Enhancement in Clustering to Find Cluster Center and Improve Cluster Quality

[1]Veerpal, [2]Sanyam Anand
[1]Student of M.Tech, LPU, Phagwara, Punjab, India
[2]Assistant Professor, LPU, Phagwara, Punjab, India

Abstract

In daily life more much amount of data is produced.Sometimes data created is uncertain and is difficult to handle. By using clustering uncertain data can be handled. Determining cluster center or say guesstimate of cluster center issues from these dissemination are expected. Problem is cracked by non-linear minimum square optimization compliant the cluster center and cluster size. Clustering has many applications in many domains Where clustering is used in various fields of our real life. It is a development which are given here and described few on it. Also is used in various fields are unknown on it. Finding cluster center and cluster sizes are very useful in many areas. Purposed work would describe the enhancement of clustering to define the cluster center to increase cluster quality. Suggested technique is very useful for practical applications and theoretical reflection for clustering problems

Keywords

Clustering, Cluster Analysis, Uncertainty, Cluster Center

I. Introduction

Data mining is defined as process of taking out of the inherent and earlier unknown and really theoretically useful information from huge amount of data. It is also called as abstraction of hidden patterns. It is used to discover pattern in data, process should be completely programmed or semiautomatic pattern discovery. And detection must be expressive. It is also a procedure of discovery hidden information in the database. This process suggest one or more computer wisdom practices to automatically analyse and mine knowledge from data contain within the database, it is part of knowledge discovery process. It applies many algorithms to large data to produce models or designs interesting to the user and extract the hidden patterns. It is also called as knowledge discovery data or say KDD process as shown diagrammatically. Data mining is taking out of hidden patterns from huge amount of data. For example ore mining, data mining is process as shown in diagram, firstly data is collected from various resources then after that data cleansing is performed is called as data preprocessing. All type of data noise is removed and then cleaned data is integrated. After data mining is performed pattern evaluation is there. Patterns are evaluated in pattern evaluation step.

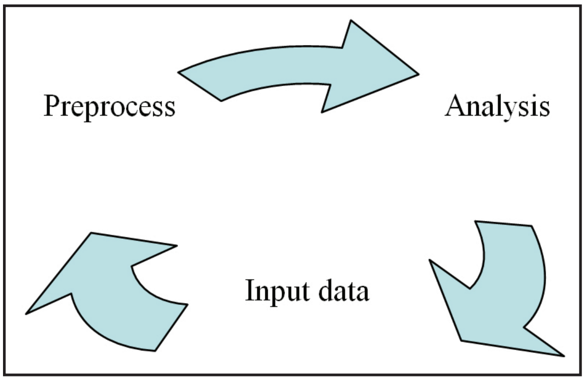

Fig. 1: Process of Data Extraction

Data mining roles: the development of discovery ideal to fit data known as data mining roles or say finctions and are of two types. Each role or function need some criteria to create one model over another role or function there are chiefly two types of model and are given below.

A. Predictive Model

It will forecast unidentified values based on the identified values. It consist of classification,time series analysis, regression. All of these are comes under supervised learning

B. Descriptive Model

It will categorize patterns in data. This descriptive model comprises clustering, association rules, sequence discovery. All of above are comes under unsupervised learning.

Data mining is needed in many areas. Daily bases many data is created in more much amount. So data mining is needed to extract useful information from large data or say huge data.Basically used for dimensionality reduction, data blast, heterogeneity of data machinery rich data.

