches represent clusters obtained on each step of hierarchical clustering.

Partitioning clustering

Partitioning algorithms construct partitions of a database of n objects into a set of k clusters. The construction involves determining the optimal partition with respect to an objective function. There are approximately kn/k! ways of partitioning a set of n data points into k subsets. An exhaustive enumeration method that can find the global optimal partition, is practically infeasible except when n and k are very small. The partitioning clustering algorithm usually adopts the Iterative Optimization paradigm. It starts with an initial partition and uses an iterative control strategy. It tries swapping data points to see if such a swapping improves the quality of clustering. When swapping does not yield any improvements in clustering, it finds a locally optimal partition. This quality of clustering is very sensitive to the initially selected partition. There are the two main categories of partitioning algorithms. They are

i. K-Means algorithms, where each cluster is represented by the center of gravity of the cluster.
ii. K-Medoid algorithms, where each cluster is represented by one of the objects of the cluster located near the center.

K-means Algorithm

The *K*-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

K-Medoid Algorithm

One of the data points in the cluster is considered as the medoid (so it is the most representative datapoint). The way the most representative data point is chosen may vary, but the most reasonable idea (that is resistant to outliers) is to pick the point that has the lowest cumulative distance to all other points. Since this is a costly operation, sometimes it is done only on a sample of the points in the cluster.

3. Principal Component Analysis

The main objective of this work is to analyze the principal components of the input variables. It is especially valuable when we have subsets of measurements that are measured on the same scale and are highly correlated. In that case it provides a few variables that are weighted combinations of the original variables that retain the explanatory power of the full original set.

In order to cluster the medical data efficiently the following steps are to be performed:

a. The principal components analysis must be performed to select the most relevant parameters with a minimal loss of information.
b. Cluster only those parameters using single linkage Agglomerative algorithm.

Taking into consideration the specificity of these techniques, the final result will be an improving of the clustering accuracy, and this thesis intended to check this supposition in a practical situation. The obtained results are presented in the following sections.

The Principal Components Analysis Method

This method of data analysis concerns the finding of the best way to represent n samples by using vectors with p variables, in such a manner so the similar samples are represented by points as close as possible. In order to find the principal components from a set of

variables, the method used is the analysis of eigenvalues and eigenvectors, which starts from a data representation using a symmetrical matrix and transforms it.

The Data Clustering

The general algorithm of hierarchical clustering has the purpose to built a chain of partitions based on the set of input data and an ultra-metric distance, in such a way so, at each step, the diameters of the classes are growing.

This general algorithm can be used in different variants, according with the formula of the ultra-metric distance used; one of these variants is the so-called Average Linkage / Group Average Algorithm.

The algorithm leads to good results when the data are separated in distinct and compact classes; its performance level highly depends on the initial partition and the number of classes to generate. The main elements which decrease its efficiency are :

- The problem of clusters validity: it is necessary to establish by experiments the optimal number of classes into a given dataset, excepting the situations when this number is pre-defined.

- The pseudo-gravitational effect: when the classes have very different sizes, the criteria function tends to favor the partitions that broke the bigger classes against the partitions that keep that classes united; in this way the points situated in border positions are often misclassified.

The typical procedure for this analysis is detailed in the figure below. The initial stage involves the reduction of the dimensionality of the data by principal components analysis (PCA). PCA is a well known technique for reducing the dimensionality of data. Discriminant function analysis (DFA) is then used to discriminate between groups on the basis of the retained principal components (PCs) and the a priori knowledge of which spectra are replicates. This process does not bias the analysis in any way. DFA is not performed on the original feature space because one can not feed co-linear variables or too many variables into DFA. The starting point for DFA is the inverse of the pooled variance-covariance matrix within a priori groups. This inverse can only exist when the matrix is non-singular, i.e., its determinant is other than zero, which implies that it is of full rank.

Singularity can be caused by collinearity, and PCA removes collinearities whilst also reducing the number of inputs to the DFA algorithm. Finally, the Euclidean distance between a priori group centres in DFA space is used to construct a similarity measure, with the Gower similarity coefficient S and these distance measures were then processed by an agglomerative clustering algorithm to construct a dendrogram. The process: Large data → PCA → DFA → HCA .

Architecture of Agglomerative clustering using PCA

PCA transforms the original set of variables to a new set of uncorrelated variables called PCs. PCA is a data reduction process and the first few PCs will typically account for >95% variance. DFA has a priori information based on spectral replicates and uses this to minimize within group variance and maximize between group variance. A similarity matrix can be constructed from the DFA space. HCA can then use this to produce a dendogram, using single linkage clustering.

