Object Oriented Model for Cluster Analysis in Environmental Risk Management

František Huňka¹, Jaroslav Ráček²

Abstract

This paper concerns the problem of object oriented modeling of cluster analysis and applying the model on environmental risk management. Cluster analysis is an exploratory method for classification objects and its usage is namely appropriate when little or nothing is known about an internal structure of examined data. Object oriented model enables to increase the possibilities of cluster analysis considerably. One of these possibilities facilitates interactive hierarchical clustering, which means adding new clusters to the current ones and continue with hierarchical clustering from the reached level of the hierarchical clustering. Actual state of clustering is stored and retrieved by using persistent store and persistent objects. Interactive hierarchical clustering and persistency of objects enable to create hypothetical classification structure of well proven data that is used for the new objects classification. Moreover object-oriented perspective enables modeling different hierarchical clustering methods with the benefit of exploiting virtual procedure mechanism.

Described approach was tested on environmental data and the results can be used in environmental data risk management. The whole application is developed using the Mjolner BETA System and the BETA language for its rich facilities of the conceptual modeling.

1. Introduction

One of the problems one can meet in management is an object classification by a large number of various kinds of attributes that are somehow connected with the examined topic (area). It is possible to solve the problem by simplifying the problem that is by taking into account only a considerably restricted amount of attributes and judge by it. But this does not bring expected results mainly because of coarse estimation. In addition all mentioned attributes does not influence the examined area in the same way.

The crisis management we are examining is dealing with localities of a given stretches that can represent potential danger of various disasters both natural, transport accidents, terrorist attacks and so on. Each locality can be in this view characterized by a larger number of different types and weights attributes. By different types we mean different types of its value expression e.g. numerical type, value given by a table and so on. In addition as mentioned early each attribute has its own value of importance. E.g. it is obvious that textile plant is less dangerous than atomic power station or an oil refinery. So in each model weighted attributes of each objects has to be used. The problem to solve is statistical determination of these weights.

Cluster analysis is an exploratory method for helping to solve classification problems. Its use is appropriate when little or nothing is known about the category structure of a body of the data. The object of cluster analysis is to sort a sample of cases under consideration into groups such that the degree of asso-

¹ Fakulty of Science, University of Ostrava, email: frantisek.hunka@osu.cz
² Fakulty of Informatics, Masaryk University Brno, email: jaroslav.racek@muni.cz
ciation is high between members of the same group and low between members of the different groups. Cluster analysis is a modern statistical method of partitioning an observed sample into disjoint or overlapping homogeneous classes, to provide an operational classification. This classification may help to:

- assist identification (e.g. diagnosing);
- formulate hypotheses concerning the origin of population;
- predict the future behavior of population type etc.

Generally clustering works with only an unlabelled pattern matrix and a proximity matrix. The pattern matrix contains vectors of attributes for each single-element cluster. The proximity matrix holds proximity values between each pair of clusters. Clustering can be subdivided into hierarchical and partitional clustering (non-hierarchical).

A hierarchical clustering is a procedure for transforming a proximity matrix into a sequence of nested partitions. Achieved results may be visualize either in the form of dendrogram or directly using eigenvector projection into two-dimensional space.

2. Object model

2.1 Motivation

As mentioned above our goal is interactive hierarchical clustering. Having a new data for hierarchical clustering the general approach is to start with the clustering again from the very beginning. This may be time-consuming process repeatedly demanding work with the huge amount of data. On the other hand interactive clustering may be used not only for simplifying the clustering process but also for setting diagnoses for the new data (newly added clusters).

Object oriented approach enables to work directly with clusters themselves, as it is more natural and advantageous. Clustering is based on the fact that between each pair of clusters there is an association (called proximity or dissimilarity value depending on chosen clustering strategy) usually expressed by real value. Euclidean distance between two clusters is a dissimilarity value, whereas the correlation is a similarity value. As we use Euclidean distance we will work with the dissimilarity value. These associations are usually arranged in triangular matrix.

In our approach we do not use integrated proximity matrix but between each pair of clusters there is an association class that holds required values. This solution fulfills our aim to work directly with clusters themselves and enables for the association class to hold more characteristic than dissimilarity value.

Our approach demands effectively storing and retrieving processed data and results of the clustering. This is solved exploiting persistent objects that are suitable for these propose and in addition enable to split the whole application into rather smaller programs with the definite target.

2.2 Main structure

The main classes of our application are: CCluster, CProximity. Association between these classes can be seen in the picture 2.1 Class diagram and 2.2 Object diagram. CProximity class serves as an association class between each pair of clusters.

Class CCluster has the following main attributes:

- list of clusters - is used for saving references of all clusters, which are grouped to the same cluster from the beginning of the clustering (at the beginning there is no reference);
- proximity list - is used for saving references to all proximities of a given cluster with the other clusters. (In short it contains references to proximities of other clusters, which do not belong to a given cluster).
- indexVector – reference to the vector of measured attributes;
• *cardinality* - actual number of clusters in the cluster list.

