Experimental Findings of Categorial Characterization of Mineral Data Using Clustering Techniques with Focus on K-Mean & SOM

Under The Guidance Of :- Mrs. Pratiksha R. Deshmukh

PHARAD DAYANAND S.
Computer Department,
Government College of Engineering
and Research,Awasari(Pune).
Pune,India.

GAWALI SHARAD S.
Computer Department,
Government College of Engineering
and Research,Awasari(Pune).
Pune, India.

MAHAJAN AMOL A.
Computer Department,
Government College of Engineering
and Research,Awasari(Pune).
Pune,India.

MULE SIDDHESHWAR P.
Computer Department,
Government College of Engineering
and Research ,Awasari(Pune).
Pune, India.

Abstract— These instructions give you guidelines for Data mining process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. Clustering of data is important feature of data mining. Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure. Various Clustering techniques are available like-Apriori, Baeysian Networks, K-means and SOM.

K-Mean and SOM(Self-Organizing Map)are two well known, unsupervised clustering algorithms. Both the algorithms require set of Training data to find cluster. We proposed a construction to compare these two algorithms. In this we compare the clustering abilities of K-Means and SOM when applied to Mineral Data. The performance of these two clustering algorithms is compared on the basis of purity and overlapping of the clusters formed after applying these algorithms to the input data.

Index Terms—About K-Mean, SOM, Mineral Data.

I. Introduction

We are living in a world full of data. Every day, people encounter a large amount of information and store or represent it as data, for further analysis and management.

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Data mining is also known as the analysis step of the knowledge discovery in databases (KDD). Knowledge discovery means to develop something new. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the data by application software, Present the data in a useful format, such as a graph or table.

II. Working and Implementation Methodology

Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering.

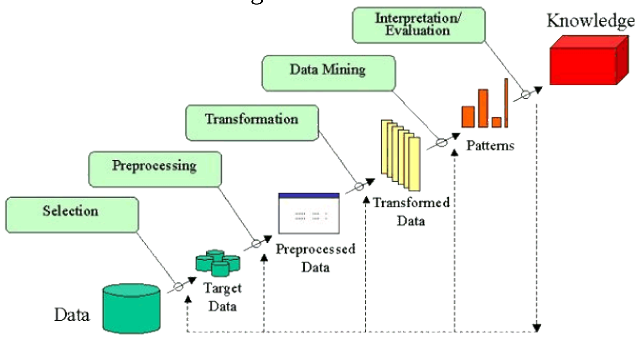

Figure 1.1: Steps of data mining process

1. Anomaly detection: This is the identification of the unusual records or data errors.
2. Association rule learning: It Searches the relationships between variables. This is sometimes referred to as market basket analysis.
3. Clustering: This is the process of finding groups and structures in the data that are in some way or another ”similar”, without using known structures in the data.
4. Classification: This is the task of generalizing known structure to apply to new data.
5. Regression: this process is used to search a function which modals the data with the least error.
6. Summarization: It provides a more compact representation of the data set, including visualization and report generation.

III. CLASSIFICATION OF CLUSTERING

Clustering algorithms can be categorized into partition based algorithms hierarchical-based algorithms, density based algorithms and grid-based algorithms. These methods vary in

1. the procedures used for measuring the similarity (within and between clusters)

2. the use of thresholds in constructing clusters

3. the manner of clustering, that is, whether they allow objects to belong to strictly to one cluster or can belong to more clusters in different degrees and the structure of the algorithm.

Irrespective of the method used, the resulting cluster structure is used as a result in itself, for inspection by a user, or to support retrieval of objects.

1. PARTITIONING ALGORITHMS
2. HIERARCHICAL ALGORITHM
3. DENSITY BASED ALGORITHMS
4. GRID DENSITY BASED ALGORITHMS
5. MODEL-BASED CLUSTERING

IV. K-MEAN CLUSTERING ALGORITHM

The KM clustering scheme used in the study is described in and its applications to remote sensing spectra can be found in . A training vector presented at the N input nodes activates the output node whose associated vector is closest, in the Euclidean sense, to the training vector. The activated node vector is updated with the rule

$V_i^{new} = V_i^{old} + \eta(V_T - V_i^{old})$…..(1)
Where V_i is the activated node vector,
V_T is the training vector that activated the node,
η is a small positive quantity called the learning rate

The clustering scheme is performed in the following steps:
1)Define M output node vectors $\in V^N$.
2)Determine the output nodes activated by the training vectors.
3)Update the activated node vectors with the rule(1).
4)Repeat steps(2) and (3) while decreasing in a linear fashion until the output vectors have quilibrated.
5)Associate the M equilibrated output vectors with vectors in the training set. The subset of training vectors associated with a single output vector is regarded as a cluster V^N

V. SOM CLUSTERING ALGORITHM

A description of the SOM scheme used in this study and its applications to several spectral libraries can be found in and An extensive bibliography of articles describing the SOM and its applications can also be found at The SOM has an input layer of N nodes and an output layer of M nodes usually arranged in a 1-D or 2-D regular array. Initially, each output node is assigned a vector $\in V^N$.

A training vector presented to the N input nodes activates the output node whose vector is closest to this input vector. During the training phase the entire training set is presented to the input nodes many times. At a given time the activated node vectors and their surrounding node vectors are updated with the rule:

$V_i^{new} = V_i^{old} + A_{ij}(V_T - V_i^{old})$…..(2)
Where V_i is the activated node vector,
V_T is the training vector.

After the training phase is complete the separation between adjacent nodes is represented by a wall whose height is the distance between the node vectors. The output layer is then partitioned into regions where each is surrounded by walls higher than a specified height and contains at least one node that is activated by the training set. The number of regions is inversely related to the wall height. The SOM clustering scheme is accomplished in the following steps:

1. Define M node vectors, arranged in a square lattice, where M 2NT.

2. Determined the output vectors which are activated by the input vectors.
3. Update the activated node vectors and their neighbors with the rule (2).
4. Repeat 3) and 4), decreasing and in a linear fashion until the output vectors equilibrate.
5. Calculate the wall heights.
6. Partition the output layer into K regions, where K is determined by a specified wall height. The region vectors are associated with K clusters in V^N.

VI. Purity of Cluster

$$\text{Purity}= 1+ (1/\log N)\sum^{n}_{i=1}(N_i/N_s)\log(N_i/N_s)\ldots(3)$$

where,
N is number of names represented,
N_i is the number of samples labeled with name i,
N_s is the total number of samples in the cluster. The ratio (N_i/N_s) is the probability of occurrence of name i in the cluster.

VII. Conclusion

We find:

1) The SOM and KM cluster differently when the number of clusters is small compared to the size of the spectral library.

2) Measures of mineralogical content and overlap reveal that at low cluster numbers KM clusters are less pure and overlap more than SOM clusters.

3) The higher SOM purities are primary due to the tendency of the SOM to to find more single member clusters at low cluster numbers

4)At low cluster numbers the KM clustering indicates a greater degree of overlap relative to the SOM clustering.

5)Clustering is the first step towards identification and classification is a intermediate step that associates unknown, unlabeled spectra with clusters derived from a set of well-known labeled spectra.

Figure Appendix

Figure 1.1: Steps of data mining process

REFERENCES

[1]. Robert Hogan, Giuseppe A. Marzo , Ted L. Roush A Comparison of Performance between Two Cluster Algorithms Applied to Mineral Spectra . 2009 IEEE Paper 1451, Updated 12 January 2009.

[2]. Elaine Rich and Kevin Knight, Shivashankar Nair, Artificial Intelligence.