4. Experimental Analysis

Experimental Analysis is intended to be of use to researchers from all fields who want to study algorithms experimentally. It has two goals: first, to provide a useful guide to new experimentalists about how such work can best be performed and written up, and second, to challenge current researchers to think about whether their own work might be improved from a scientific point of view. Efficient implementations make it easier to support claims of practicality and competitiveness. Faster implementations allow one to perform experiments on more and/or larger instances or to finish the study more quickly.

Secondary protein data structure

In data mining one often encounters situations where there are a large number of variables in the database. In such situations it is very likely that the subsets of variables are highly correlated with each other. The accuracy and reliability of a classification are prediction model will suffer if we include highly correlated variables or variables that are unrelated to the outcome of interest because of over fitting. In model development also superfluous variables can increase cost due to collection and processing of these variables. The dimensionality of a model is the number of the independent or input variables used by the model. One of the key steps in data mining is therefore finding ways to reduce dimensionality without sacrificing accuracy.

In this research, Secondary Protein Data Structure are taken as a sample data to reduce the dimensionality and cluster those measurements by applying agglomerative clustering.

Proteins make up one of the four groups of macromolecules. There are four structures in a protein. The primary structure of a protein is its amino acid sequence. This sequence can range anywhere from around twenty to more than forty thousand amino acids. The secondary structure of a protein consists of how these amino acids fold, for example alpha helixes and beta sheets. The tertiary structure of a protein consists of sub units and how they are arranged in three-dimensional space. The quaternary structure of a protein consists of multiple subunits. The secondary,

tertiary and quaternary structures of protein are all determined by the primary structure. An important task in biology is to predict the secondary structure of the protein. Protein structure determines function and proteins basically control everything about living organisms. Therefore if every structure of a protein can be predicted simply by the amino acid sequence then the function can be known and this gives rise to what information is encoded in the genome. This will allow scientists to create better treatments for certain protein deficiencies and genetic disorders.

PCA with Agglomerative clustering have been used in order to classify proteins into families. The following table shows the sample data for amino acid families with the (GLY,THR,VAL, etc) are the 3 char codes for the amino acids.

Table 1: The Secondary Structure of Amino Acids

0.583459	..	..	0.3531	GLY
0.190529	..	..	0.799772	VAL
0.015414	..	..	0.63379	THR
0.022533	..	..	0.618395	GLY
0.177807	..	..	0.81956	GLY
0.559249	..	..	0.067452	VAL
0.401447	..	..	0.823502	THR
0.785853	..	..	0.382277	THR
0.169737	..	..	0.209338	THR
0.833929	..	..	0.175505	THR
0.123778	..	..	0.517261	THR
0.823983	..	..	0.355042	GLY
0.790152	..	..	0.846917	THR
0.190626	..	..	0.157966	VAL
0.039714	..	..	0.742466	GLY
0.917679	..	..	0.887824	VAL
0.572234	..	..	0.882561	GLY
0.34013	..	..	0.778362	THR
0.239943	..	..	0.476678	THR
0.258705	..	..	0.262235	GLY

These families depend upon the amino acid sequence of the protein, which in turn describe their function. The results are then analyzed from a plotted graph that is produced by showing the clusters. The above said three amino acid families in a database can be shown as pie graph.

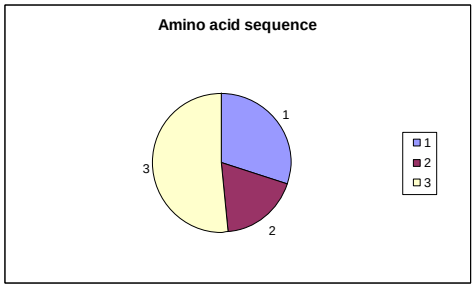

Figure 2. Three Amino acid families

There are many protein databases available. These databases allow the user to enter a sequence and then similar sequences from proteins are returned. Agglomerative clustering are used to cluster the protein sequences according to similarities in the amino acid sequence. The sequences cannot be compared to directly because the lengths of the protein sequences vary. Therefore two proteins can be very similar in pattern but be of different sizes.

The Secondary structure of Globular Protein data base has been taken from the UCI Machine Learning Repository Data base. This is a data set used in this study to predict the secondary structure of certain globular proteins. The idea is to take a linear sequence of amino acids and to predict, for each of these amino acids, what secondary structure it is a part of within the protein. There are three choices:

alpha-helix
beta-sheet
random-coil.

