SOCIETY OF ACTUARIES

Article from:

Forecasting & Futurism

December 2014 – Issue 10

PREDICTIVE MODELING SERIES

Data Clustering
And Its Application in Insurance

By Richard Xu and Dihui Lai

Predictive modeling is a very general term that includes many statistical algorithms to find relations in historical data for the purpose to predict future behavior. Some of these algorithms, such as data clustering, can be found useful for insurance applications.

Clustering analysis is a process of identifying patterns within a set of objects and grouping the objects with similarities into clusters. This unsupervised classification technique is ideal for exploratory analysis and is widely applied in fields such as object recognition and market segmentation. Additionally, this analytic method is also finding its way in supporting insurance, as more and more emphasis has been put on data driven decision-making. In this article, we are going to give a brief review on this method and demonstrate its power with an application.

INTRODUCTION

Data clustering groups objects into meaningful categories whose members exhibit similarities. Objects categorized in the same group are more similar to each other than to those in other groups. The similarity between objects is defined by a certain measure, such as distance between objects, dense areas of the data space, or other particular statistical distributions. Since each object is represented in a high-dimensional space, all these criteria have to be calculated in a multivariate method.

Clustering analysis can be broadly categorized as an unsupervised algorithm where data is not labeled. Explained mathematically, the input data have only variables x_{ij}, but no target variable y_j (where i is the index for data fields and j index for data records). For situations where no knowledge about segmentation exists or it is impossible to label all data points for a large dataset, clustering analysis is a very powerful way to discover data structure and relationship.

As the name indicates, the main purpose of clustering is to help organize and describe the objects of interest. We can use the knowledge to naturally classify objects for deeper understanding, to explore the underlying data structure, or to organize the data. Clustering has been widely used in many fields to achieve these goals. Examples of clustering applications include identifying hierarchical systems in biology,

information retrieval from the Internet, determining weather patterns in atmosphere and ocean for climatology, establishing book categories in libraries, and identifying common features and variations of disease conditions in psychology and medicine.

Business, including insurance industry, can also benefit from the application of clustering analysis. Very large amounts of information on current and potential customers have been collected. Clustering can be used to segment customers into a small number of groups for marketing activities. For example, market analysts can use cluster analysis to partition the general population of consumers into segments to better understand the relationships between different groups of consumers for marketing purposes.

CLUSTERING PROCEDURE

A simple clustering task can normally be completed in three steps: feature extraction, proximity measure definition and clustering/grouping.[1]

Feature extraction: This is a procedure of determining the features that best represent an object. For example, the most effective features to identify a person could be name, date of birth and gender. However, the effective feature could change depending on the question we are addressing. Considering the health condition of a person, the most useful feature might rather be his heart rate and blood pressure.

Proximity measure: Once objects are represented in their feature space, we need to determine the similarity measure between objects.[2] The most common measure is the Euclidean distance and is defined as

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{n}(x_{A,i} - x_{B,i})^2}$$

for two objects A and B in an n-dimensional feature space

The more similar two objects are, the smaller the distance in their feature space. The distance measure can also be defined as the sum of the absolute differences between two objects along each dimension. This is known as Manhattan distance, represented by the formula. The measure is related to the walking distance (number of blocks) between two points in a city and is therefore also called city-block distance. Alternative metrics, such as Chebyshev, Mahalanobis or Canberra distance, could be useful measures depending on the nature of the problem.

Clustering/Grouping: Clustering is the major process where we determine how each object is assigned to certain groups. Hierarchical clustering and partition clustering are common methods.

Hierarchical clustering is an iterative process of connecting objects based on distance. For example, at the beginning, each object is considered to be a cluster of its own. Then the pair of clusters with the shortest distance (most similar objects) is linked to form a new cluster. This linkage procedure then continues on the newly formed clusters until no further cluster can be established. However, the resulting hierarchical structure does not provide a unique way of clustering and the decision on the number of clusters can be challenging.[3]

Comparing to hierarchical clustering, partition clustering does not produce hierarchical structure and is therefore less computationally intensive. The algorithm, however, requires the user to determine the number of clusters before the analysis which requires caution.[4] A well-known algorithm of partition clustering is the k-mean. The algorithm uses k pre-determined points in the feature space as the center of a cluster. Each object in the feature space is then assigned to the closest cluster center. The new cluster center is then calculated using the members in the current cluster. The procedure continues until convergence (when each cluster no longer has changing objects).

