International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: SJIF: 5.995
Available Online at www.journalijcar.org
Volume 7; Issue 2(G); February 2018; Page No. 10097-10100
DOI: http://dx.doi.org/10.24327/ijcar.2018.10100.1696

Research Article

LINEAR AND NON-LINER CLUSTERING ALGORITHMS FOR DATA MINING APPLICATIONS

Neelakantappa M*

Department of IT, BVRIT, Narasapur, TS

A R T I C L E I N F O

Article History:
Received 12th November, 2017
Received in revised form 13th December, 2017
Accepted 3rd January, 2018
Published online 28th February, 2018

Key words:

Data Mining, Linear and Non Linear Clustering

A B S T R A C T

Data mining is an important field of computer science which analyzes the large data sets and derives meaningful conclusions. In this paper we explained various data mining tools that are available in the present day market. We explained the concept of clustering, also surveyed different algorithms belonging to both linear and non linear clustering and explained them theoretically.

Copyright©2018 ***Neelakantappa M.*** *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

INTRODUCTION

Data mining is a process of analyzing data and forming meaningful conclusions from large sets of data. The data which is collected from multiple sources is integrated into single data storage is called target data. These sources may not be identical that is source may be of heterogeneous. Thus the process of discovering conclusions or knowledge from these large data sets is known as knowledge discovery. The data which is relevant to make analysis is decided and retrieved from these large data sets. Then it is pre-processed and transformed into the required standard format. Thus data mining algorithms are then applied to extract pattern or rules from which we interpret knowledge or information.

Definition

"Data Mining represents a process developed to examine large amounts of data that is routinely collected. The term also refers to a collection of tools used to perform the process. Data collected from various areas such as marketing, health, communication, etc… are used in data mining."

Tools for Data Mining

There are different types of tools available for the purpose of data mining. Some of the tools which are available today in the market are explained in this section.

R: It is an open source programming language environment available for statistical computing and graphics. It provides graphical and statistical techniques which includes linear and non-linear modeling, time-series analysis, and classification. R is widely adopted by the researchers for statistical software development and data analysis. Extensibility and data visualization are key features of R.

Rattle GUI

It is one of the graphical user interface for data mining written using the R programming language. It presents statistical and visual results of data and transforms data that can be readily modelled builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores on new datasets.

Orange

Orange is open source data mining, visualization software which helps beginners and experts for their analysis. It helps in designing of data analysis process through visual programming or python scripting. It represents most major algorithms for data mining and contains different visualization from scatter plots, bar charts trees to dendrograms, networks and heatmaps. It remembers user's choices, suggest most used combinations and intelligently chooses which communication channels to use. It has specialized add-ons like Bioorange for bio informatics. Even multidimensional data can become more sensible in 2D, especially with clever attribute ranking and selections.

Rapid-I Rapid Miner

Rapid-I Rapid Miner is one of the open source tool for data mining which is available as a data mining engine for the combination into own products. It has ability to run on major platforms and operating system. It is powerful but intuitive GUI for designing analysis process. It offers data integration, analytical ETL, data analysis and reporting on one single suite.

Corresponding author:* **Neelakantappa M
Department of IT, BVRIT, Narasapur, TS

It provides a graphical process design for standard tasks and scripting language for random operations.

Tanagara

Tanagara is open source data analysis software which is used for different purposes which proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases. The main purpose of Tanagara is to make the use data mining software in a better and easy way by conforming to the present practice of software development and allowing to analyze either real or synthetic data. The another purpose is to propose an architecture that allows the users to add to their own data mining method which helps in comparing their performances. It acts as experimental platform in order to do the essential work, dispensing them to deal with unpleasant part of data management. One more purpose is to give the direction to a developer mainly who is a beginner in diffusing a possible methodology for building this kind of software. It can be treated as a pedagogical tool for learning programming techniques since it permits to access the source code, to look pattern of software how it was built, the problems to avoid, necessary steps of the project, tools and code libraries used for the project.

Clustering

Clustering means partitioning the given data set and forming groups where similar kinds of objects are brought under one group while different objects into another group. Clustering or cluster analysis do not refers to a specific algorithm but it is a task to be solved. Here the formation of clusters is dependent on the type of the algorithm we use. It is used to segment the data. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high. The advancement of clustering algorithms lead to their use in wide variety of applications in different domains including image processing, computational biology, mobile communication, medicine, etc….

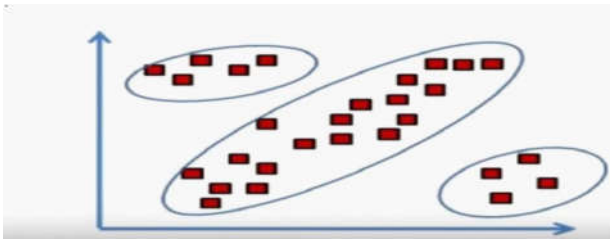

Figure 1 Formation of clusters

The above figure demonstrates the formation of clusters. There are three clusters. The red color items are the data objects. Each cluster is formed basing on some parameters. Mostly the parameter would be the distance between the objects. The clustering algorithms can be classified into two kinds.