The same procedure which implemented for the head measurements are applied for the protein secondary structure data set.

Conclusion and Future Work

The principal component analysis is a statistical method well known as very efficient for dimension reduction being given a number of parameters, this technique helps to select the most relevant parameters with a minimal loss information. Taking in consideration the specificity of these techniques , this study assumed that , selecting the most relevant medical parameters by a principal component analysis and then clustering only those parameters.

The final result will be an improving of the clustering accuracy. In this work, the user focused on PCA for Dimensionality Reduction. It was observed that for all data sets, the use of PCA resulted in a higher accuracy. It was also observed that PCA is more effective for severe dimensionality reduction. Each seems to use a different approach to perform a low dimensional faithful representation of a high dimensional data.

This work experimented for the data sets with 15 attributes. The experiment captures most of the variability in the data, it seems the output was shown in dendrogram with less than half the number of original dimensions in the data.

Directions for future work include considering other types of high-dimensional data to gain a further understanding of the type of data for which each of the two dimensionality reduction techniques is best suited, as well as considering other dimensionality reduction techniques. In future this result can be compared with the other techniques for reducing dimensionality of data to be used by a nearest neighbor classifier on certain image data sets and certain number of micro array data sets.

References

[1] Chernick, M.R., Friis, R.H., "Introductory Biostatistics for the Health Sciences", John Wiley & Sons Publ., 2003.
[2] Zhou, X.H., Obuchowski, N.A., McClish, D.K., "Statistical Methods in Diagnostic Medicine", John Wiley & Sons Publ., 2002.
[3] C. Dascălu, Boiculese, L., "The Usefulness of Algorithms Based on Clustering in the Diagnosis Finding in Medical Practice", in Lecture Notes of the ICB Seminars - Statistics and Clinical Practice, editors: L. Bobrowski, J. Doroszewski, E. Marubini, N. Victor, Warsaw, 2000, pg. 53 – 56.
[4] Alsabti, K., Ranka, S., Singh, V., "An Efficient K-Means Clustering Algorithm", in Proceedings of the 1st Workshop on High-Performance Data Mining, 1998.
[5] David Hand, Heikki Mannila, Padhraic Smyth – "Principles of Data Mining" Prentice-Hall of India Private Limited, New Delhi, 2001.
[6] T. Cox and M. Cox. "Multidimensional Scaling", Chapman and Hall, London, 1994.
[7] C.J.C. Burges, "Geometric Methods for Feature Extraction and Dimensional Reduction", Kluwer Academic Publishers, 2005.
[8] C.J.C. Burges, "Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners, " Kluwer Academic Publishers, 2005.
[9] Sholom M.Weiss , Nitin Indurkhya, " Predictive Data Mining", Morgan Kaufman Publishers , 1998.
[10] Jackson, J. E., A, "User's Guide to Principal Components", John Wiley and Sons, Inc., 1991, p. 592.
[11] Hyvarinen A, Oja E.,"Independent component analysis: algorithms and applications",A paper in Neural Network Journal, 2000; 13: 411-30.
[12] J. B. Tenenbaum, V. de Silva, and J. C. Langford. ,"A global geometric framework for nonlinear dimensionality reduction". Science, 2000
[13] S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." Science, 2000.
[14] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms", IEEE Transactions on Evolutionary Computation, 4(2): 164 - 171, 2000.
[15] Miguel A. Carreira-Perpinan, "A review of dimension reduction techniques", technical report CS-96-09, Dept. Of Computer Science, University of Sheffield
[16] Imola K. Fodor," A Survey of Dimension Reduction Techniques", LLNL technical report, June 2002
[17] Jonathon Shlens, "A Tutorial on Principal Component Analysis", Systems Neurobiology Laboratory, University of California, (Dated: December 10, 2005; Version 2)
[18] Fahim A.M, Salem A.M., Torkey F.A., Ramadan M.A "An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775 (Online) 2006.
[19] Xiang Chen1, 3 and Robert F. Murphy "Objective Clustering of Proteins Based on Subcellular Location Patterns" Journal of Biomedicine and Biotechnology • 2005:2 (2005) 87–95 • DOI: 10.1155/JBB.2005.87.
[20] D.P. Berrar, W. Dubitzky, M. Granzow "Singular value decomposition and principal component analysis".A Practical Approach to Microarray Data Analysis. 2004.