Comparative Analysis of Clustering Algorithms for Large-Scale Data sets using Self-Organizing Maps

Gyan Chand Sharma[1]*, Dr. Mohit Gupta[2]

[1] Research Scholar, University of Technology

[2] Associate Professor, Department of Computer Science, University of Technology

Abstract - However, SOM is also known as one of clustering techniques, since dimensionality reduction may also be seen as reducing (or clustering) input data to lower dimensions (or clusters). This research aims to group new enrolled students to a high school based on their academic grades using a SOM learning algorithm. The goal of cluster analysis is to identify distinct groupings within the data. The objects that belong to the same group ought to be comparable to one another, while those belonging to different groups ought to be as dissimilar to one another as is practicable. When dealing with clustering difficulties, one is particularly interested in the characterization of the clusters through the use of prototypes, which can be objects that are typical, representational, or representative in nature. There are a variety of ways to quantify differences between things. In this particular piece of research, the Euclidean distance was utilized as the comparative tool".

Keywords - Clustering, Data Sets, Organization Map

- X -

INTRODUCTION

The most significant method for classifying a mountain of information into manageable and understandable heaps is called cluster analysis, or CA for short. It is a tool for the reduction of the amount of data that creates subgroups that are more controlled than individual data points (Divya & Mohan 2013). In a manner analogous to factor analysis, it investigates the whole complement of inter-associations that exist between variables. There is a relationship between classification and both cluster analysis and discriminant analysis. On the other hand, the latter requires prior information of membership in each cluster in order to categorise new instances. This is necessary for the classification process. In CA, there is no prior knowledge of the elements that correlate to each cluster. An examination of the data helps to determine the categorization, often known as clusters. Following that, subsequent multivariate studies may be carried out on the clusters by themselves as separate groups (Wu et al 2009).

Clustering is actually the process of classifying items into a number of distinct groups. To be more specific, clustering is the process of segmenting a dataset into a subset of objects with the goal that the data present in each subset may possibly share certain similarities and frequently be located in close proximity to one another based on a certain defined distance measure. Clustering of data is a method that falls under the umbrella of broad approaches to statistical data analysis. It is utilised most frequently in a variety of fields, such as bioinformatics, pattern identification, image analysis, and pattern recognition (Osmar 1998).

PRIMARY CLUSTERING STEPS

In the next section, the essential stages involved in clustering will be described. The preprocessing and feature selection procedures, the similarity measure, the clustering method, the result validation, the result interpretation, and the application steps are all involved in the clustering process.

Preprocessing and Feature Selection

The majority of clustering models begin with the assumption that n-dimensional feature vectors adequately capture each and every data item. Performing this action requires selecting an appropriate characteristic and taking responsibility for calculating the underlying principles of the desired feature set. The decision to determine a separation of all the characteristics that are available in order to reduce the dimensionality of the problem space is usually appealing. It typically calls for a high-quality contract of data analysis and knowledge on the surrounding area.

Similarity Measure

A set of objects is grouped into more than a few clusters in the clustering process, so that comparable substances will be in the same cluster and dissimilar ones will be in different clusters. The similarity measure plays an important role in this process and plays an important role in the clustering process. In the process of clustering, each item is represented by its characteristics, and the degree of similarity between groups of objects is determined by a resemblance objective. This is a purpose that accepts as input two different collections of data items and produces as output the similarity measure that exists between those collections.

Result Validation

It is necessary to iterate back to some earlier point in the process if the outcome does not make any sense at all. It is possible that doing an analysis of the tendency to cluster will be helpful in determining whether or not clusters actually exist. It is important to keep in mind that certain clusters will be produced by any clustering method, regardless of whether or not naturally occurring clusters already exist.

Result Interpretation and Application

Data compression, theory generation (looking for patterns in the clustering of data), hypothesis testing (such as verifying feature correlation or other data properties through a high degree of cluster formation), and prediction are some examples of typical applications of clustering (once clusters have been formed from data and characterized, novel information substance can be able to be classified by the characteristics of the cluster to which they would belong).

