ANALYSIS AND SURVEY ON CLASSIFICATION PERFORMANCES OF DATA MINING CLUSTERING ALGORITHMS FOR REMOTELY SENSED MULTISPECTRAL IMAGE DATA

1 M.Amalmary
PhD Research Scholar , Hindusthan College of Arts & Science , TamilNadu, India.
2 Dr.A.Prakash,
Professor , Hindusthan College of Arts & Science . Coimbatore , TamilNadu, India.

Abstract

Data Mining is characterized as a technique used to extract and mine the undetectable, meaningful information from mountain of data. Clustering is an important technique that has been presented in the area of data mining. Clustering is characterized as a method used to aggregate similar data into a set of clusters based on some common characteristics. K-means is one of the popular partitional based clustering algorithms in the area of research. The impact factor of k-means is its straightforwardness, high productivity and scalability. However, is also contains number of limitations: random selection of initial centroids, number of cluster should be initialized and impact by anomalies. In perspective on these lacks, this paper shows a study of enhancements done to traditional k-means to handle such limitations.

Keywords: Cluster, Supervised, Data Mining, Classifications, Prediction, Regression.

1. Introduction

Data Mining is a technique used to extract and mine the undetectable, meaningful information from mountain of data. The term data mining is also relevantly utilized as Knowledge Discovery in Database, Knowledge designing. Based on the patterns we look for the Data Mining models and tasks are partitioned into two main categories Predictive models and Descriptive Models. Whereas the Predictive Model is utilized to anticipate the feasibility of result, the other Descriptive model is utilized to depict the important features of dataset. The kinds of Predictive model are classification, regression, prediction and time arrangement analysis. The various models incorporated into illustrative model are clustering, summarization, Association principles and sequence discovery. Clustering an unsupervised learning technique established in the area of data mining. Clustering or cluster analysis can be characterized as a data reduction apparatus used to create subgroups that are more manageable than individual datum. Generally, clustering is characterized as a procedure utilized for organizing/gathering a large amount of data into meaningful gatherings or clusters based on some similarity between data.

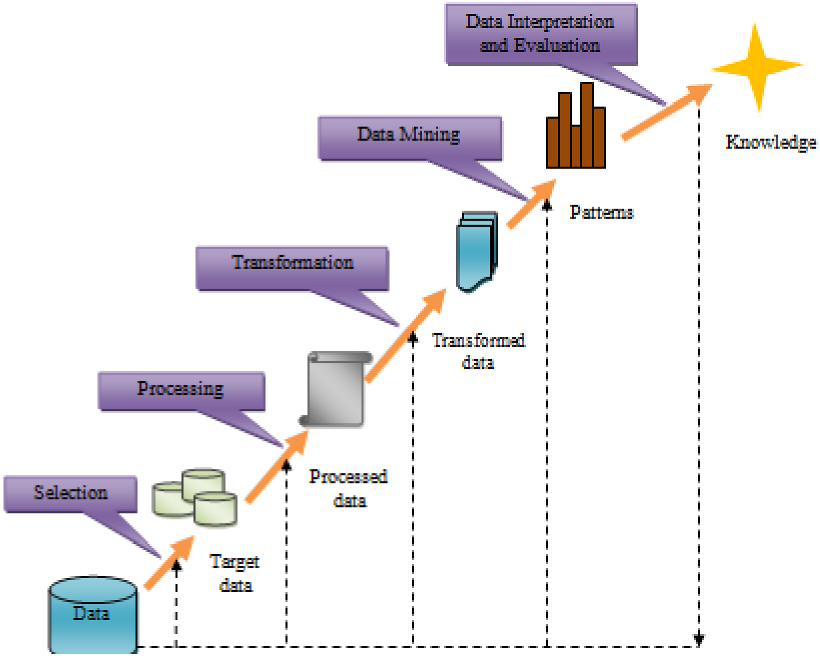

Figure 1: Knowledge Discovery Process

Clusters are the gatherings that have data similar on basis of common features and dissimilar to data in other clusters. The applications areas where clustering plays an important job are machine learning, image processing, data mining, marketing, content mining. The terms clustering and classifications are always confused with each other, since they are two separate terms. Whereas Clustering is unsupervised learning process because the subsequent clusters are not known before the execution which infers the absence of predefined classes in clustering. On the other hand classification is a supervised learning process because of quality of predefined classes. The high quality clustering is to obtain high intra cluster similarity and low entomb cluster similarity. There are number of clustering algorithms that are utilized to cluster the data. Fig.1 shows knowledge discovery iterative procedure including working of data mining. The said functionalities are measured to see the kind of patterns to be found in data mining tasks, Data Mining tasks can be categorized into two categories.

