Implementation of Clustering Data Mining Techniques to Improve Accuracy of Instances using PCA: A Case Study on Income Tax Payee

Dr. Gurpreet Singh, Vanshita Sharma

Department of Computer Sc.&Engg.
St. Soldier Institute of Engg. & Tech. Jalandhar (Punjab), India

Abstract

In this research, we use clustering algorithms to mine the data of tax and extract the information about tax payee by using some traditional algorithms K-MEANS, SOM and HAC algorithms from clustering it produce the better results with the analysis of traditional algorithms and compare it by applying on tax dataset. Clustering method will use for make the clusters of similar groups to extract the easily features or properties from samples of tax datasets? This comparison is able to find clusters in large high dimensional spaces efficiently. It is suitable for clustering in the full dimensional space as well as in subspaces. Experiments on both synthetic data and real-life data show that the technique is effective and also scales well for large high dimensional datasets.

Keywords: HAC, SOM, K-Means, PCA.

I. INTRODUCTION

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

In this paper, we are working on tax payer dataset and find out the persons that they are paying tax or not. Generally, the payee is the person to whom you make the payment, regardless of whether that person is the beneficial owner of the income. However, there are situations in which the payee is a person other than the one to whom you actually make a payment.

Principal Component Analysis (PCA) is a powerful and popular multivariate analysis method that lets you investigate multidimensional datasets with quantitative variables. It is widely used in biostatistics, marketing, sociology, and many other fields.

It is a projection method as it projects observations from a p-dimensional space with p variables to a k-dimensional space (where k < p) so as to conserve the maximum amount of information (information is measured here through the total variance of the dataset) from the initial dimensions. PCA dimensions are also called axes or Factors. If the information associated with the first 2 or 3 axes represents a sufficient percentage of the total variability of the scatter plot, the observations could be represented on a 2 or 3-dimensional chart, thus making interpretation much easier. The Principal Component Analysis, a Data Mining tool PCA can thus be considered as a Data Mining method as it allows to easily extract information from large datasets. There are several uses for it, including:

- The study and visualization of the correlations between variables to hopefully be able to limit the number of variables to be measured afterwards;
- Obtaining non-correlated factors which are linear combinations of the initial variables so as to use these factors in modeling methods such as linear regression, logistic regression or discriminant analysis;
- Visualizing observations in a 2- or 3-dimensional space in order to identify uniform or atypical groups of observations.

II. THE K-MEANS ALGORITHM

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation.

The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then recomputed and the process begins again. Here’s how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently.
depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

The Self-Organizing Map (SOM) Algorithm has proven to be one of the most powerful algorithms in data visualization and exploration. Application areas include various fields of science and technology, e.g., complex industrial processes, telecommunications systems, document and image databases, and even financial applications. The SOM maps the high-dimensional input vectors onto a two-dimensional grid of prototype vectors and orders them. For a human interpreter, the ordered prototype vectors are easier to visualize and explore than the original data.

The Self-Organizing Map (SOM) has proven to be one of the most powerful algorithms in data visualization and exploration. Application areas include various fields of science and technology, e.g., complex industrial processes, telecommunications systems, document and image databases, and even financial applications. The SOM maps the high-dimensional input vectors onto a two-dimensional grid of prototype vectors and orders them. For a human interpreter, the ordered prototype vectors are easier to visualize and explore than the original data.

HAC (Hierarchical Agglomerative Clustering) is a clustering method that produces “natural” groups of examples characterized by attributes. A tree, called a dendrogram, where successive agglomerations are showed, starting from one example per cluster, until the whole dataset belong to one cluster, describes the clustering process.

The main advantage of HAC is the user can guess the right partitioning by visualizing the tree, he usually prune the tree between nodes presenting an important variation. The main disadvantage is that requires the computation of distances between each example, which is very time consuming when the dataset size increases.

Tax Payee Dataset Used

<table>
<thead>
<tr>
<th>Attributes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial no</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>total income</td>
<td></td>
</tr>
<tr>
<td>Tax Payable</td>
<td></td>
</tr>
<tr>
<td>Income from Salary/Pension</td>
<td></td>
</tr>
<tr>
<td>Income from One House Property</td>
<td></td>
</tr>
<tr>
<td>Income from other sources</td>
<td></td>
</tr>
<tr>
<td>Gross Income</td>
<td></td>
</tr>
<tr>
<td>Add Agriculture</td>
<td></td>
</tr>
<tr>
<td>Gross total income</td>
<td></td>
</tr>
<tr>
<td>Deduction Under Chapter VI A</td>
<td></td>
</tr>
<tr>
<td>Total taxable income</td>
<td></td>
</tr>
<tr>
<td>Exempt income</td>
<td></td>
</tr>
<tr>
<td>Tax on Total Income</td>
<td></td>
</tr>
<tr>
<td>Income chargeable at %</td>
<td></td>
</tr>
<tr>
<td>tax payable</td>
<td></td>
</tr>
<tr>
<td>Less agriculture income</td>
<td></td>
</tr>
<tr>
<td>Tax on Total Income</td>
<td></td>
</tr>
<tr>
<td>education cess 3%</td>
<td></td>
</tr>
<tr>
<td>Taxable Income</td>
<td></td>
</tr>
<tr>
<td>Total TDS and TCS deduction</td>
<td></td>
</tr>
<tr>
<td>TAX PAYEE</td>
<td></td>
</tr>
</tbody>
</table>

III. IMPLEMENTATION AND RESULTS

Fig 1: Applying the SOM in data mining. Post-processing the SOM extracts qualitative or quantitative information of the data. Visualization and clustering provide qualitative information, while modeling and monitoring give quantitative information resulting in deeper understanding of the system behavior.

Fig 2: Load dataset. We have 500 instances and 16 attributes.

Fig 3: Check the integrity of the dataset by computing some descriptive statistics indicators. We insert the DEFINE STATUS component into the diagram using the...
shortcut into the toolbar. Then set all the variables as INPUT.

Fig 4: The number of instances into each cell is displayed. See that 45.20 % (Error Rate) of the TSS (Total Sum of Squares) is explained by the partitioning into 6 classes. I can compare this result to those of the other methods below.

Fig 5: Individuals who are in adjacent cells are also close in the original representation space. This is one of the main interests of this method. Let us check this assertion on the Tax dataset I cannot visualize the dataset into the original space. So we use a PCA in order to obtain a 2D representation. We try to visualize the relative positions of groups (clusters) in the scatter plot. I add the PRINCIPAL COMPONENT ANALYSIS component (FACTORIAL ANALYSIS tab) after the KOHONEN-SOM 1 component. Click on the VIEW menu. The first two factors representation space accounts for the 47.04% of the total variability. It seems weak, but on this dataset, it is enough to represent properly the instances.

Fig 6: Correspondence between the proximities into the SOM map and the proximities into the 2 first factors of PCA.

Fig 7: The DENDROGRAM tab of the visualization window is very important. By clicking on each node of the tree, I obtain the ID of the pre-clusters supplied by the SOM algorithm.

IV. CONCLUSION

In this review paper, study is being done on partitioning clustering algorithms and hierarchical clustering algorithms. The features of K-Means, SOM, HAC, PCA clustering algorithms are used. The comparison of algorithms is done with the existing algorithms K-Means, HAC, and SOM on Tax dataset using Tanagra data mining tool. The results by changing the number of clusters value specifies with PCA gives better performance clustering by reducing the sum of square error and improve accuracy on instances.

REFERENCES