Class *CProximity* has the following main attributes:

• *references* - to a pair of clusters;
• *dissimilarity value* - real value expressing actual association between two clusters;
• *uniMeasure* - alternative real value expressing proximity getting by applying of Hall’s method on the *pattern matrix*.

Clusters themselves are arranged in a linked list of clusters. Each cluster has a unique reference (through *indexVector*) to its vector of the measured attributes and is part of the *pattern matrix*.

---

**Fig. 1: Class diagram**

```
+---------------------+
| CCluster            |
|                    |
| ClusterA           |
|                    |
| ClusterB           |
|                    |
| CProximity         |
+---------------------+
```

**Fig. 2: Object diagram**

```
+---------------------+          +---------------------+          +---------------------+
| ClusterA: CCluster  |          | ClusterB: CCluster  |
|          |  Proximity: CProximity  |
|          |                  |
|          |   ClusterA: CCluster  |
|          |   ClusterB: CCluster  |
|          |   dissimilarityValue: @real |
|          |   uniMeasure: @real |
```

---

448
The structure of the pattern matrix is designed effectively. There is no problem to create "submatrix" composed of a chosen number of vectors if necessary. The pattern matrix that holds measured values of clusters is used in our approach for explicit calculation of dissimilarity, visualization, and for some clustering methods that need input data (e.g., Hall methods [10]). During the hierarchical clustering this link list of clusters is gradually reduced as clusters are grouped together. Subsequently, all the other lists in the class CCluster itself are updated in each step of the clustering. Finally, we can get only one big cluster or we can stop clustering at some meaningful level of the clustering.

2.3 Persistent store and its application
In the design of the persistent store and persistent objects we have to take into account the fact that persistence is type orthogonal. This means that when an object is saved, all the other objects reachable from that object are saved too. Similarly, when an object is read, all the objects reachable from that object are also read by need.

In our application, we create one persistent store that contains required persistent objects. We make persistent cluster list (contains clusters themselves), minimum proximity list (contains minimum dissimilarity value found in each step of clustering), pattern matrix and eigenvector matrix (used for two dimension projection). Cluster list is important for interactive clustering itself though minimum proximity list is necessary for the graphical output in the form of dendrogram.

Using persistent store was motivated by the amount of the data that has to be processed and by the necessity to split the application into rather smaller programs. Lazy fetch support in BETA enables to delay object fetch from secondary storage until an object is actually needed. When using lazy fetch, only the persistent root and a few more objects are initially fetched from the secondary storage. Fetch of the other reachable objects is deferred until their state is needed. At that time, the object is transparently fetched from the secondary storage.

Moreover, using persistent objects allows sharing data among different program executions, which enables reasonable splitting of the whole application into smaller parts.

2.4 Software solution and experimental data
Our application is done on the Mjolner BETA System that is an integrated and interactive general-purpose software development environment that supports industrial strength programming using object-oriented programming in the BETA programming language.

We choose this system mainly for these reasons:
• it supports the entire software development process including object-oriented design and implementation, user interface construction, persistence, graphics programming.
• the BETA language consists of a small number of concepts and the power of the language is the orthogonality of these concepts. This system provides powerful mechanisms for modeling.
• virtual procedures and virtual classes mechanisms provide powerful means for describing a number of cluster methods.
• it provides facilities for persistent objects. Clustering is known for working with huge amounts of data. For easy and effective manipulation with the data, we use persistent stores with persistent objects mechanism.

In our application, we exploit medical data that were collected during a project focused on diseases caused by smoking. Each person is characterized by 17 medical attributes gained by medical screening. The Trencin University in Slovakia provides this data [7].
3. Cluster analysis algorithm

3.1 Basic Algorithm of Processing

A hierarchical classification is a nested sequence of partitions, which corresponds to the recursive type of algorithm. The recursive type of algorithm needs the pattern matrix just only during introductory calculation of the single dissimilarities. Later it exploits the results of the previous step of clustering. There is also the other type of algorithm, the explicit algorithm. Contrary to the recursive type of algorithm the explicit type of algorithm works with the pattern matrix in each step of clustering, which is a bit awkward.

These two types of algorithms were deduced for each hierarchical clustering method. Generally the recursive type of algorithm is mostly used for it does not require connection to the pattern matrix.

Interactive clustering is a process when we on purpose combine the recursive type of algorithm with the explicit one. The explicit type of algorithm is exploit in the phase when a new data (clusters) are added and there are missing dissimilarities among the current structure of the clusters and the new ones. It can be also used just only for calculation of the dissimilarity values. Interactive clustering is usually done at some meaningful level of the hierarchical clustering when we temporary stop clustering process and add new clusters. The dissimilarity value expresses proximity between former cluster and the new one. The newly added data (clusters) can be used either for diagnoses purposes or for extending given structure of the clusters.