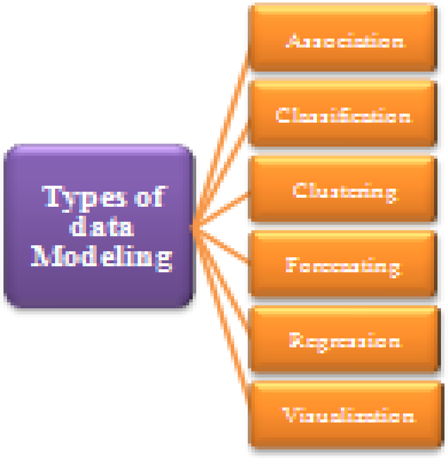

Figure 2: Data Modeling

Cluster analysis is operation on large collection of data which finds likenesses between information according to the attributes find in the information and gathering comparative information objects into clusters. An efficient clustering method will convey high quality clusters. We classified two main concepts on similarity that is high intra-class similarity have cohesive within clusters and low between class similarity have particular between clusters. The nature of a gathering result depends on upon both the comparability measure used by the strategy and its usage. The nature of a gathering strategy is likewise measured by its capacity to locate a couple or the majority of the inaccessible patterns. In information retrieval (IR), cluster analysis has been utilized to create gatherings of reports with the goal of enhancing the coherence and power of retrieval, or to determine the semantic of the literature of a field. The terms in an archive collection can also be clustered to show their relationships.

