Cluster analysis

Shengping Yang PhD, Gilbert Berdine MD

Women with coronary heart disease (CHD) might not experience chest pain. To avoid incorrect diagnosis and delayed treatment, is it possible to classify women by certain characteristics to facilitate better diagnosis?

Cluster analysis has been widely used in biomedical research, including using high throughput data to explore disease subtypes, especially unknown subtypes, clustering patients into different groups based on symptoms experienced, making predictions on patient outcomes using clinical and/or genomic information, etc. In general, the goal of a cluster analysis is to group subjects that have similar characteristics into the same group, while maximizing differences across groups. There are many algorithms for defining clusters or groups, but these algorithms fall mainly into two categories: supervised clustering and unsupervised clustering. Supervised and unsupervised clustering are directly analogous to supervised and unsupervised machine learning.

1. Supervised clustering

Supervised clustering maps input variables onto predefined clusters of an output variable (outcome). For our example, input variables might be age, history of chest pain, and diagnosis of diabetes, while the output variable might be incidence of myocardial infarction or death. An outcome variable can have two or more discrete classes. In general, the goal of supervised classification is to assign all subjects to one of the predefined classes (clusters) of the outcome variable. For example, we might define outcome clusters of "positive myocardial infarction with death", "positive myocardial infarction who survive", "negative myocardial infarction with death", and "negative myocardial infarction who survive." A number of algorithms have been developed for performing supervised classification, including logistic regression, decision trees, support vector machines, neuron networks, etc. Often, a training data set is used to train the classification model, to determine the best parameters for mapping input variables onto outcome clusters. The model is then "validated" by using these parameters on a new data set to determine the variance between the model predictions and the observed results. Depending on the number of outcome classes and the specific goal of a study, different algorithms can be used.

Corresponding author: **Shengping Yang**
Contact Information: Shengping.yang@ttuhsc.edu
DOI: 10.12746/swrccc.v6i26.504

1.1 Logistic regression

If there are two classes in the outcome set (binary outcome), then a logistic regression can be used to calculate the probability of a subject's belonging to one class or the other. Subsequently, predictions can be made based on a cut-off value (w.r.t. probability or odds ratio) obtained from a training dataset. We have previously presented the application of a logistic regression model from the perspective of evaluating the association between risk factor(s) and a binary outcome, and from the perspective of supervised learning, a logistic regression can be viewed as a classification algorithm.

While a logistic regression can be used for predicting binary outcome, it does not allow the outcome to have more than two discrete classes. Multinomial logistic regression does not have this limitation, however, some perhaps unrealistic assumptions have to be made in its application.

1.2 Decision tree

Compared to a logistic regression, which does not provide graphic output, a decision tree is featured with a tree-like graphical structure that includes root node, branches, and leaf nodes (see an example decision

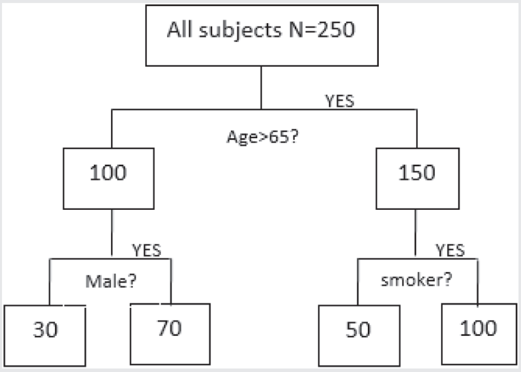

Figure 1. A decision tree example.

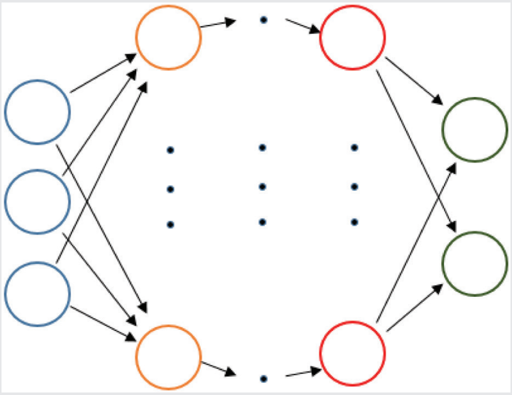

Figure 2. A neuron network example.

tree in Figure 1), to help visualize classification result. The goal of a decision tree is to divide the data/ subjects into smaller sets (subsets) based on information in the data, e.g., patient age, gender, so that the majority of subjects in each subset belong to the same outcome class. In the women with CHD study, ideally, once all subjects are divided into subsets, the majority of some of the subsets are women with CHD, while the majority of other subsets are women without CHD. This will greatly facilitate the diagnosis of CHD in women because the probability of having CHD for women with certain characteristics can be substantially higher than those with other characteristics. Computationally, a decision tree is constructed by minimizing the impurity of subsets with respect to the outcome cluster. However, the computational detail is beyond the scope of this article.

A decision tree is easy to visualize, intuitive to interpret, and straightforward to implement in practice. However, it is challenging to choose the best tree depth/number of terminal subsets, and perform pre- and post-pruning.