1. Linear Clustering Algorithms
2. Non Linear Clustering Algorithms

In sections IV and V we see the algorithms of the above mentioned linear and non linear clustering algorithms.

Linear Clustering Algorithms

Fuzzy C-mean Clustering

Fuzzy c-means (FCM) is a method of clustering which allows a piece of data to belong to two or more clusters. This method was developed by Dunn in 1973 and improved by Bezdek in 1981. It is mostly used in pattern recognition.

Limitations

1. With fuzzy c-means, the centroid of a cluster is computed as the mean of all points, weighted by their degree of belonging to the cluster.
2. The degree of being in a certain cluster is related to the inverse to the distance to the Cluster.
3. The performance depends on initial centroids that are formed.

Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation

The image segmentation is defined as the dividing of an image into non-overlapped, consistent regions which are similar with respect to some characteristics such as gray value or texture. Fuzzy c-mean (FCM) is one of the most populous methods for image segmentation and its success chiefly attributes to the introduction of fuzziness for the belongingness of each image pixels. Compared with crisp, hard segmentation methods, FCM is able to continues to have more information from the original image. One disadvantage of standard FCM is not to consider any spatial information in image context, which makes it to be very sensitive to noise and other imaging artifacts.

Now we aim to include the spatial constraint into the FCM algorithm forming FCM_S. But this FCM_S calculates the neighborhood term in each iteration step, which is very much time-consuming. In order to reduce the computational loads of FCM_S, Chen and Zhang proposed two variants, FCM_S1 and FCM_S2, which simplified the neighborhood term of the objective function of FCM_S. These two algorithms introduce the extra mean-filtered image and median-filtered image respectively, which can be computed in advance, to replace the neighborhood term of FCM_S. Thus the execution times of both FCM_S1 and FCM_S2 are considerably reduced.

Disadvantage

The time taken to segment the image depends on its size.

Hierarchial Document Clustering

It uses Unweighted Pair Group Method with Arithmetic Mean (UPGMA). It uses a top down approach which is called as the divisive approach. It starts with all the data objects in the same cluster and repeatedly splits a cluster into many smaller clusters until a certain termination condition is satisfied.

Limitation

Methods in this category usually suffer from their inability to perform adjustment once a merge or split has been performed Frequent Itemset based Hierarchial Clustering: The above limitation can be eradicated using the method of Frequent Item set based Hierarchical Clustering(FIHC). FIHC uses only the global frequent items in document vectors, drastically reducing the dimensionality of the document set. Experiments show that clustering with reduced dimensionality is more efficient and scalable. FIHC can cluster 100K documents within several minutes while HFTC and UPGMA even do not produce a solution to clustering problem. FIHC is scalable and accurate. The clustering accuracy of FIHC consistently better compared to other methods. FIHC allows the user to specify an optional parameter, the desired number of clusters in the solution. The

cluster tree provides a logical organization of clusters which facilitates browsing the documents. Each cluster is attached with a cluster label that summarizes the documents in the cluster. There is no separate requirement of post-processing for generating these meaningful cluster descriptions.

Non Linear Clustering Algorithms

Efficient Kernel Clustering Using Random Fourier Features

Non linear shapes in the real world data sets can be handled by the kernel functions. Kernel-based clustering techniques, which use a nonlinear distance function, has the ability to find clusters of different densities and distributions, characteristics built-in, in many real data sets. However, their quadratic computational complexity provides them non scalable to very large data sets. We use Fourier maps for kernel clustering. The fourier map is proposed for huge scale classification. They set data points into a high dimensional non linear manifold where the clusters tend to be separable. The algorithm requires the calculation of nXn matrix where n is the size of data set. The data is mapped into a low-dimensional randomized feature space spanned by m basis vectors called as the Fourier components drawn from the Fourier transformation of the kernel. function. The inner product of the data points in this feature space approximates the kernel similarity between them. A linear learning algorithm is then applied to this data based on the random Fourier features, while obtaining performance similar to that of the kernel-based learning algorithm. This technique has been successful in applying to diverse learning tasks including classification, regression and retrieval.

Given the vector representations of the data points based on the random Fourier features, we apply the k-means algorithm to find the clusters in the data. When compared to the kernel k-means algorithm, the clustering algorithm achieves an approximation error of O(1/?m), where m is the number of Fourier components.