Clustering Algorithm

The algorithms used in clustering are more generalised approaches that make use of certain similarity measurements as subroutines. The precise methods for clustering that are used are determined by the intended qualities of the clustering that is produced in the end. Complexity in terms of both time and space is another factor to take into account. An attempt is made by a clustering algorithm to discover natural groupings of components (or data) that have some resemblance with one another. The procedure for clustering also determines the dataset that serves as the group's centroid. The majority of algorithms examine the distance between a location and the cluster centroids to decide whether or not the point belongs to a cluster. According to Franti and Kivijarvi (2000) and Santhosh et al. (2012), the output of a clustering algorithm is essentially a statistical description of the cluster centroids together with the number of components that are contained inside each cluster. Over the course of time, a great variety of alternative approaches to clustering have been developed. The following paragraph will provide further explanation of these methods.

Partitioning Clustering

A database that contains n objects can be partitioned into k disjoint clusters by using partitioning clustering techniques. Every cluster has at least one item in it, and every object has a very specific home among the available clusters. In order to identify effective methods for clustering and partitioning, first split the dataset into initial divisions, and then, using several iterations, maximise the quality of the clusters that have been created. During each cycle, certain items are transferred from one cluster to another, which ultimately results in an improvement to the clustering's overall quality. The algorithm comes to an end when it determines that there is no way to improve the quality by moving any of the objects. Take note that the techniques for dividing and clustering are frequently heuristic, and that it is not certain that you will discover the clustering that has the highest quality.

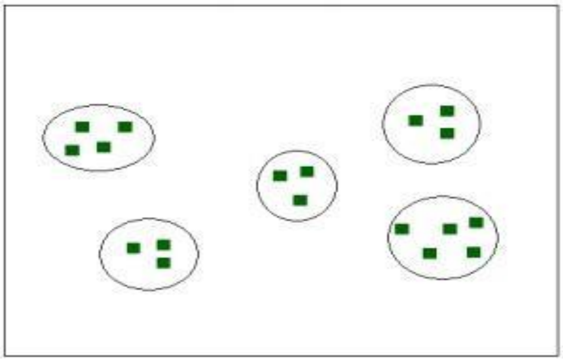


Figure 1 Partitional clustering

k-means clustering and k-medoids approaches like PAM and CLARANS (Clustering Large Application Based upon Randomized Search) are the most notable examples of this direction. The data is partitioned into non-overlapping groups and then clustered using the partitioned clustering method such that each data item belongs to exactly one cluster, as shown in Figure 3. It is determined in advance how many clusters there will be as well as the criteria for splitting the data.

Hierarchical Clustering

A hierarchical breakdown of the dataset is produced by the use of hierarchical clustering A hierarchical clustering organises data in such a way that it is possible for smaller clusters to be contained inside larger, more broad clusters. Agglomerative clustering and divisive clustering are the two methods that may be used to perform hierarchical clustering. A bottom-up methodology is used when agglomerative clustering is performed. Each each thing may be thought of as its own cluster. The subsequent stage involves combining the two clusters that are geographically located in close proximity to one another. Repeating the process of merging clusters occurs until either an exhaustive tree of clusters is constructed or a termination condition is satisfied, whichever comes first. The name for this type of cluster diagram is a dendrogram, and it includes clusters of varied sizes.

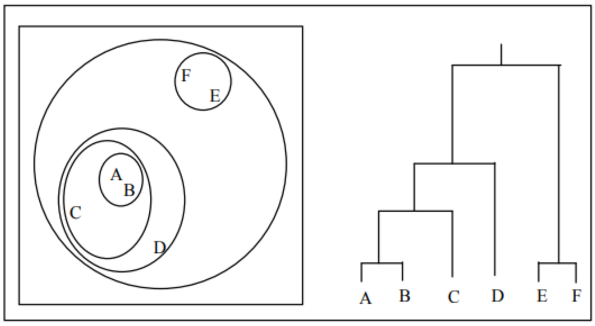


Figure 2 Hierarchical clustering

The clustering process is finished if either every cluster has precisely one item in it or a termination condition has been satisfied. The decision that an item belongs to a cluster in fundamental hierarchical clustering techniques cannot be changed after it has been made. Despite the fact that this trait makes it easier to locate clusters, the quality of the clustering may deteriorate as a result. Therefore, there are more advanced methods that do an analysis of the object links, such as in CURE or Chameleon or that also employ iterative relocation, such as BIRCH. Hierarchical clusters are clusters that are merged concentrically, as seen in Figure 1.4. This means that any data item will be present in several hierarchical clusters. The results of hierarchical clustering can either agglomerate or divide the data. The first step of the divisive clustering method is to place every data item in the same cluster. After that, a nested clustering is created by the original cluster to further split it until each data item is contained within its own cluster.