2. Literature Survey

[1] Amit Gupta, Naganna Chetty, Shraddha Shukla (2015) proposed model for enhancing the performance of K-means data clustering method and Naïve Bayes data classification method. Clustering plays an important job in data mining by categorizing objects into gatherings based on their similarity. Cluster analysis or clustering isn't an algorithm. It is a task which incorporates several distinct algorithms. Clustering is exceptionally essential in data mining. K-means clustering is most broadly utilized clustering algorithm which is utilized in many areas such as information retrieval, PC vision and pattern recognition. Naïve bayes classification assumes that each attribute is autonomous and contributes equally to the model. The general methodology for naïve bayes classifier utilized WEKA data preprocessing device for applying the naïve bayes classifier on the datasets utilized K-means clustering assigns n data points into k clusters with the goal that similar data points can be assembled together. It is an iterative method which assigns each point to the cluster whose centroid is the nearest. Then it again calculates the centroid of these gatherings by taking its average. K Means clustering algorithm applied on the chose attribute set produces significant enhancements in the values obtained for the original dataset. The results obtained for abalone and iris datasets shows an enhancement in the accuracy by 3.49% and 2% separately for the chose attribute set. **[2] Xiangyang Li, Nong Ye** (2006) proposed a data mining algorithm based on supervised clustering to learn data patterns and utilize these patterns for data classification. This algorithm enables a scalable incremental learning of patterns from data with both numeric and nominal variables. It is conceivable that a natural large cluster may be partitioned into several small clusters by framework cells, if the large cluster covers the area of several lattice cells. Much like a traditional hierarchical clustering algorithm, we utilize a supervised gathering strategy to check if any two clusters nearest to each other have the same target class and thus can be assembled into one cluster (as illustrated in Fig. 1). In this gathering system, a solitary linkage method is utilized, where the distance between two larger clusters is characterized as the distance between their nearest "points," the original clusters in our case. The weighted Euclidean distance is utilized for this calculation. Nominal to numeric. A multicategorical nominal variable can be converted into multiple binary variables, utilizing 0 or 1 to speak to either a categorical value absent or present in a data point. These binary variables can then be handled as numeric variables. A shortcoming of this method is that when there are many conceivable categorical values for a nominal variable, the method must deal with a large number of binary variables that have high reliance among them. The value of the original numeric variables should be scaled into [0,1] to make them compatible with the new binary variables. Numeric to nominal. The K-mode algorithm in is a rearranged version of the above K-prototype algorithm. Utilizing certain methods to convert numeric variables to nominal variables, this K-mode algorithm handles only nominal variables. **[3] Yuting Wan, Yanfei Zhon, Ailong Ma** (2018) proposed a completely automatic spectral– spatial fuzzy clustering method utilizing an adaptive multi objective memetic algorithm (AMOMA) for multispectral remote sensing imagery. A completely automatic spectral– spatial fuzzy clustering approach utilizing an adaptive MA is proposed for remote sensing image multi objective clustering. FAS2FC_AMOMA is achieved in two steps: an ADL and an ACL. In the ACL, the framework of a multi objective MA is utilized for the spectral– spatial remote sensing image clustering. Meanwhile, the GLS is set as the local-search-based approach, which is integrated into the global-search-based method (jDE). The expressions of the clustering objective functions for clustering the remote sensing imagery should be known. With the improvement of ongoing decades, the clustering objective functions are many and various, yet they should speak to the structure of the clustering data. In remote sensing image clustering, XB and Jm records have been broadly used. The values of XB and Jm are limited through constantly updating the membership μij and the clustering focuses Ui until it meeting the ceasing condition. Equation indicates that Jm calculates the variances within each cluster and they are then summed, which can speak to the level of compactness of the test data. Thus, a favorable clustering result can be obtained when the value of Jm is lower. In addition, from, XB considers the separation between the clusters in the term Sep(U), and lower value also indicates a favorable outcome for a data set. The target of the ADL is to automatically obtain the optimal number of cluster focuses, without the utilization of any earlier information. In addition, for upgrading the Jm and XB functions, jDE is used. The optimal number of cluster focuses is obtained, and thus we never again need to manually include the number of clusters into the ACL. In addition, because of the Jm and XB functions, not being limited concurrently, and the fact that the spatial information isn't considered, the world class individual chose in the ADL cannot be regarded as the final solution. In the ACL, we initialize the population and encode the individuals by randomly choosing the gray values of the pixels in the remote sensing image. Then, to take the spatial information into account, the two objective functions—XB and Jm_S—are simultaneously optimized by the multi objective memetic optimization approach. In addition, jDE and GLS, which are global-search method and local search operator, are utilized to meet the concept of the MA. Finally, to achieve completely automatic clustering, the angle-based knee point selection method is used. **[4] Chen Yang, Lorenzo Bruzzone, Fengyue Sun, Laijun Lu, Renchu Guan, and Yanchun Liang** (2010) proposed ld be considered for data analysis. In this paper, we propose a novel image clustering method [called fuzzy statistics-based affinity propagation (FS-AP)] which is based on a fuzzy statistical similarity measure (FSS) to extract land-spread information in multispectral imagery. Fluffiness is an inborn characteristic of remote-sensing imagery. Probabilistic clustering techniques utilize the concept of memberships to portray the degree by which a vector belongs to a cluster. The utilization of memberships gives probabilistic methods more realistic clustering than hard or fresh techniques. In conventional fuzzy classification, pixels can belong to several classes with various degrees of membership, which is the case when class descriptions overlap, e.g., within the sight of blended pixels. Pixels whose feature values are within these overlapping ranges can be viewed as ambiguous pixels. Although fuzzy concepts make it conceivable to portray these ambiguities, the main aim of each classification is to characterize classes as unambiguously as conceivable. Conventional fuzzy clustering, like the fuzzy K-means, needs the given cluster numbers, and the clustering results strongly rely upon the initial sequence of samples. There are two main inadequacies associated with fuzzy K-means, namely, inability to distinguish exceptions from non anomalies by weighing the memberships and attraction of the centroid toward the exceptions. Both lacks together are alluded to as "noise affectability." Moreover, conventional fuzzy schemes are based on maximum and minimum paradigms. The majority of the cluster analysis results are thus easy to trap in local optimizations which increase random city and cause trouble in getting accurate results. Multispectral and hyper spectral remote-sensing images often have broad entomb band correlations. Subsequently, the images may contain similar information and have similar spatial structures. At the same time, multispectral and hyper spectral images have their very own special characteristics, namely, the spatial variability of the spectral signature. According to this, we present the accompanying statistical characteristics, which are based on fuzzy statistics. **[5] Vibha L, P Deepa Shenoy, Venugopal K R, L M Patnaik** (2009) proposed automatic segmentation of the satellite image into unmistakable regions and further to extract tree tally from the vegetative area is displayed. Checking of trees also includes a pre-processing stage. Here the area of intrigue is in RGB format, force processing is included here. Here thresholding the certain values are done to distinguish the area of intrigue for example start of spring season as this was considered to be the trees. The unwanted area is converted into a darker level, thus converting the image into a grayscale image, which is later converted to a binary image. The algorithm proposed for tree including works two clearly. The main assumptions are phases in the primary phase we apply image enhancement and i) Large trees (total height greater than 6m) are target of smoothing strategy. The image enhancement is done utilizing the analysis. Automatic thresholding method, where vegetative area is extracted based on power method. It consists of setting ii) The base shape of crown projection of trees out of sight values for pixels beneath the threshold range while area of study is circular. the remaining are made equal to 1. In the following phase an iii) The research analysis does not take into consideration template for the tree is created called image matrix along with the distinctions in tree species. a mapping function. The two are finished to get a tree matrix. In this particular application, there are three regions of intrigue: the Land area, the Residential area and the Vegetations area. Great segmentation quality is achieved. Some encouraging experimental results have been obtained in satellite images Different Segments taken as from Quick Bird. We herein propose an efficient algorithm for extraction of automatic information on satellite of images. The segmentation procedure is important for tackling the problem of thematic mapping with utilization of remote sensing data, since binary images of high quality is required. **[6] Manjun Qin, Fengying Xie , Wei Li, Zhenwei Shi and Haopeng Zhang** (2018) proposed a novel dehazing method based on a deep convolutional neural network (CNN) with the residual structure is proposed for multispectral remote sensing images. An epic haze removal method based on the CNN is proposed for multispectral remote sensing imagery. In the designed network, multiple CNN individuals with the residual structure are utilized to learn the mapping from the hazy image to the clear image for various levels of haze samples. Through the fusion unit, the yields of CNN individuals are adaptively consolidated to yield the final reestablished image. In addition, considering the wavelength correlation property of atmospheric scattering, a wavelength-subordinate haze synthesis method based on the Rayleighs law is proposed to generate the labeled data to train the network. The designed network is start to finish, and the haze in the multispectral images can be successfully and adaptively expelled. A start to finish haze removal framework based on a CNN is proposed for multispectral remote sensing images, in which multiple CNN individuals are connected in parallel to learn the mapping from the hazy image to the clear image, and a fusion unit is utilized to adaptively join these individuals' yields to generate the final reestablished image. 2) The designed CNN individual utilizes multiscale convolutions to mine the multi scale features of haze and adopts the residual structure to diminish the learning trouble. Better performance is obtained. 3) another wavelength-subordinate haze simulation method is proposed to generate hazy multispectral images near real conditions, with which to train the designed network, progressively accurate dehazing results can be obtained. **[7] Mohamed Anis Loghmari, Mohamed Saber Naceur, and Mohamed Rached Boussema** (2006) proposed another approach based on a two-level source separation (TLSS), which consists of a spectral separation along the diverse utilized bands and a spatial separation along neighboring pixels of each image band. It is outstanding that in real images, like those covering agricultural fields, there is an important spatial correlation between neighboring pixels. This situation arises much of the time, and mainly when the ground area characterized by the IFOV contains many surface sorts that are relatively large to the spatial sensor resolution. Thus, spatially neighboring pixels are likely to belong to the same class.

