ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Analysis of Various Data Mining Techniques: A Survey

Santhoshi Rupa Duvvi [1], Vyshnavi Nagireddy [2]

Assistant Professor, Department. of CSE, Avanthi Institute of Engineering and Technology, Vizianagaram, India [1]

Assistant Professor, Department. of CSE, Avanthi Institute of Engineering and Technology, Vizianagaram, India[2]

ABSTRACT: The process of data mining includes an analytical process that makes use of data and determines association by identifying patterns within the data. There are various task classifications involved in it like association, classification, prediction, clustering, sequential patterns and regression. This paper typically concentrates on concepts like clustering and Image processing and there by summarizes about clustering approaches. Clustering has been a prominent research issue in data mining due to its benefits in diverse applications and is extended in many research aspects. Various approaches within it are discussed here such that it brings various algorithms and its usage at one point..

KEYWORDS: Data mining, clustering, k-means, image processing, data.

I. INTRODUCTION

Data mining is a concept that extracts patterns which makes use of voluminous data sets that can be used by various sections of people in world for applications namely database management, information science, statistics and machine learning etc [1]. Data mining is usually classified into predictive and descriptive shown in fig 1. Clustering is a process where related data sets are grouped together which are disjoint [2]. Clustering is further organized into categories like hierarchical and partitioning clustering. It used in various applications like medical applications, image processing, pattern recognition, disaster prediction, social network analysis, stock exchange and market exchange etc. Medical applications in clustering have more impact on research as detection and prediction of any disease can be utilized to reduce the risk.

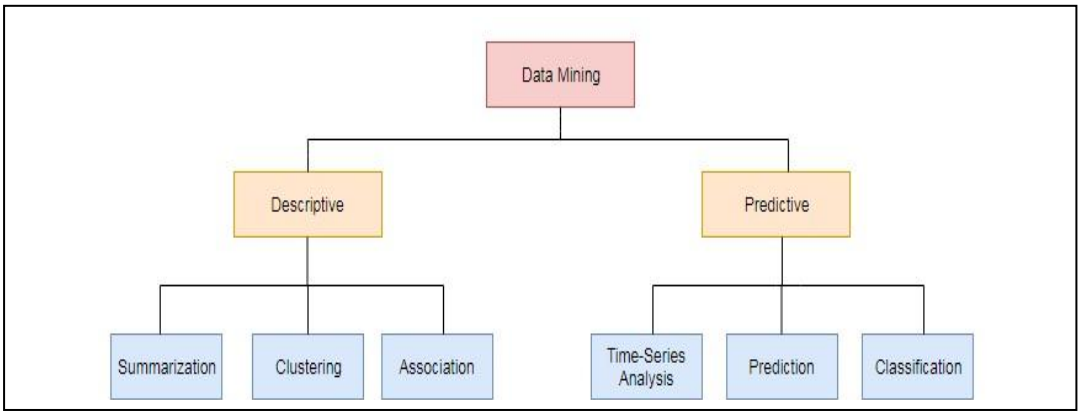

Fig. 1. Classification of Data mining

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Image processing, is a mechanism of categorizing image into different groups as mentioned in fig. 2. Many researchers have been working in present and past in the domain of image segmentation by making use of clustering. In image processing, feature extraction is a notable aspect of dimensionality reduction. Feature extraction includes rationalizing number of resources which are necessary to detail a large set of data precisely. Image processing becomes a prominent aspect in medical sciences where it is employed to withdraw the ROI(region of interest) from background.