C. Data Mining Process

Data mining process contains basically six steps and are listed below

1. formulate problem
2. collect data
3. represent data in the form of labels
4. learn the model or say predictor
5. evaluate the model
6. fine tune the model as needed

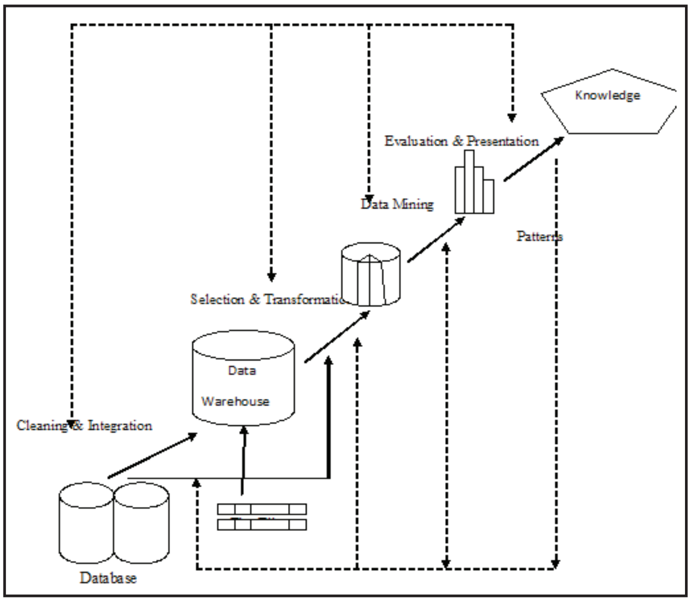

Fig. 2: Process of Data Mining

II. Clustering

Clustering is the mission of alignment objects in such a way that the objects in one group are similar to each other than those of another group. It is main task of investigative data mining and a common practice for data investigation used in many fields counting machine learning pattern recognition, image analysis, information analysis and bioinformatics.

Cluster

cluster is said to be assemblage of data object where two types of clustering resemblances are there. Within the opposite classes resemblance is objects are different to objects in other cluster. However within the class resemblance is that in which objects are more similar to same cluster. Clustering has basically methods that are given below.

A. Partitioning method
B. Density based method
C. Hierarchal method
D. Grid based method
E. Model based method

III. Literature Review

A. An Efficient Uncertain Data Point Clustering Based on Probability-Maximization Algorithm

Handling uncertain data is much more difficult. Uncertain data means that data which has no certainty in them. Data are uncertain which has no proper location. Means moving objects like persons and many living things like animals. Probability distributions are used to define uncertain data objects. Many scenarios are there that are used to define clustering of uncertain data objects. Basic examples of uncertain data objects are as marketing research and one another example is weather station monitor weather conditions. Accordingly some distributions there is need to cluster the uncertain objects.

Data uncertainty is represented with the help of probability theory. Probability density functions are there to represent an object. In this portioning method and density based clustering is used to clustering the uncertain data.Portioning method contains k-mean clustering and density based clustering contain DBSCAN that are used in this paper to handle uncertain data or say clustering of uncertain data.

In this purposed algorithm data sets are firstly preprocessed and in preprocessing step basic components are sizes,classes, attributes, standard deviations. After preprocessing step probability maximization algorithm is performed. In PM algorithm two partitions are there one is true partition and second is clustering results.

B. Handling Uncertainty and Clustering in Uncertain Data Based on KL Divergence Technique

Many problems in data are comes due to uncertainty of data. From those problems clustering is one of them. Due to this uncertainty of data there are many problems in clustering. Previously known methods i.e portioning method and density based method adopted to handle uncertain data and cluster that uncertain data into a single cluster that is the objects in one cluster are similar to one another. Portioning method and density based methods are that which use or say based on geometric distance between objects.

In purposed algorithm probability distributions are used which are the mandatory features of uncertain objects. Purposed algorithm Kullback-Liebler divergence is used. This algorithm is used to measure the similarity between objects and integrate portioning and density based method to cluster uncertain objects. FCM method is used in this to cluster uncertain data objects and show effectiveness of the data objects. FCM is as described fuzzy c-mean clustering for data with acceptance. Data objects are represented by probability distributions. And probability distributions is described by probability mass function and objects that's are in continuous distributions are described by probability density function. In this paper basically KL-divergence is integrated with portioning and density based clustering to handle uncertain data and clustering of uncertain data

