Neural Networks for Clustering Analysis of Molecular Data

Lin Wang\textsuperscript{1}, Minghu Jiang\textsuperscript{2}, Xiaofang Tang\textsuperscript{3}, Qiuqi Ruan\textsuperscript{3}, Baozong Yuan\textsuperscript{3}, Frank Noe\textsuperscript{4}

\textsuperscript{1}School of Electronic Eng., Beijing Univ. of Post and Telecom., Beijing, 100876, China
\textsuperscript{2}Lab of Computational Linguistics, Dept. of Chinese Language, Tsinghua University, Beijing, 100084, China
\textsuperscript{3}Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
\textsuperscript{4}Interdisciplinary Center for Scientific Computing, University of Heidelberg, 69210 Heidelberg, Germany

Email: jiang.mh@tsinghua.edu.cn

Abstract

In this paper hierarchical cluster and the competitive learning cluster are compared by using molecular data of large size sets. We construct a reproducible matrix to evaluate the quality of clustering, and dead nodes problem of the competitive learning network is solved by the conscience mechanism. The experimental results show that the hierarchical clustering can represent a multi-level hierarchy which show the tree relation of cluster distance, the competitive learning network has a good clustering reproducible and indicate the effectiveness of clusters for molecular data.

1. Introduction

Cluster analysis is a way to examine similarities and dissimilarities of observations. Data often fall naturally into groups or clusters, similar inputs should be classified as the same cluster and dissimilar inputs should be classified as different clusters. Clustering is realized by unsupervised learning, no predefined classes, that is partitioning a set of variables into classes or hierarchical groups of classes, those within each cluster are more closely related to one another than objects assigned to different clusters. The data are classed into subgroups or clusters, such that the distance of data items within the same cluster (intra-cluster variance) is small and the distance of data items stemming from different clusters (inter-cluster variance) is large [1]. The cluster analysis is more appropriate for some aspects of biological learning, human and social sciences and related areas. There are several major methods of clustering -- hierarchical clustering, k-means clustering and neural networks clustering etc. The goal of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered.

In this paper, Section 2 presents the hierarchical clustering, competitive learning network clustering and dead nodes problem. Section 3 describes experimental methods and results, compares and analysis the virtues and shortcomings of the two clustering methods. Section 4 summarizes the conclusions.

2. Cluster algorithms and models

2.1. Competitive learning model

Competitive learning can be realized by a neural network, the outputs of the network by competitive learning can reflect the correlation of the input data. If the data is clustered, the network will classify unseen inputs perfectly, and reflect in the input-output mapping. The arrangement of network outputs provides a flexible knowledge representation.

Competitive learning network usually consists of an input layer of \(N\) fan-out nodes and an output layer of \(K\) processing nodes. The nodes of output layer are full connected to that of input layer. The learning algorithm is shown as follows:

Step 1: Initialize the small random weights.
Step 2: Each output neuron (node) calculates the distance between the input vector and the weight vector connecting the input to it, if the \(i^{th}\) index of the closest competitive neuron is found and set its output activation to 1, the activation of all other competitive neurons to 0.
Step 3: The weight of the only winning neuron is updated so as to make the weights of the winning neuron closer to the current input pattern.
Step 4: Check if the stopping criterion is satisfied, else go to Step 2.

The result is that the winning node is more likely to win the competition again when a similar input is presented, and less likely to win when a very different input is presented. As more and more inputs are presented, the weight of each output node, which closest to a group of inputs, is updated toward those input vectors. Eventually, if there are enough output nodes, every cluster of similar inputs will have a node that outputs 1 when an input vector in the cluster is presented, while all others output 0. Thus, the competitive network clusters the input data.

2.2 Dead nodes problem

Unlike the supervised training, the competitive learning network has not general. Its defect is that dead nodes which these neurons with initial weights far from any input vectors and never win the competition, the result in their weights never get to learn, and these dead nodes do nothing, no matter how long the training is continued. The dead nodes increase with the number of competitive layer nodes increasing. To prevent this from happening, add a negative bias to each node and decreasing the bias each time when the node wins; this will make it harder for
2.4 Quality of Clustering

2.4.1 Intra and inter-class similarity. The high quality clusters have the characteristics with high intra-class similarity (or small within-cluster distance) and low inter-class similarity (or big between-cluster distance), which are evaluated by their scattering matrices [3]. The intra-class scattering matrix is defined as:

\[ S_{\text{intra}} = \sum_{i=1}^{N} P(w_i)E \{ (M_i - M_0) (M_i - M_0)^T \mid w_i \} \]

where, \( P(w_i) \) is the a priori probability of class \( w_i \), \( M_i \) is the mean vector (centroid) of class \( w_i \), \( M_0 = \sum_{i=1}^{N} P(w_i)M_i \), \( N \) is the number of classes. It reveals the scattering of samples around their respective class centroids. And the inter-class scattering matrix is defined as:

\[ S_{\text{inter}} = \sum_{i=1}^{N} P(w_i)(M_i - M_0) (M_i - M_0)^T \]

The diagonal items of \( S_{\text{intra}} \) and \( S_{\text{inter}} \) reveal the intra- and inter-class separability of individual features. If the diagonal item of \( S_{\text{intra}} \) is small while that of \( S_{\text{inter}} \) is large, then the corresponding feature has good class separability. The off-diagonal items of \( S_{\text{intra}} \) and \( S_{\text{inter}} \) reveal the correlation between different features.