This results in finding homogeneous regions in the classified images. The aim of the spatial-separation method is to discriminate between the diverse classes related to these homogeneous regions so as to associate only one physical theme to each source image. The generation of artificial images that save the characteristics of satellite images can be made from numerous points of view with varied degrees of realism. Because we deal with agricultural fields, each of our area intrigue must be homogeneous as for one class. another approach to detect, recognize, and analyze the compositional information from remotely detected images. This approach is exhibited under the sourceseparation method based on Bayesian estimation. In addition to a portion of the details of implementation, we showed that the Bayesian approach gives us the likelihood to push further a portion of the limitations of the classical technique in the source separation. **[8] Saroj K. Meher, B. Uma Shankar, and Ashish Ghosh** (2007) proposed o use the extracted features obtained by the wavelet transform (WT) rather than the original multispectral features of remote-sensing images for land spread classification. To deal with the non stationary behavior of signals, many research works have been carried out. Efforts are made to conquer the disadvantages of Fourier transform that assumes the signal to be stationary within its total range of analysis utilizing short-time Fourier transform and WT. WT stretches out the single-scale analysis to multi scale analysis that deteriorates the signal in multiple scales, where each scale speaks to a particular coarseness of the analyzed signal. WT endeavors to distinguish both the scale and space information of the occasion simultaneously which make it increasingly helpful for analysis of remote-sensing images. Further, various distinguishable characteristics like spatio-geometric information, energy at various scales, etc., which are normally called the signature of a particular land spread in remote-sensing images, are safeguarded in the WT-based decomposition performed with orthogonal basis. Selection of the training samples is made according to an earlier assumption of the land-spread regions. After learning the classifier with these training samples, it is utilized to classify the land fronts of the whole image. Various multispectral remote sensing images from IRS-1A and SPOT are utilized. However, we have included only two images because these bear diverse characteristics like spatial resolution, number of bands, and wavelengths, while they have similar land-spread classes. Also, a synthetic image is utilized to help our objective. The enhancement in performance of the classification scheme is confirmed utilizing one synthetic image and two remote sensing images. Various performance measures that are utilized to evaluate the classification results indicate that the addition of WF enhances the classification accuracy.