C. Density Based Algorithm for Discovering Clusters in Large Spatial Database With Noise

Clustering algorithms are used for class identification purpose basically. Spatial clustering has many requirements like domain knowledge to represent the input parameters and cluster of arbitrary shapes are discovered in spatial databases but there requirements are not fulfilled by well known clustering algorithms or say these well known clustering algorithm can not fulfill the requirements of spatial databases. Spatial database means data related to space some spatial database systems are used to manage the spatial data. For identifying class clustering algorithms are used. In spatial data many requirements are there like minimum knowledge domain for determining input parameters, arbitrary shaped cluster discovery, mainly efficiency on the large databases. For clustering in spatial databases new algorithm is purposed i.e density based algorithm DBSCAN algorithm for spatial databases is used. In portioning method various partitions of data is done in many clusters like K-mean and K-mediod methods are there in portioning method. But is not useful for spatial databases due to not fulfill of the requirements of spatial databases. An algorithm CLARANS is used in clustering i.e clustering large application based on randomized search. It is efficient and more effective it is a very improved algorithm of K-mediod algorithm

But it is not much more efficient for spatial databases. DBSCAN algorithm is used in spatial databases. DBSCAN algorithm take only one parameter as a input and support user to determine an appropriate value for it.

D. Ontology –Based Access to Probabilistic Data

Ontology is used for querying probabilistic data. There are much more uncertainty in data that is in probabilistic data. Ontology is used in many applications like managing data such data which is extracted from web. Or say web data is managed by ontology approach. In this system one most important task is there i.e query rewriting in first order logic is an important tool.

Ontology based data access is an active area of logic research description. In short it is also called as OBDA i.e ontology based data access. It provides meaning or say semantics for complete data. And has capabilities that give more complete answer to the queries.In this paper extension of OBDA is presented over here that capture uncertain data with the help of probabilistic data model. It will replace the uncertain answers with the certain answers with the help of probabilistic computations that are more certain than that before.New approach relates to probabilistic database system i.e PDBMs in same that OBDA relate to RDBMS. At last it can be said that by using ontology uncertainty of data can be handled and data is managed by ontology. Means to say that ontology has many application in uncertain data management.

E. Topography of Multivariate Normal Distributions

For fitting a high dimensional data always multivariate normal distributions are used. It is explained that a topography of that in the sense of features as a density, it can be analyzed in lower dimensions by using a ridgeline manifold which contain all the critical points and ridges of that density. A evaluation of that plot on ridgeline explain the key features of the mixed density.Additionally by using that ridgeline uncovering the functions which determine the number of modes of mixed density. When there are only two components or parts are mixed with each other. Firstly follow the analysis and that give the curvature of function that can be used to prove the number of modality theorms. In this survey for understanding the topography of mixed normal distributions.There are many power full tools that has been developed. That Tools are more and more powerful tools for understanding the topology of multivariate normal mixtures model. Tools become more powerful and has more power in case of k=2 in this case problem can be reduced from D dimensions down to 1 one can plot a simple plot in investigating the key features i.e density in any problem. This can be described analytically on some certain cases.

F. A survey on Soft Subspace Clustering

Subspace clustering is a very good clustering technology to identify clusters based on there associations with subspace in high dimensional space subspace clustering can be classified into 2 different categories.Hard subspace clustering (HSC) and soft subspace clustering (SSC). HSC is been extensively studied by scientific community. SSC are relatively new but more attention on them due to better adaptability. In this paper comprehensive survey on existing SSC algorithm and recent developments use presented over here. SSC has mainly three types conventional subspace clustering, independent SSC, Extended SSC. As discussed three main categories of SSC are as listed below:

1. Conventional SSC

Conventional feature allowance clustering algorithms with all the clusters distribution the same subspace and a shared common mass or say weight.