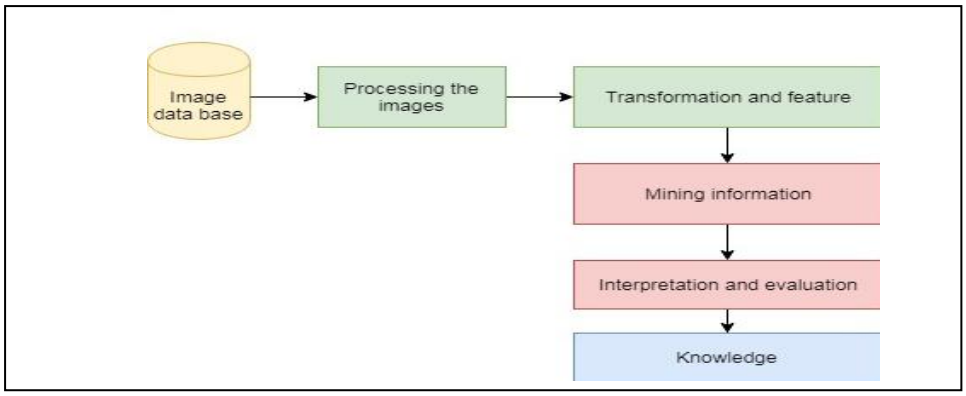

Fig. 2. Image Processing Mechanism

It has practices further namely threshold based, cluster based,edge based and neural network based. Amid these procedures, clustering method can be considered as most productive technique. Clustering is massive method which has various types like K-means clustering, Fuzzy c means clustering, mountain clustering method and subtractive clustering method. Amongst them, K-means is prominent due its advantages like simple and computationally fast than the hierarchical clustering. The procedure in common for a k-means clustering is, the count of variable. The results are varied upon cluster result for distinct number of clusters. Hence it is mandatory initialize the suitable count of cluster.

Various algorithms that use these clustering mechanisms are broadly classified as evolutionary algorithms, swarm intelligence algorithms and nature inspired algorithms etc. This survey focuses on a detailed overview of clustering analysis done by researchers in recent years.

This paper is ordered as follows section 2 describes regarding literature survey where various algorithms and their procedures are discussed, section 3focuses on diverse domains in Data mining and section 4 concludes the survey.

II. LITERATURE SURVEY

In Image Processing, clustering is preferbly advantageous for optimizing the search span of images in the database. K-means can be examned here, it is technique which divides deposit of data into certain set of groups. The procedure holds segregating collection of data into k number of data group. The given set of data is categorized into k number of clusters that are disjoint. K means algorithm consists of two distinct phases. In phase one, k means calculates the k centroid where as in phase two, it considers individual point to the cluster which has nearest centroid from the data point. There are distinct techniques to determine the distance of the nearest centroid and most used methods is Euclidean distance. In this method, grouping is done and then recalculations are made to get a new centroid later a new Euclidean distance is calculated between each center and each data point and assigns the points in the cluster which have minimum Euclidean distance. By its centroid every cluster in the partition interpreted by its member objects and by its centroid. Each cluster is the point to which the sum of the distances from all the objects in that cluster is minimized. So K-means is an iterative algorithm in which it minimizes the sum of distances from each of its object to its cluster centroid, over all clusters. We also have computational complexity which needs to be reviewed while formulating the k means clustering. [3]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Subtractive clustering is one of the methods to detect the nearest data point to find a centroid of cluster in view of density of data points. Process of compressing includes evaluating number and primary position of the cluster hub. It allocates the data stretch into meshing point and determine the unit of capacity for each data point base on distance to the actual data point. In a meshing point with many data points holds adjacent potential value. If meshing point have high density value, it will choose the initial cluster and reduce its potential value; the next cluster center will mesh with many data point nearby other than first cluster center grid points. procedure then has actions like acquiring new cluster center and minimizing the density of surrounding mesh point. The process is iterated until possiblity of grid point's ranks under the threshold. [4]

Clustering algorithms are been widely used in image processing sector. Such algorithms in clustering are suitable for specific set of images like images related to medical reports, microscopic images etc. On a similar discussion, algorithm namely Adaptive Fuzzy-k means clustering can be applied for image processing which can be processed using image databases. Fuzzy k means includes the concept of fuzzy logic and results optimal resluts compared to several clustering algorithms. A both qualitative and quantitative analysis of fuzzy k means algorithm provides a best segmentation fulfilment for diverse varieties of images and various count of segmented areas. The fuzzy algorithm yields best as compared to other clustering methods in image procesing. [5]

In [6] PSO (Particle swarm optimization) is distance based outlier detection mechanism is mentioned. PSO is a swarm intelligence algorithm where it has its advantages due to its fast convergence, simple implementation, less parameters and requirements for objective functions etc. And hence it got extended with multiple modifications and used in various applications namely clustering in WSN (Wireless sensor networks) for energy optimization, velocity adaptation. Each particle is called as a swarm. It inspired by social behavior of particles where group of birds and school of fish behavior is examined. Then, particle will have a fitness that is evaluated using a fitness function. Then comes the pbest(personal best) and gbest(global best) of a particle based on its position in the swarm. When a particle moves, the positions get updated such that pbest is the individual best and gbest is the best position among the recorded pbest. Positions in this PSO gets update for each iteration where velocity is calculated. In the proposed wok mentioned in [6], the degree of outlying in the cluster is studied by considering the distance between data point and k-nearest neighbor set now the PSO proposed algorithm is applied to detect outliers based on the outlying degree. Where a comparison is done using the threshold value which is proposed in the form of an equation named threshold equation.