OBJECTIVES

1. To overcome the problems with discrediting attributes.

2. To reduce the problem of date mining the cluster numbers.

3. To propose animate grated frame work for visualized, exploratory data clustering, and pattern extraction in mixed data.

RESEARCH METHODOLOGY

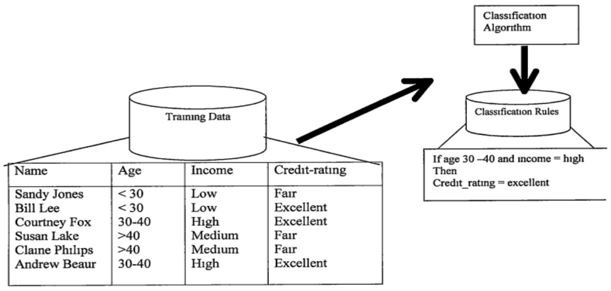
In this portion of the thesis, the ID3 method is investigated for its potential use in producing categorization rules for the census database. It was decided that a person's income would serve as the class attribute, and a decision tree was built using the ID3 method, as was covered in Section 6.2. It is possible to acquire classification rules for the class attribute. Knowing characteristics like as age, worker class, industry code, race, hispanic origin, and geography allows one to make educated guesses about the hourly salary that will be paid in the future.

Classification

"Both classification and prediction are types of data analysis that can be used to either extract models that describe significant data classes or to predict future data trends. However, while classification models continuous-valued functions and predicts categorical labels (or discrete values), prediction models categorical labels and predicts discrete values. It is presumed that every tuple belongs to one of the specified classes, and the class label property is the one that decides which class each tuple belongs to. Data tuples are sometimes referred to as samples, examples, and objects when discussing them within the framework of categorization. This stage of the process is referred to as supervised learning since the labels of each training sample are already known. The fundamental technique for decision tree induction, known as ID3, is known as the greedy algorithm. It builds decision trees in a top-down, recursive, divide-and-conquer fashion. In this implementation of the method, the categorical and discrete-valued qualities that are taken into consideration for the classification process are all categorical".

DATA ANALYSIS

Rules for categorization, decision trees, or mathematical formulae may be used to express the learnt model. For instance, if a database of consumer credit information is provided, one may develop categorization criteria to classify clients as either having exceptional or acceptable credit ratings. A greater comprehension of the information included in the database may be attained by using these rules, which can also be utilised to classify future data samples.



Name	Age	Income	Credıt-ratıng
Sandy Jones	< 30	Low	Faır
Bill Lee	< 30	Low	Excellent
Courtney Fox	30-40	Hıgh	Excellent
Susan Lake	>40	Medium	Faır
Claıne Phılıps	>40	Medıum	Faır
Andrew Beaur	30-40	Hıgh	Excellent

If age 30 –40 and ıncome = hıgh
Then
Credıt_ratıng = excellent

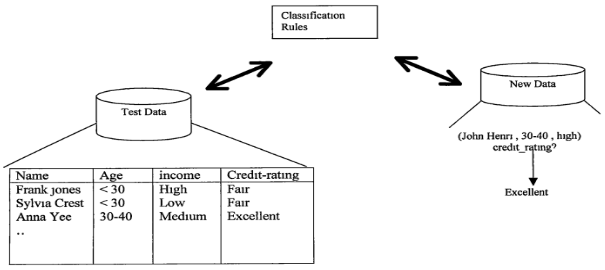
Figure 3 Learned Model

B) Second Step

The categorization system relies on the model. To begin, the predicted accuracy of the model (or classifier, if you prefer) is determined in the manner that will be detailed further below.

Predictive Accuracy of the Learned Model

The percentage of a given test set's samples that are correctly classified by a model is referred to as the model's accuracy on that particular test set. For each test sample, the known class label is contrasted with the learned model's class prediction for that sample to determine a model's accuracy. In the machine learning literature, such data are also referred to as "unknown" or "previously unseen data." If the accuracy of the model is deemed to be satisfactory, it can be applied to the classification of future data tuples or objects for which the class label is unknown. This will allow the model to classify data or objects for which the class label is unknown. For instance, the classification rules learned in Figure 5 from the analysis of data from existing customers can be used to predict the credit rating of new or future customers (i.e., customers who have not yet been seen), as shown in Figure 6. This is because new or future customers have not yet been seen.