1.3 Neuron network

Neuron networks were originally developed to simulate the human brain. A neuron network in general consists of input, output, and hidden neurons. In Figure 2, blue circles represent input neurons, orange and red circles represent hidden neurons, and green circles represent output. Hidden neurons can have more than one layer. Some non-linear functions can be used to calculate the probability of each output class. After a learning process, a neuron network can be used to make predictions for new subjects. Very often predictions made by neuron networks outperform predictions made by linear regression. However, a disadvantage of using neuron networks is that the whole process in not intuitive, and thus the results are difficult to interpret.

Logistic regression, decision trees, and neuron networks are just a few examples of supervised classification. Because the outcome classes are predefined, it is thus feasible to choose factors that might be associated with outcome in a study to improve prediction. However, in unsupervised classification, that becomes more challenging.

2. Unsupervised classification

Unsupervised classification is more exploratory in nature because there is no known outcome cluster or variable. In general, the goal of unsupervised classification is to assign each subject into one of the finite and naturally formed clusters. The commonly used algorithms for unsupervised clustering include hierarchical clustering, mixture models, k-mean clustering, self-organizing maps, principle component analysis, etc. Because the total number of clusters is not known, it is challenging to evaluate accuracy of the clusters generated, especially for data with higher dimensions.

2.1 Hierarchical clustering

Hierarchical clustering may be represented by a tree-like (e.g., dendrogram) graphical structure that splits subjects into small subsets. Clusters are defined based on the path length necessary to connect elements of the cluster. Subjects within a cluster are closer to other subjects in the same cluster than they are to subjects in other clusters. The height of the nodes often represents to what extent subjects are similar to each other, the larger the height of the nodes connecting two subjects, the greater the difference between the two subjects. A cut-off value for separation distance can be used to classify subjects into clusters. In general, the larger the cut-off value, the fewer the number of clusters, and the larger the number of subjects in a cluster. In the example dendrograms, if a larger cut-off value is used (Figure 3; red rectangles), then all subjects are clustered into two groups, and if a slightly smaller cut-off value is used (Figure 3; blue rectangles), then all subjects are clustered into three groups. It is not straightforward to determine the best cut-off if the underlying true cluster structure is not clear. However, if there are true differences among subjects, then such a dendrogram provides clear visualization of the true underlying structure.

Due to the exploratory nature of hierarchical clustering, it is often difficult to determine what information (e.g., how many and what factors) should be used for calculating similarity across subjects. In fact, subjects can be clustered into entirely different groups if different information is used. In addition, hierarchical clustering is sensitive to outlier observations.

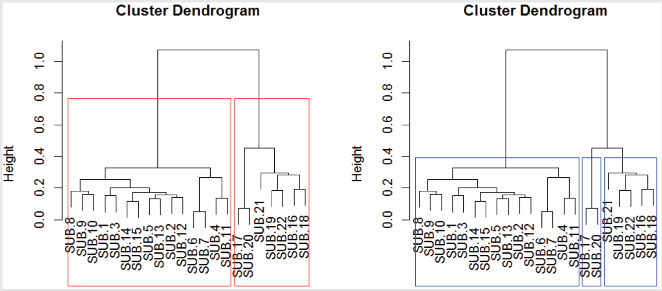

Figure 3. Dendrogram examples. Left: subjects were clustered into 2 groups; Right: subjects were clustered into 3 groups.

2.2 K-mean clustering

The goal of a k-mean clustering is to partition all subjects into a pre-determined number (k) of clusters based on minimizing the Euclidean distance between a subject and the centroid of the cluster that the subject belongs to. However, since subject partition is based primarily on minimizing the within-cluster variance, the cluster means converge towards the cluster center, thus the clusters are expected to (artificially) have similar size regardless of the nature of true underlying clusters. Consequently, this might result in incorrect classification. In addition, unlike logistic regression and decision trees, where the outcome clusters are known (always the same in repeated clustering applications), the clusters defined in k-mean clustering can sometimes change, which causes difficulties in interpretation.

There are substantial differences between supervised and unsupervised clustering. In supervised classification, the outcome classes are predefined, while in unsupervised classification, there is no known outcome in advance. If the goal of a study is to make predictions, then supervised classification is best; otherwise, if the goal is to explore potential data associations, then unsupervised classification is best. Some computational algorithms, such as regression, are best suited for supervised clustering. Some computational algorithms used for unsupervised clustering will require making decisions on certain cut-off values. In some cases, the same algorithm (with modification), such as neuron networks, might be used for either supervised or unsupervised clustering depending on the goal of a study.

Keywords: cluster analysis, decision tree, neural network, k-mean

From: Departments of Pathology (SY) and Internal Medicine (GB) at Texas Tech University Health Sciences Center in Lubbock, Texas
Submitted: 10/5/2018
Conflicts of interest: none

References

1. Bogumił K, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Central European J Operations Research 2017;26(1):135–15.
2. McCulloch W, Pitts W. A logical calculus of ideas immanent in nervous activity. Bulletin Mathematical Biophysics 1943; 5(4):115–133.
3. Rokach L, Maimon O. Clustering methods. Data mining and knowledge discovery handbook. Springer US 2005, pp. 321–352.
4. Xu R, Wunsch D. Survey of clustering algorithms. IEEE transactions on neural networks 2005;16(3):645–678.
5. Yang S, Berdine G. Categorical data analysis – logistic regression. The Southwest Respiratory and Critical Care Chronicles 2014;2(7):51–54.