Conclusion

Popular data clustering algorithms in data mining are specifically hierarchical clustering, partitioning clustering, thickness based clustering(expectation maximization algorithm),grid based clustering, model based clustering and the soft computing methods. Soft computing techniques contain genetic algorithms, simulated annealing, fuzzy clustering; neural networks are as of late helpful to determine the issues in data mining from large databases. They attempt to give us the ease results, and hence the technique will speed - up. Data mining is the very high-quality area for learning the clustering algorithms. Clustering play a crucial job in many applications. The commonly utilized efficient clustering algorithm is k-means clustering. Clustering Algorithm is an important subject of research now a days in data mining. This paper have displayed a review of latest research work done in this area. However clustering algorithms is still at the stage of exploration and advancement. The review concludes that many upgrades are basically required on clustering to enhance problem of cluster initialization, cluster quality and productivity of algorithm.

References

1. Amit Gupta, Naganna Chetty, Shraddha Shukla, “A Classification Method to Classify High Dimensional Data”, 978-1-4673-9354-6/15/$31.00 ©2015 IEEE.

2. Xiangyang Li, Nong Ye, “A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 36, NO. 2, MARCH 2006.

3. Yuting Wan, Yanfei Zhon, Ailong Ma, “Fully Automatic Spectral–Spatial Fuzzy Clustering Using an Adaptive Multiobjective Memetic Algorithm for Multispectral Imagery”, 0196-2892 © 2018 IEEE.

4. Chen Yang, Lorenzo Bruzzone, Fengyue Sun, Laijun Lu, Renchu Guan, and Yanchun Liang, “A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in Multispectral Images”, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 48, NO. 6, JUNE 2010.

5. Vibha L, P Deepa Shenoy, Venugopal K R, L M Patnaik, "Robust Technique for Segmentation and Counting of Trees from Remotely Sensed Data", 2009 IEEE International Advance Computing Conference.

6. Manjun Qin, Fengying Xie , Wei Li, Zhenwei Shi and Haopeng Zhang, "Dehazing for Multispectral Remote Sensing Images Based on a Convolutional Neural Network With the Residual Architecture", 1939-1404 © 2018 IEEE.

7. Mohamed Anis Loghmari, Mohamed Saber Naceur, and Mohamed Rached Boussema, "A Spectral and Spatial Source Separation of Multispectral Images", IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 44, NO. 12, DECEMBER 2006.

8. Saroj K. Meher, B. Uma Shankar, and Ashish Ghosh, "Wavelet-Feature-Based Classifiers for Multispectral Remote-Sensing Images", IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 45, NO. 6, JUNE 2007.

9. Salman Khan , Anthony P. Doulgeris , Salvatore Savastano , Raffaella Guida, "AUTOMATIC CLUSTERING OF MULTISPECTRAL DATA USING A NON-GAUSSIAN STATISTICAL MODEL", 978-1-4799-5775-0/14/$31.00 ©2014 IEEE.

10. J. Senthilnath, Sushant Kulkarni, J.A. Benediktsson, and X-S. Yang, "A Novel Approach for Multi-Spectral Satellite Image Classification Based on the Bat Algorithm", IEEE Geoscience and Remote Sensing Letters (GRSL), 13(4), 599-603 (2016).

11.Xuejian Sun, Lifu Zhang, Senior Member, IEEE, Hang Yang, Taixia Wu, Yi Cen, and Yi Guo," Enhancement of Spectral Resolution for Remotely Sensed Multispectral Image" IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

12.Ahmed Tahraoui, Radja Kheddam, Abdenour Bouakache, Aichouche Belhadj-Aissa," Affinity Propagation for Unsupervised Classification of Remotely Sensed Images", 3rd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP'2017 May 22-24, 2017, Fez, Morroco.

13.Pablo Morales-Álvarez, Adrián Pérez-Suay, Gustau Camps-Valls," Remote Sensing Image Classification With Large-Scale Gaussian Processes", IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING2017.

14.Fan Hu, Gui-Song Xia, Zifeng Wang, , Xin Huang, Liangpei Zhang, Hong Sun,” Unsupervised Feature Learning Via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification”, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 8, NO. 5, MAY 2015.