III. DATA MINING IN DIVERSE DOMAINS

Table 1. Analysis on Data mining techniques in various Applications

Authors	Method	Area Of Application
[7]	Association	Genetic variation in health sciences
[8]	Association	Simulation of common disease genes
[9]	Association	Crime Pattern mining
[10]	Clustering	Text Mining
[11]	Clustering	Mining approaches for disease
[12]	Clustering	prediction
[13]	Classification	Image Segmentation
[14]	Classification	Early detection of diabetes
[15]	Classification	Sentiment analysis
		Fraud detection

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

IV. CONCLUSION AND FUTURE WORK

Applications in various aspects of data mining are referred such that it can be understood that mining can be surveyed in various projects. It is observed that image processing can be extended as image mining and clustering is a key in initial procedure of data mining. With the use of k-mean and its variants in fuzzy k-means algorithm, Subtractive clustering there is advantage of cost reduction that means time complexity is minimized. This paper engages with the application of standard k means and fuzzy k-means clustering algorithms in image processing and PSO in clustering. Apart from this an analysis of various data mining applications in research are discussed. With this study it is observed that Data mining tasks can be applied in current research applications with diverse domains.

REFERENCES

1. Nagappan, M., et al. "Heart Disease Prediction Using Data Mining Technique." (2019).
2. P Parwekar, Pritee, and Vyshnavi Nagireddy. "Comparative Analysis of SGO and PSO for Clustering in WSN." *2018 International Conference on Computing and Network Communications (CoCoNet)*. IEEE, 2018.
3. Chouhan, Preeti, and Mukesh Tiwari. "Image Retrieval Using Data Mining and Image Processing Techniques." *Image (IN)*3.12 (2015).
4. Dhanachandra, Nameirakpam, Khumanthem Manglem, and Yambem Jina Chanu. "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm." *Procedia Computer Science* 54 (2015): 764-771
5. Sulaiman, Siti Noraini, and Nor Ashidi Mat Isa. "Adaptive fuzzy-K-means clustering algorithm for image segmentation." *IEEE Transactions on Consumer Electronics* 56.4 (2010): 2661-2668.
6. Wahid, Abdul, and Annavarapu Chandra Sekhara Rao. "A distance-based outlier detection using particle swarm optimization technique." *Information and Communication Technology for Competitive Strategies*. Springer, Singapore, 2019. 633-643.
7. Risch, Neil, and Kathleen Merikangas. "The future of genetic studies of complex human diseases." *Science* 273.5281 (1996): 1516-1517.
8. Wu KS, Tanksley SD (1993) Abundance, polymorphism and genetic mapping of microsatellites in rice. Mol Gen Genet 241 : 225—235
9. Usha, D., and K. Rameshkumar. "A complete survey on application of frequent pattern mining and association rule mining on crime pattern mining." *International Journal of Advances in Computer Science and Technology* 3.4 (2014).
10. Stanisz, Tomasz, Jarosław Kwapień, and Stanisław Drożdż. "Linguistic data mining with complex networks: a stylometric-oriented approach." *Information Sciences* 482 (2019): 301-320.
11. Ghorbani, R., and R. Ghousi. "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction." *International Journal of Data and Network Science*3.2 (2019): 47-70.
12. Turkar, Hemantkumar R., and Nileshsingh V. Thakur. "Performance Comparison of Clustering Algorithms Based Image Segmentation on Mobile Devices." Cognitive Informatics and Soft Computing. Springer, Singapore, 2019. 581-591.
13. Das, Himansu, Bighnaraj Naik, and H. S. Behera. "Classification of diabetes mellitus disease (DMD): a data mining (DM) approach." Progress in Computing, Analytics and Networking. Springer, Singapore, 2018. 539-549.
14. Kang, Mangi, Jaelim Ahn, and Kichun Lee. "Opinion mining using ensemble text hidden Markov models for text classification." *Expert Systems with Applications* 94 (2018): 218-227.
15. Jurgovsky, Johannes, et al. "Sequence classification for credit-card fraud detection." Expert Systems with Applications 100 (2018): 234-245.