Impact of Feature Selection Methods in Hierarchical Clustering Technique: A Review

Prof. Mrs. J.R.Prasad, Prof. R.S.Prasad, Dr. U.V.Kulkarni

Abstract
The issues in clustering, the unsupervised classification of patterns, have been addressed in many contexts and by researchers in many disciplines. Its broad appeal and application as one of the steps in exploratory data analysis is well-known. Literature on the pattern clustering methods, algorithms is surveyed. The authors have tried to explore the performance of hierarchical clustering method in this paper.

Index Terms
Cluster, Dissimilarity, Feature Extraction, Pattern representation.

I. INTRODUCTION.

Data analysis procedures can be classified as either exploratory or confirmatory, based on the availability of appropriate models for the data source, but a key element in both types of procedures (whether for hypothesis formation or decision-making) is the grouping, or classification of measurements based on either (i) goodness-of-fit to a postulated model, or (ii) natural groupings (clustering) revealed through analysis. Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity.

Mrs. J.R.Prasad is working as an Assistant Professor at the Department of Computer Engineering, VIIT, Pune. She is a member of IAENG: No. 62346

Prof. R.S.Prasad is Assistant Professor and Head of the Department of Computer Engineering, VIIT, Pune. He is a member of IAENG: No. 62347

Dr. U.V.Kulkarni is a Professor at Department of Computer Engineering and Dean, Academics at SGGS, Nanded.

Steps of a Clustering Task

Typical pattern clustering activity involves the following steps [1]:
(1)Pattern representation (optionally including feature extraction and/or selection),
(2)Definition of a pattern proximity measure appropriate to the data domain,
(3) Clustering or grouping,
(4) Data abstraction (if needed), and
(5) Assessment of output (if needed).

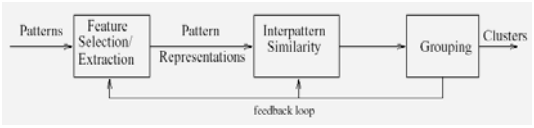

Fig. 1: Steps of a clustering task

II. ROLE OF PATTERN REPRESENTATION, FEATURE SELECTION AND EXTRACTION.

There are no theoretical guidelines that suggest the appropriate patterns and features to use in a specific situation. Indeed, the pattern generation process is often not directly controllable; the user's role in the pattern representation process is to gather facts and conjectures about the data, optionally perform feature selection and extraction, and design the subsequent elements of the clustering system. Because of the difficulties surrounding pattern representation, it is conveniently assumed that the pattern representation is available prior to clustering. Nonetheless, a careful investigation of the available features and any available transformations (even simple ones) can yield significantly improved clustering results. A good pattern representation can often yield a simple and easily understood clustering; a poor pattern representation may yield a complex clustering whose true structure is difficult or impossible to discern.
The features can be subdivided into the following types [2]:
(1) Quantitative features: e.g.

(a) Continuous values (e.g., weight);
(b) Discrete values (e.g., the number of Computers);
(c) Interval values (e.g., the duration of an Event).
(2) Qualitative features:
(a) Nominal or unordered (e.g., color);
(b) Ordinal (e.g., military rank or qualitative evaluations of temperature ("cool" or "hot") or sound intensity ("quiet" or "loud")).

Quantitative features can be measured on a ratio scale (with a meaningful reference value, such as temperature), or on nominal or ordinal scales. One can also use structured features [3] which are represented as trees, where the parent node represents a generalization of its child nodes. For example, a parent node "vehicle" may be a generalization of children labeled "cars," "buses," "trucks," and "motorcycles." Further, the node "cars" could be a generalization of cars of the type "Toyota," "Ford," "Benz," etc. A generalized representation of patterns, called *symbolic objects* was proposed in [4]. Symbolic objects are defined by a logical conjunction of events. These events link values and features in which the features can take one or more values and all the objects need not be defined on the same set of features.

It is often valuable to isolate only the most descriptive and discriminatory features in the input set, and utilize those features exclusively in subsequent analysis. Feature selection techniques identify a subset of the existing features for subsequent use, while feature extraction techniques compute new features from the original set. In either case, the goal is to improve classification performance and/or computational efficiency. Feature selection is a well-explored topic in statistical pattern recognition [5]; however, in a clustering context (i.e., lacking class labels for patterns), the feature selection process is of necessity ad hoc, and might involve a trial-and-error process where various subsets of features are selected, the resulting patterns clustered, and the output evaluated using a validity index. In contrast, some of the popular feature extraction processes (e.g., principal components analysis [6]) do not depend on labeled data and can be used directly. Reduction of the number of features has an additional benefit, namely the ability to produce output that can be visually inspected by a human.

