A Data Mining Approach to Computational Taxonomy.

G.Perichinsky-R.Garcia Martinez
Departamento de Computación
Facultad de Ingeniería
Universidad de Buenos Aires
{gperi,rgm}@mara.fi.uba.ar

Abstract

This study investigates an approach of knowledge discovery and data mining in insufficient databases. An application of Computational Taxonomy analysis demonstrates that the approach is effective in such a data mining process. The approach is characterized by the use of both the second type of domain knowledge and visualization. This type of knowledge is newly defined in this study and deduced from supposition about background situations of the domain. The supposition is triggered by strong intuition about the extracted features in a recurrent process of data mining. This type of domain knowledge is useful not only for discovering interesting knowledge but also for guiding the subsequent search for more explicit and interesting knowledge. The visualization is very useful for triggering the supposition.

Key Words

Data mining, Knowledge discovery, Insufficient database, Taxonomy, Clustering

1. Introduction

We have a chance of making a clustering of a set of objects using their states or characters values. Some data mining technologies are applied to discovering the knowledge useful for the clustering analysis. This leads to investigation of effective technologies in discovering the target knowledge. The Data Matrix is, however, primarily used for obtaining the distances between objects. Therefore, the database seems insufficient for the cluster structure analysis. This leads to another investigation of appropriate technologies for data mining in insufficient databases.

In performing the data mining in insufficient databases, domain knowledge is especially effective not only in extracting interesting knowledge but also in guiding and containing the search for the interesting knowledge. Data visualization also significantly assists the data mining process as an interface between human and computer for iterative mining based on the domain knowledge.

In the course of data mining for the clustering analysis, it is noted that two types of domain knowledge are required:

The first is the domain knowledge which is typically defined and usually provided by some domain experts, in this study by Numerical Taxonomy researchers, and applications. The data mining problem involves many contextual constraints to be taken into account, which are only in experts' mind but not explicitly represented anywhere. The first type of domain knowledge brings to mind such important constraints.

The second is the domain knowledge which is newly defined in this study and deduced from supposition about background situations of a domain. The data mining process yields many incomplete features which can never be discarded to discover the target knowledge. The supposition is triggered by strong intuition about such features. The second type of domain knowledge is useful for guiding and containing the subsequent search for more explicit and interesting knowledge in the data mining process in insufficient databases.

To direct the search for the target knowledge, interaction is required between human relating to the domain knowledge and computer to do the search.

This leads to an iterative process of data mining with a preferred hierarchy for the interested set of data. This processing provides perhaps the best opportunity for the knowledge discovery in insufficient databases.

The discovered knowledge is primarily classified into two. One class is deeply concerned with the characteristics of the objects, subject matter of the classification. The other relates to a set of objects which belong to each cluster, and the characteristics of the clusters.

The knowledge includes interesting features extracted in the stages of data mining. Each feature can be associated with quantitative information to indicate how the feature is distinguishable. A

quantitative measurement can be defined as a frequency of the figure appearing in the related data. This measurement does, however, never directly indicate importance of the feature to the target goal. Fortunately, all of the features extracted in this study are shown in the visualizations. These are all comprehensible as they are, because these are all indicated according to amounts typically employed by domain experts. Therefore, the visualizations are offered as they are to the domain experts. It is left to them bow the extracted knowledge is used for the target goal.
The insufficient database potentially leads to problematic mining if one does not do preprocessing that is appropriate to the goal at hand. The approach studied here can be viewed as a preprocessing stage followed by the so called discovery induction learning algorithms such as ID3 and C4.5.

2.Data mining process

2.1. Original database.

The Data Matrix is stored in a data format, as values of data of the attributes, of the Original Database. The numbers of attributes increase or decrease to bring some sorts together to genre of clustering is then convenient for facilitating the knowledge discovery.

2.2. The first stage

Attention is typically paid in the application to several primary items such as number of attributes for the stabilization of the clusters. These can be displayed by visualizations from the original database in this first stage. This leads to discovery of primary knowledge, that is, clustering invariants or which attributes are taken into account and so forth.

2.3. Numerical Taxonomy.

We infer an analogy of the taxonomic representation in dynamic relational database.
We have explain the theoretical development of a domain's structured Database and how they can be represented in a Dynamic Database.
Immediately we apply our model to the structural aspects of the taxonomy, applying Scaling Methods for domains.
We define numerical methods used for establishing and defining clusters by their taxonomic distances.
We shall let C_{jk} stand for a general dissimilarity coefficient of which taxonomic distance, d_{jk}, is a special example. Euclidean distances will be used in the explanation of clustering techniques.
We use clustering strategy of space-conserving or the space-distorting strategies that appears as though the space in the immediate vicinity of a cluster has been contracted or dilated and if we return to the criterion of admission for a candidate joining an extant cluster, this is constant in all pair-group method.
Thus we can represent the data matrix and to compute the resemblance of normalized domains.
The steps of clustering are the recomputation of the coefficient of similarity for future admission followed by the admission criterion for new members to an established cluster.
The strategies of both space-conserving and space-distorting that appear in the immediate vicinity of a cluster either contract or dilate the space, and this is constant in all pair-group methods.

2.4. Dispersion

Once a typical value it is known of the variable of the states of the characters, it is necessary to have a parameter that give an idea of how scattered, or concentrated, are their values respect to the mean value.
It is considered to the variance as a moment of second order and represents the moment of inertia of the distribution of objects (mass) with respect to their gravity center: centroid.
The normalization of the states of the character causes that the average of all character will be of value zero and variance of unitary value.