CONTINUED ON **PAGE 24**

APPLICATION: RISK SEGMENTATION ON FOREIGN TRAVEL

The risk involved in international traveling is of particular interest to life insurance companies. It is important to understand the possible risks associated with countries in a quantitative way. Here we investigated 205 countries, each of which is represented by 25 feature variables including life expectancy, HIV prevalence, Communicable Disease Death Rate, and GDP. (Table. 1).

Category	Feature Variables
Life Expectancy	Life Expectancy
Health	Maternal Mortality; Infant Mortality; Underweight Children; Adult Obesity; HIV Prevalence; Communicable Disease Death Rate; Physician Density; Sanitation; Drinking Water; Hospital Beds
Safety/Security	Traffic; Homicide; Military Conflicts; Foreign Deaths; Occupational Accidents
Environment	Carbon Dioxide; Particulate Matter Concentration
Infrastructure	Internet Users; Mobile Phone; Road Density
Economic	GDP Per Capita (PPP); Corruption; Education-Expected Years of School; Gini Index

Table 1: Features that are used as variables to describe the risk characters of a country. The data are collected from several different sources, including the World Bank, the U.S. State Department, the CIA World Fact-book, the World Health Organization, World Economic Forum and the United Nations. The missing values in the data set are imputed using the bootstrap expectation maximization (EM) algorithm.

To understand the possible risk ensembles formed between different countries, we explore the 25-dimensional feature space with iterative hierarchical clustering. We use Euclidean distance as a measure of the similarity between countries. A clustering of six groups is selected from the hierarchical structures constructed by the algorithm. With the aid of principal component analysis,[5] we mapped the feature space to a two-dimensional plane where we are able to see the relations between the countries (Figure 1). Countries from clusters one and four are quite distinct from the rest and are easy to separate by visual inspection. Please note that Figure 1 only plots clusters in two-dimensions. If we could include more dimensions, the clusters would appear more distinct.

The resulting clusters exhibit intuitive appeal. European countries such as Germany, Spain and the United Kingdom share a cluster with the United States, Australia and Hong Kong, whereas countries such as Afghanistan and Pakistan occupy a different cluster. The resulting analysis largely supports judgmental categorization of countries based on expected risk from foreign travel.

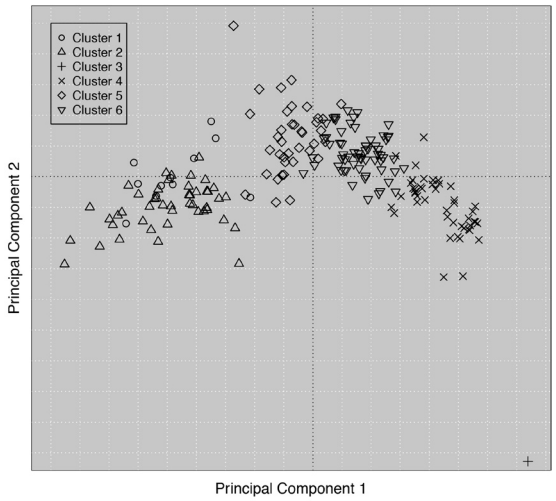

Figure 1: Projection of the feature space onto a two-dimensional principal component plane. Each dot represents a country and the color indicates the clusters that each country belongs to.

CONCLUSION

Data sets with high dimensions are normally difficult to understand. Data mining techniques can be helpful tools in exploring the underlying structures. This paper described how clustering methods could be used in an example within the life insurance industry. Further applications of the clustering method could help us better understand the client behavior, market segmentation, and customer classification.

REFERENCES

1. Jain AK, Murty MN, FLYNN PJ; Data Clustering: A Review; ACM Computing Surveys, Vol. 31, No. 3, 1999. https://ai.vub.ac.be/sites/default/files/data-clustering.pdf

2. Green PE and Rao, VR; A Note on Proximity Measures and Cluster Analysis; Journal of Marketing Research, Vol. 6, No. 3, 1969

3. Salvador S and Chan P; Determining the Number of Clusters/Segments in Hierarchical Clustering/ Segmentation Algorithms; Tools with Artificial Intelligence, 2004.

4. DUBES, R. C. 1987. How many clusters are best?—an experiment. Pattern Recogn. 20, 6 (Nov. 1, 1987), 645–663.

5. Wold S, Esbensen K, Geladi P; Principal component analysis; Chemometrics and Intelligent Laboratory Systems, 2 (1987) 37-52.

Richard Xu

Richard Xu, FSA, Ph.D., is senior data scientist at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *rxu@rgare.com.*

Dihui Lai

Dihui Lai, Ph.D., is data scientist analyst at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *dlai@rgare.com.*