Name	Age	income	Credit-rating
Frank jones	< 30	High	Fair
Sylvia Crest	< 30	Low	Fair
Anna Yee	30-40	Medium	Excellent
..			

Figure 4 Prediction

Classification by Decision Tree Induction

A decision tree is a tree structure that is similar to a flow chart in that each internal node symbolises a test that is being performed on an attribute, each branch represents a result that is being determined by the test, and the leaf nodes represent classes or class distributions. The root node of a tree is the node that is at the very top of the tree. Rectangles are used to represent the internal nodes, whereas ovals are used to represent the leaf nodes. In order to correctly categorise a sample whose characteristics are unknown, the values of the sample's attributes are compared to those in the decision tree. From the root node, a route is followed until it reaches a leaf node, which is where the class prediction for that sample is stored. Converting decision trees into categorization rules is a straightforward process. Figure 3 depicts a generic decision tree for your perusal. It is a graphical representation of the idea known as "buys computer," which determines whether or not a client is likely to make a purchase of a computer.

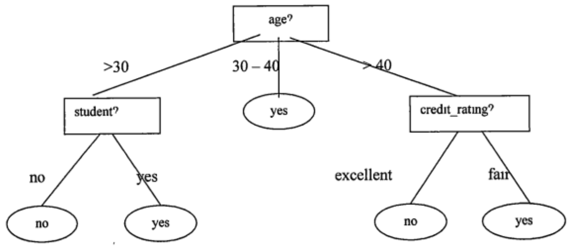


Figure 5 Decision tree

"ID3 Algorithm"

"The ID3 algorithm probably is the most popular algorithm in data mining It uses information gain as a criterion to find a suitable attribute to partition the universe until all granules can be understood or expressed by a formula Much effort has been made to extend the ID3 algorithm in order to get a better classification result The C4 5 proposed by Quinlan himself and fuzzy decision tree are among them Figure 4 shows the learning algorithm of ID3 The ID3 algorithm is studied and is used to generate classification rules for the census database. IF all cases in the training set belong to the same class THEN return the value of the class ELSE Select an attribute a to split the universe Divide the training set into non-empty subsets, one for each value of attribute a Return a tree with one branch for each subset, each branch having a descendant sub tree or a class value produced by applying the algorithm recursively for each subset in turn".

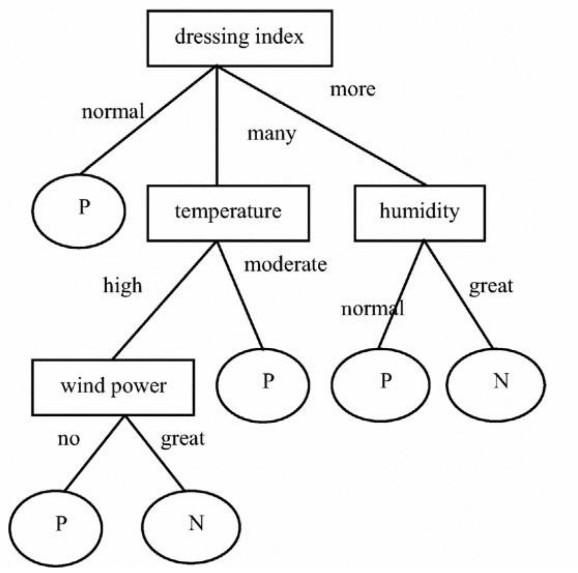


Figure 6 The learning algorithm of ID3

CONCLUSION

The current work explores three areas of data mining, using census database, viz.

- Large data clustering

- Data summarization
- Mining association rules
- A general procedure for data mining of census data is proposed as a sequence of clustering, prototype selection Each stage is sustained by computing CA
- Use of medoids and leaders as representative objects are shown to be successful
- Computation of prototypes with dissimilarity based thresholds is demonstrated.
- association rule generations by three different algorithms is demonstrated
- SOM , for visualizing the clusters is explored
- Predicting the pattern from the database was accurate to 98%
- ID3 algorithm was demonstrated to generate classification rules
- Apart from data mining algorithms biometric approach of palm print verification was implemented and its performance proved to be more accurate when compared to other biometric approaches.