Kernel Sparse Subspace Clustering[9]

Image processing applications requires processing and representation of high dimensional data. The high-dimensional data can be better represented by a low dimensional subspace. For instance, it is well known that the set of face images under all possible adornment conditions can be well approximated by a 9-dimensional linear subspace. Similarly, the route path of a rigidly moving object in a video and hand written digits with different variations also lie in low dimensional subspaces. One can view the collection of data from different classes as samples from a union of low-dimensional subspaces. In subspace clustering, given the data from a union of subspaces, the objective is to find the number of subspaces, their dimensions, the segmentation of the data and a basis for each subspace. One of the advantages of these methods is that they are robust to noise and barricade. This is well supported by theoretical analysis.

An Efficient Minimum Spanning Tree Based Clustering Algorithm

The Minimum Spanning Tree[MST] algorithm forms a base for many clustering algorithms. MST has been studied for clustering by several researchers in different fields of biological data analysis, image processing and pattern recognition. One approach of the MST based clustering algorithm is to first construct MST over the complete dataset. Then the inconsistent edges are repeatedly removed to form connected components until some terminating condition is met. The resultant connected components represent the different clustering groups. The another approach is to maximize or minimize the degrees of the vertices, which is usually computationally expensive. In this algorithm the weight for each is usually considered as the distance between the end points forming the edge. On removal of the longest edge results into two-group clustering. Removing the next longest edge outcomes into three-group clustering and so on

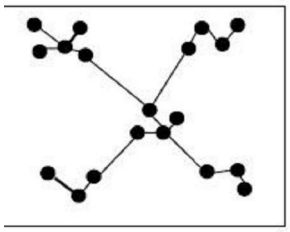

Fig 2a A MST containing all data points

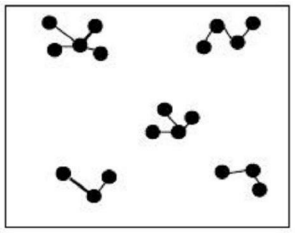

Fig 2 b Clusters after removing longest edges

The advantage of these MST based algorithms when compared to other algorithms like K-Means is that it do not requires the priori of knowledge like number of clusters, cluster seeds etc…, The intra ratio is better when compared with K-Means algorithm.

Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH)

It is one of the efficient data clustering algorithms for handling very large databases. BIRCH[11] incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources and also by considering the available memory and time as constraints. It can perform clustering with a single scan of data. It improves by performing the additional scans on the data. It can also handle noise which mean data points that are not part of underlying pattern. The algorithm considers only metric attributes.

The advantages of BIRCH over other algorithms are:

1. BIRCH is local in that each clustering decision is made without scanning all data points or all currently existing clusters. It uses natural closeness that reflect the natural closeness of points, and at the same time, can be incrementally maintained during the clustering process.
2. BIRCH exploits the observation that the data space is usually not uniformly occupied, and hence not every data point is equally important for clustering purposes. A dense region of points is treated collectively as a single cluster. Points in sparse regions are treated as outliers and removed optionally.
3. BIRCH makes full use of available memory to derive the finest possible clusters.
4. Proper parameter setting is useful to increase efficiency of BIRCH algorithm.

CONCLUSION

Data Mining is used to analyze the large data sets and to draw the conclusions. There are various data mining tools available. The paper had emphasized on importance of clustering and

also explained different linear and non linear algorithms theoretically. It also addressed the various parameter issues regarding to the clustering of data on very large data sets.

References

1. Sangeta Goele, Nisha Chandana, "Data Mining Trend In Past, Current and Future" *International Journal Of Computing and Business Research,* in Proc. I-Society 2012.
2. "Data Mining Tools and Trends - An Overview", S.Hameetha Begum, *International Journal of Emerging Research in Engineering and Technology*.
3. Fuzzy C- Means by Balaji K Juby N Zacharias
4. "Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation" , Weiling Cai Songcan Chen, Daoqiang Zhang, Science Direct
5. Hierarchical Document Clustering, Benjamin C. M. Fung, Ke Wang, and Martin Ester, Simon Fraser University, Canada
6. "Efficient Kernel Clustering Using Random Fourier Features", Radha Chitta, Rong Jin and Anil K. Jain, IEEE 20th Conference
7. "Kernel Sparse Subspace Clustering" by Vishal M. Patel and Rene Vidal IEEE2014 Conference
8. "An Efficient Minimum Spanning Tree based Clustering Algorithm", Prasanta K Jana and Azad Naik
9. "BalancedIterative Reducing And Clustering Using Hierarchies (BIRCH)", Tian Zhang, Raghu Ramakrishnan, Miron Livny.

How to cite this article:

Neelakantappa M (2018) 'Linear And Non-Liner Clustering Algorithms for Data Mining Applications', *International Journal of Current Advanced Research,* 07(2), pp. 10097-10100. DOI: http://dx.doi.org/10.24327/ijcar.2018.10100.1696

****