In that process we use persistent store both for the storing and retrieving the structure of the clusters. After interrupting clustering all-important objects are stored into the persistent store and we may print numerical value of clustering or execute a program for visualization.

During clustering it is possible for the user to move backward in nested sequence of the hierarchical clustering and get to the very meaningful level of clustering.

3.2 Detailed algorithm for hierarchical clustering

In this part we focus on description only one mostly used algorithm both in its recursive and explicit form. It concerns Ward-Wishart algorithm [9], which is based on the square of Euclid metrics. The recursive form of the algorithm works only with one pair of clusters with minimum dissimilarity value and is described by following formula:

\[
D(U, R) = \frac{1}{|R| + |U|} \left[ (|U| + |P|)D(U, P) + (|U| + |Q|)D(U, Q) - |U|D(P, Q) \right]
\]

where

- \(R\) is a new cluster composed of \(P\) and \(Q\) clusters.
- \(U\) represents all the other current clusters in the application.
- \(|R|\) represents a cardinality of a given cluster.
- \(D(U, R)\) is a new proximity value between \(U\) and \(R\) clusters.

Explicit form of Ward-Wishart algorithm is given by increase of \(I_{pq}\) value of objective function \(E\) that accompanies unification of the clusters \(P\) and \(Q\). If we notify \(P \cup Q = R\), increase of the objective function is

\[
I_{pq} = E_r - E_p - E_q
\]

where \(E_p\), \(E_q\), \(E_r\) are values of the objective function of the clusters \(P\), \(Q\), \(R\). Function \(E_A\) of cluster \(A = \{A_1, A_2, \ldots, A_t\}\) created by \(t\) clusters \(A_i = (a_{i1}, a_{i2}, \ldots, a_{it})\), \(i = 1, 2, \ldots, t\), is defined:

\[
E_A = \sum_{i=1}^{t} \sum_{j=1}^{t} (a_{ij} - \bar{a}_j)^2, \text{where} \quad \bar{a}_j = \frac{1}{t} \sum_{i=1}^{t} a_{ij}
\]
4. Environmental crisis management

In the introduction attributes of the examined areas was mentioned. Now a deeper description of these attributes will be made. In the following text a notion of geographical grid will be used. The term represents a square area of 10 by 10 kilometers. The aim of environmental crisis management is to predict a potential threat of some catastrophe such as earthquake, floods, traffic accident, wildfire, chemical accident, terrorist attack and so on.

Therefore the examined attributes correspond to the purpose of the crisis management. The most important attributes indicating a risky grid (area) could be the following:

- altitude
- settlement density
- average temperature
- sum of rainfall
- number of farms and agricultural enterprises
- number of enterprises divided by industry type
- traffic density
- percentage of built-up area, agricultural land, woods and body of water
- average age of citizens
- percentage of citizens in productive age
- education of citizens
- seismic stability of land

As can be seen the given attributes have not the same measured units. In order to adjust measured units statistical standardization has to be performed. The other important issue we are now facing is specifying individual attributes weight. As it is not a simple process we are solving it in different variants and we also contemplate to use simulation for this purpose. As it is not a deterministic process we are trying to suggest weight intervals for modeling purpose. When the specification of attributes weight will be finished gathering well statistically proven data (grid) may start. These well proven grids will be used for the creation of hypothetical clustering structure that will serve as so called “knowledge base” use in expert systems.

During clustering we have to taken into account also the other fact that it is important to select proper computing strategy depending on the type of the data. In general there are many different methods suitable for the different type of data.

5. Conclusions

Classification is one of the human’s ability to organize complexity in terms of hierarchies. Clustering is a means to focus on similarities between a numbers of phenomena and to ignore their differences, i.e. clustering is a classification of phenomena. Object oriented model for cluster analysis methods can be very helpful in situations where great number of different attributes of individual objects have to be classified. Using object-oriented perspective brings several advantages in the process of clustering. First of all a layer design can be used for the model creation. Model based on layers is not only better intellectually comprehensible and clearly organized but it is possible the model to be easily modified and extended. So the model is suitable for making experiments. The other assets are connected with internal structure of the model. It can work directly with the clusters themselves that is more natural and effective than work only with similarity matrix and in its consequence it enables to work with other taxonomy of measuring proximity of clusters. Virtual mechanism facilitates to model different cluster analysis methods. Using object-
oriented perspective we can create more complex structure that can be used not only for interactive hierarchical clustering but also can be used for calculation the other clustering characteristics.

Object persistency and proven data are the basis for hypothetical classification structure. The structure is used for finding characteristic properties of new unknown objects. Using object oriented model for cluster analysis in environmental crisis management seems to be promising progress.

Bibliography

Lukasová, A., Šarmanová, J.: Cluster Analysis Methods, SNTL Praha 1985  