If we take as value of the dispersion to the variance σ^2_d , we express the principle of minimal square.

2.5. Clusters and Spectra.

In discussing Sequential, Agglomerative, Hierarchic and Nonoverlapping (SAHN) clustering procedures we make a useful distinction between the types of measure.
Given two clusters **J** and **K** that are to be joined, the problem is to evaluate the dissimilarity between the resulting joint cluster and additional candidates **L** for further fusion. The fused cluster is denoted **(J,K)** , with $\mathbf{t_{j,k} = t_j + t_k}$ OTUs (taxonomic operational unit taxa or set of objects).
The cluster center or centroid represents an average object, which is simply a mathematical construct that permits the characterization of the Density, the Variance, the taxon (one OTU) radius and the range as **INVARIANT** quantities.
The states of the taxonomic characters in a class, defined ordinarily with reference to the set of their properties, allow one to calculate the distances between the members of the class. The distances can be established by the similarity relationship among individuals (obtaining a Matrix of Similarity that has been computed).
Considering characteristic spectra, in addition to the states of the characters or attributes of the OTUs, we introduce here the new **SPECTRAL** concepts of i)**OBJECTS** and ii)**FAMILY SPECTRA.**
Within the taxonomic space this method of clustering delimits taxonomic groups in such a manner that they can be visualized as characteristic spectra of an OTU and characteristic spectra of the families.
We define an individual spectral metric for the set of distances between an OTU and the other OTUs of the set. Each one provides the states of the characters and, therefore, is constant for each OTU, if the taxonomic conditions do not change (in analogy with the fasors).
The spectrum of taxonomic similarity is the set of distances between the OTUs of the set, that determine the constant characteristics of a cluster or family, for a given type of taxonomic conditions.
The importance of having an individual taxonomic spectrum (ITS) emerges from the fact that it gives information concerning the properties of the individuals through the states of their characters.
Invariants are found that characterize each cluster. Among them we mention the variance, the radius, the density and the centroid.
These invariants are associated with the spectra of taxonomic similarity that identify each family.

3 The second stage

There is another set of primary items that are typically taken into account in the clustering method. These includes Invariants, normalizations, distances, and so forth.
Another database is then required together with the original database, which includes information about the Matrix of Similarity.
It appear some results from the visualization in this second stage.
The visualization elucidates interesting relations between different couples of these items. These features also lead to interesting knowledge.
This means that the modified database is used in the second stage of data mining.

4. The third stage

The previous two stages relate to the first type of domain knowledge stated in the previous studies, while this third stage of data mining has a direct relationship with the second type of domain knowledge.
An interesting feature of the objects characteristics groups is extracted in the first stage of data mining. This leads to supposition about a wide variety of background situations of the groups. It is supposed that most of objects have a relationship through the distances between them. The second type of domain knowledge is deduced from such a supposition, which is used for guiding

the subsequent search in this stage for more explicit and interesting knowledge.
There is another information that the clustering has been computing through a iteration around the invariants. Based on the two types of domain knowledge, the search in this stage is directed to compute with couples or tuples of attributes. These clustering combinations are discovered using links techniques in order to groups the OTUs.

5.The forth stage

The features extracted in the third stage of data mining bring to some of the second types of domain knowledge, as described in the previous paragraph. This domain knowledge allows to direct this forth stage search to different combinations in the light of each step of the clustering. The genre of clustering is introduced in this stage to give more explicit and interesting features.
This interesting knowledge suggests an important decision to the clustering techniques to attach more importance to cluster combinations preferred by the links between them.

6. The discovered knowledge

The discovered knowledge is primarily classified into two. One class is deeply concerned with the characteristics of the objects, subject matter of the classification. The other relates to a set of objects which belong to each cluster, and the characteristics of the clusters.
All of the features extracted in this study are shown in the visualizations. These are all comprehensible as they are, because these are all indicated according to amounts typically employed by the domain experts. Therefore, the visualizations are offered as they are to the researchers. Thus, it is left to them how the extracted knowledge is used for the clustering.

7. The proposed approach

The approach of data mining proposed in this study is summarized above. The second type of domain knowledge is indicated as the one triggered by the results from each stage and by the discoveries. This type of domain knowledge is recognized as being very useful for the data mining in insufficient databases in the course of this study.
The approach studied here can be viewed as a preprocessing stage followed by the so called discovery driven induction algorithms.

8. Conclusion and future work

This study investigates an approach of knowledge discovery and data mining in insufficient databases. An application to CLUSTERING analysis demonstrates that the approach is effective in such a data mining process. The approach is characterized by the use of both the second type of domain knowledge and visualization. This type of knowledge is newly defined in this study and deduced from supposition about background situations of the domain. The supposition is triggered by strong intuition about the extracted features in a recurrent process of data mining.
This type of domain knowledge is useful not only for discovering interesting knowledge but also for guiding the subsequent search for more explicit and interesting knowledge. The visualization is very useful for triggering the supposition.
The approach yet relies on human ability to employ the domain knowledge through the interface between human and computer. It is left to future work to investigate which parts of the approach are, and how these arc, successfully implemented in data mining tools. It should be noted that applications recognized as successful invariably require the cooperation of domain analysts and developers of generic data mining tools.
It is often that the data used for knowledge discovery are not collected for the mining of knowledge, but a by product of other tasks. We hope that the approach proposed here will become a good reference for other various applications.