2.4.2 Reproducible matrix. Because priori probability \( P(w_i) \) is usually unknown, the results of clustering can be evaluated by reproducible matrix. If the same data are trained repeat with different initial conditions, similarity of classes size distribution need be evaluated, i.e., a similar assignment of data points to classes should be measured for different initial conditions, assume that input data have 5000 samples, the algorithm is shown as follows:

Step 1: Repeat the clustering 5 times using random initial conditions each time;
Step 2: Setup an upper or lower triangular matrix, having 5000*5000/2 entries. In each entry \((i, j)\), count the number of times that data sets \(i\) and \(j\) have been assigned to the same cluster.
Step 3: Count the number of \((i, j)\) – pairs with values 2–3, judge if the intermediate values (2–3) are very rare. If the number of \((i, j)\) – pairs with values 2–3 below 10\%, then the clustering is to a good degree reproducible, and the clustering is thought to be successful. Ideally, only values of 0 and 5 should occur, which would mean that results are perfectly reproducible.

3. Experimental results

Two different molecular data sets are trained:

1) DATA-I: Molecular dynamics trajectories, 5000 samples of 42 raw features which are the backbone torsion angles for 21 selected residues.
2) DATA-II: Molecular dynamics trajectories, 5000 samples of 63 raw features which are the x/y/z-coordinates for 21 selected C\(_6\) coordinates.

The distribution and standard deviation of DATA-I and Data II are shown from Fig. 1 to Fig. 4, respectively. The corresponding hierarchical cluster and the heat map of hierarchical clustering (with correlation as the distance metric and average linkage used to generate the hierarchical tree) from DATA-I and Data II are shown from Fig. 5 to Fig. 8, respectively. By comparing results of distributions and standard deviations, it is easy to know that the DATA-I has larger fluctuant and standard deviations than DATA-II. Between-cluster distance of DATA-I is far larger than that of DATA-II.

The experimental results of competitive learning network are shown in Fig. 9 and Fig. 10. These figures show clustering distributions of competitive learning network, column is the number of samples for some a class, number of output nodes is 25, number of iterations is 40, Kohonen learning rate is 0.001, conscience learning rate is from 0.0001 to 0.0005. Fig. 9 show that some dead neurons exist, i.e., some neurons may not always get allocated, Fig. 10 show dead neurons are canceled by conscience learning algorithm.

In order to evaluate the cluster effect of competitive learning network, the same data are trained repeat with different initial conditions, similarity of classes size distribution is calculated, the numbers of \((i, j)\) – pairs with values 0–5 by using competitive learning network with 5 different initial conditions is calculated, Kohonen learning rate ≈0.001, Conscience learning rate = 0.0005. The numbers of \((i, j)\) – pairs with values 0–5 in an upper triangular matrix of 5000*5000/2 entries are respectively...
it is easily to obtain the number of $(i, j)$ – pairs with values $2\sim 3$ equal to $9.15\%$ (below $10\%$). The experiment shows that competitive learning network has a good clustering reproducible.

Fig. 2. Standard deviation of Data-I, 5000 samples

Fig. 3. Distribution of Data-II, 5000 samples

Fig. 4. Standard deviation of Data-II, 5000 samples

Fig. 5. Hierarchical cluster of 5000 samples of Data I

Fig. 6. Heat map of hierarchical clustering with correlation as the distance metric and average linkage used to generate the hierarchical tree from 5000 samples of Data I
4. Conclusions

The hierarchical clustering can represent a multi-level hierarchy which show the tree relation of cluster distance. The competitive learning network can cluster the input data, it only adapts to the node of winner, the winning node is more likely to win the competition again when a similar input is presented, thus similar inputs are clustered into same a class and dissimilar inputs are clustered into different classes, competitive learning network has a good clustering reproducible, the conscience learning algorithm can effectively cancel the dead nodes. These clustering methods all indicate the effectiveness of clusters for molecular data of large size sets.

Acknowledgement

This work was supported by National Natural Science Key Foundation of China (No. 60331010) and National Natural Science Foundation of China; Tsinghua University 985 Research Foundation; the Excellent Young Teacher Program, Ministry of Education, China; and State Key Lab of Pattern Recognition Open Foundation, Chinese Academy of Sciences.

References