This paper investigates the role of pattern representation, feature selection and the feature extraction processes on the performance of the Hierarchical clustering algorithm.

III. HIERARCHICAL AGGLOMERATIVE CLUSTERING ALGORITHM

(1) Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.
(2) Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.
(3) If all patterns are in one cluster, stop. Otherwise, go to step 2.

Hierarchical divisive algorithms start with a single cluster of all the given objects and keep splitting the clusters based on some criterion to obtain a partition of singleton clusters.

The most challenging step in clustering is feature extraction or pattern representation. Pattern recognition researchers conveniently avoid this step by assuming that the pattern representations are available as input to the clustering algorithm. In small size data sets, pattern representations can be obtained based on previous experience of the user with the problem. However, in the case of large data sets, it is difficult for the user to keep track of the importance of each feature in clustering. A solution is to make as many measurements on the patterns as possible and use them in pattern representation. But it is not possible to use a large collection of measurements directly in clustering because of computational costs. So several feature extraction/selection approaches have been designed to obtain linear or nonlinear combinations of these measurements which can be used to represent patterns. To illustrate this we present the following example.

Example

Consider 8 different objects belonging to the same category e.g. movie category, Lets choose the movies as follows:

The eight movies are:

1. To Live
2. The Shawshank Redemption
3. Top Gun
4. Four Weddings and a Funeral
5. Mulan
6. Lion King
7. Eighteen Springs
8. Brave heart

Let us create an 8 x 8 proximity matrix in which each row and column represents a data object (e.g., name of a movie) using the method of paired comparisons. Enter a number from 0 to 10 in position (i,j) that reflects our opinion of how similar objects i and j are for j > i. Assume the matrix is symmetric and put zeros on the main diagonal. Let 0 mean "identical" and let 10 mean "nothing in common." Let us not try to establish objective criteria - just use our gut feelings. However, let us try to be consistent in our judgments.

The dissimilarity matrix D is as shown below in Fig. 2.

$$\begin{pmatrix} 0 & 5 & 9 & 7 & 9 & 8 & 4 & 3 \\ & 0 & 9 & 6 & 10 & 3 & 9 & 5 \\ & & 0 & 4 & 5 & 5 & 9 & 8 \\ & & & 0 & 3 & 4 & 2 & 9 \\ & & & & 0 & 4 & 9 & 5 \\ & & & & & 0 & 8 & 7 \\ & & & & & & 0 & 8 \\ & & & & & & & 0 \end{pmatrix}$$

Fig. 2: Dissimilarity matrix D

Now a final dissimilarity matrix by breaking all ties in the matrix of part (a) is as follows. To break a tie, we have considered the set of pairs having the same dissimilarity. We have added or subtracted small amounts until no two dissimilarities are the same and then ranked the order the results, so the proximity matrix contains the numbers 1,2,...,28 above the main diagonal with 1 for pair with most in common and 28 for the least similar pair. So, the ties free dissimilarity D' is as follows as shown in Fig. 3.

$$\begin{pmatrix} 0 & 9 & 21 & 15 & 22 & 17 & 5 & 2 \\ & 0 & 23 & 14 & 28 & 3 & 24 & 10 \\ & & 0 & 6 & 11 & 12 & 25 & 18 \\ & & & 0 & 4 & 7 & 1 & 27 \\ & & & & 0 & 8 & 26 & 13 \\ & & & & & 0 & 19 & 16 \\ & & & & & & 0 & 20 \\ & & & & & & & 0 \end{pmatrix}$$

Fig. 3: Ties free Dissimilarity matrix D'

We will then generate a hierarchical clustering of the patterns as follows.

(i) Start with each data object assigned to a different class (cluster).
(ii) Repeat step (iii) until all the objects are assigned to one cluster. Let each iteration of step (iii) represent the next clustering level.
(iii) Of all object pairs (i,j) such that i and j belong to different clusters, find the pair (u,v) with the least dissimilarity value. Combine the two clusters containing u and v.
(iv) Sketch the above hierarchical grouping in a graph (show the levels at which the various clusters merge).

A. Hierarchical clustering

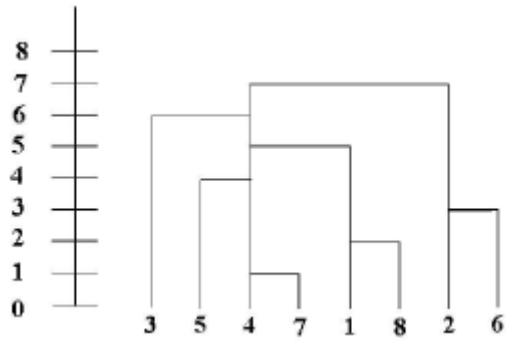

Fig. 4: Hierarchical clustering for movies data.