REFERENCES

1. A. Imron, "Management of School-based Students" (Manajemen Peserta Didik Berbasis Sekolah), Malang: Universitas Negeri Malang, 2012.

2. T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological Cybernetics, vol. 43, no. 1, pp. 59--69, 1982.

3. I. Y. Purbasari, F. T. Anggraeny and N. Harianto, "Classification of broiler chicken eggs using support vector machine (svm) and feature selection algorithm," in International Joint Conference on Science and Technology, Nusa Dua, 2018.

4. M. Maimunah and T. Rokhman, "Classification of Declining Quality of Chicken Eggs Based on the Color of Shells Using Support Vector Machine "(Klasifikasi Penurunan Kualitas Telur Ayam Ras Berdasarkan Warna Kerabang Menggunakan Support Vector Machine)," Informatics for Educators and Professionals, vol. 3, no. 1, pp. 43-52, 2018.

5. D. Nurdiyah and I. A. Muwakhid, "Comparison of Support Vector Machine and K-Nearest Neighbor for Fertile and Infertile Egg Classification Based on Glcm Texture Analysis " (Perbandingan Support Vector Machine dan K-Nearest Neighbor Untuk Klasifikasi Telur Fertil Dan Infertil Berdasarkan Analisis Texture GLCM), Jurnal Transformatika, vol. 13, no. 2, pp. 29-34, 2016.

6. S. Lakho, A. H. Jalbani, M. S. Vighio, I. A. Memon, S. S. Soomro and S. Q. N, "Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network," Journal of Computing and Mathematical Sciences, vol. 1, no. 2, pp. 11-19, 2017.

7. F. Anggraeny, I. Purbasari and E. Suryaningsih, "ReliefF Feature Selection and Bayesian Network Model for Hepatitis Diagnosis," in International Conferences on Information Technology and Business (ICITB), Bandar Lampung, 2017.

8. F. T. Anggraeny, "Prediction of Student's Academic Achievement using Artificial Neural Network (Prediksi Prestasi Akademik Mahasiswa dengan Metode Jaringan Syaraf Tiruan)," in National Seminar of Information Technology Roles in Food, Chemical, and Manufacturing Industries to Support Development (Seminar Nasional Peran Teknologi Informasi di Bidang Industri Pangan, Kimia, dan Manufaktur dalam Menunjang Pembangunan), Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, 2009.

9. S. Isljamovic and M. Suknovic, "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study from Faculty of Organizational Sciences," in The Eurasia Proceedings of Educational & Social Sciences (EPESS), Konya, Turkey, 2014.

10. O. L. Usman and A. O. Adenubi, "Artificial Neural Network (ANN) Model for Predicting Students' Academic Performance," Journal of Science and Information Technology, vol. 1, no. 2, pp. 23-37, 2013.

11. E. Y. Obsie and S. A. Adem, "Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study," International Journal of Computer Applications, vol. 180, no. 40, pp. 39-47, 2018.

12. L. Rahmawati, A. D. Cahyani and S. S. Putro, "Utilization of SOM-IDB cluster method as an Analysis of Scholarship Acceptance Analysis (Pemanfaatan metode cluster SOM – IDB sebagai Analisa Pengelompokan Penerimaan Beasiswa)," University of Trunojoyo Madura, Bangkalan, 2013.

13. N. Hendayanti, G. Putri and M. Nurhidayati, "

Accuracy of Classification of STMIK STIKOM Bali Scholarship Recipients with Hybrid Self Organizing Maps and K-Mean Algorithms (Ketepatan Klasifikasi Penerima Beasiswa STMIK STIKOM Bali dengan Hybrid Self Organizing Maps dan Algoritma K-Mean)," VARIAN Journal, vol. 2, no. 1, pp. 1-7, 2018.

14. M. Bara, N. Ahmad, M. Modu and H. Ali, "Self-organizing map clustering method for the analysis of elearning activities," in Majan International Conference (MIC), Muscat, Oman, 2018.

15. Y. Lee, "Using Self-Organizing Map and Clustering to Investigate Problem-Solving Patterns in the Massive Open Online Course: An Exploratory Study," Journal of Educational Computing Research, vol. 57, no. 2, pp. 471-490, 2019.

Corresponding Author

Gyan Chand Sharma*

Research Scholar, University of Technology