Let us list ten characteristics (features) of these 8 objects (movies) as shown in Fig. 4. These features can be based on nominal, ordinal, interval or ratio scales. Thus each of the objects is represented by a ten dimensional feature vectors.

The ten binary characteristics or features are:
1. Foreign language Film
2. Romance
3. Cartoon
4. Tragedy
5. Historical story
6. longer than 2.5 hours
7. Based on a famous Novel
8. Large number of roles
9. Impressive music
10. Instructive

The feature vectors for the movies are:
x1 : (1001001000);
x2 : (0000010001);
x3 : (0100000010);
x4 : (0100000010);
x5 : (0010100100);
x6 : (0110000011);
x7 : (1101001000);
x8 : (0001110111);

ISBN: 978-988-98671-8-8

The 8 x 8 proximity matrix by using the appropriate distance between a pair of patterns as the dissimilarity Distance matrix, H as shown in Fig. 5.

$$\begin{pmatrix} 0 & 5 & 5 & 5 & 6 & 7 & 1 & 7 \\ & 0 & 4 & 4 & 5 & 4 & 6 & 4 \\ & & 0 & 0 & 5 & 2 & 4 & 6 \\ & & & 0 & 5 & 2 & 4 & 6 \\ & & & & 0 & 5 & 7 & 5 \\ & & & & & 0 & 6 & 6 \\ & & & & & & 0 & 8 \\ & & & & & & & 0 \end{pmatrix}$$

Fig. 5: Distance matrix, H.

$$\begin{pmatrix} 0 & 11 & 12 & 13 & 19 & 25 & 2 & 26 \\ & 0 & 5 & 6 & 14 & 7 & 20 & 8 \\ & & 0 & 1 & 15 & 3 & 9 & 21 \\ & & & 0 & 16 & 4 & 10 & 22 \\ & & & & 0 & 17 & 27 & 18 \\ & & & & & 0 & 23 & 24 \\ & & & & & & 0 & 28 \\ & & & & & & & 0 \end{pmatrix}$$

Fig. 6: Tie-free matrix, H'

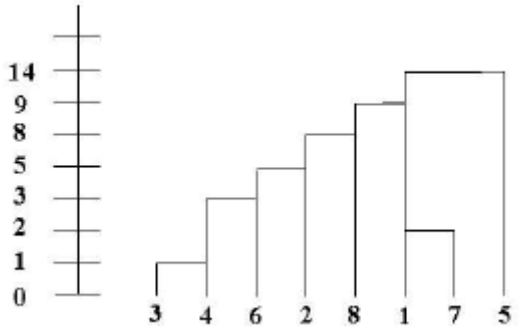

Fig. 7: Hierarchical clustering

IV. CONCLUSION.

The first clustering is more reasonable than the second clustering. If we cut the first clustering at level 3.5, five groups are formed, in which the similar movies are grouped. For example, the romance movies, i.e., movies 4 and 7, are in a group, and the instructive movies, i.e., movies 2 and 6, are in a group. The second clustering makes less sense comparatively. This suggests that if the features are not representative, the clustering accuracy will be impaired.

Weights are needed to reflect different importance of the features. For example, some people don't think it matters much if a movie is longer or shorter than 2.5 hours, so this criterion should have a small value of weight.

REFERENCES

[1] Jain, A. K. and Dubes, R. C. 1988. *Algorithms for Clustering Data*, Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

[2] Gowda, K. C. and Diday, E. 1992. *Symbolic clustering using a new dissimilarities measure. IEEE Trans. Syst. Man Cybern. 22*,368–378.

[3] Michalski, R., Stepp, R. E., And Diday,E. 1983. *Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Trans. Pattern Anal.Mach. Intell. PAMI-5*, 5 (Sept.), 396–409.

[4] Diday, E. 1988. The symbolic approach in clustering. In *Classification and Related Methods*, H. H. Bock, Ed. North-Holland Publishing Co., Amsterdam, The Netherlands.

[5] Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., New York, NY.

[6] Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*. 2nd ed. Academic Press Prof., Inc., San Diego, CA. 803–818.

[7] Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., New York, NY.

[8] Hoffman, R. and Jain, A. K. 1987. Segmentation and classification of range images. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-9*, 5 (Sept. 1987), 608–620.

[9] Jain, A. K. And Dubes, R. C. 1988. *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.