2. Independent SSC

Multiple feature weighting clustering algorithms with all the clusters having their own weight vectors, i.e., each cluster has an independent subspace, and the weight vectors are controllable by different mechanisms

3. Extended SSC

Algorithms extending the CSSC or ISSC algorithms with new clustering mechanisms for performance enhancement

Comprehensive survey of SSC is presented over here. These are systematically categorized into three categories, XSSC, ISSC, CSSC there are explained along with subcategories are explained in detailed. It is seen that transfer learning and multi-view learning will play an important role in development of SSC in future. A thorough understanding of SSC algorithm and insight into the advancement of SSC can be obtained through this survey.

III .Conclusion

Much more amount of data produced or generated in daily life is uncertain and handling that uncertain data is quite difficult. Digital data is produced or say generated in daily life a real time example of this digital data is weather it is different in different places all over the world or say on earth. By using clustering algorithms these type of uncertain data can be handled. Density based and portioning clustering methods have been used to handle such data but there is no proper accuracy and quality of cluster is not so good while handling these type of uncertain data. There is need for further enhancement in clustering algorithms to handle digitized data .

IV. Acknowledgment

I wish to thank all people who has directly or indirectly helped me in completing this paper work successfully. I express my gratitude toward my thesis mentor and also toward the Head of Computer Science Department for there valuable suggestions and constant guideline during this paper work and also acknowledge the research work done by all researchers in field across the internet for maintaining valuable documents and resource on internet.

References

[1] RayS, LindsayBG,"The topography of multivariate normal mixtures", Ann Stat 33, pp. 2042–2065, 2005.

[2] Guo GD, Chen S, Chen LF,"Soft subspace clustering with an improved feature weight self-adjustment mechanism", Int J Mach Learn Cybe, 2011.

[3] J. Han, M. Kambler,"Data Mining: Concept and Techniques".

[4] M. Ester, H-P. Kriegel, J. Sander, X.Xu,"A Density-Based Algorithm for Discovering Clusters in Large Spatial databases with Noise", Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

[5] H.P.Kriegel, M. Pfeifle," Hierarchical Density-Based Clustering of Uncertain Data", Proc. IEEE Int'l Conf. Data Mining (ICDM), 2011.

[6] Jean Christoph Jung, Carsten Lutz University Bermen, Germany, "ontology-based access to Probabilistic Data".

[7] Martin Easter, Hans-Peter Kriegel, Jorge Sander, Xiaowei Xu,"Density Based Algorithm for Discovering Clusters in Large Databases with Noise".

[8] Jiawei Han, Micheline Kamber, Jian Pei,"Data mining Concepts and Techniques".

[9] Jiang, Jian Pei, Yufei Tao, XueminLin,"Clustering Uncertain Data Based on some time Probability Distribution Similarity", IEEE Transactions on KDE, Vol. 25, No. 4, April 2013.

[10] H. P. Kriegel, M. Pfeifle,"Hierarchical Density-Based Clustering of Uncertain Data", Proc. IEEE Int'l Conf. Data Mining (ICDM).

[11] T.Imielinski, W.L.Lipski Jr.,"Incomplete Information in relational Databases", J. ACM, Vol. 31, pp. 761-791, 1984.

[12] S. Kullback, R.A. Leibler,"On Information and Suffficiency", The Annals of Math Statistics.

[13] Kanika Lakhani, Gaurav Girdhar,"Data Mining and data Ware Housing.

[14] J.B. MacQueen,"Some Methods for Classification and Analysis of Statistics and Probability", 1967.

[15] B.W. Silverman,"Density Estimation for Statistics and Data Analysis", Chapman and Hall,1986.

[16] Ian H.Witten, Eibe Frank,"Data mining Practical machine learning tool and technique".

[17] Pang-Ning Tan, Michael SteinBach, Vipin Kumar: Data Mining Introduction.

[18] Trevor Hastie, Robert Tibshirani, Jerome Friedman, "Elements of Statistical Learning Data Mining Inference and Prediction".

[19] Gordon S. Linof, Michael J.A Berry,"Data Mining techniques".

[20] J.Xu, W.B. Croft,"Cluster-Based Language Models for Distributed Retrieval", Proc. 22nd Ann, Int'l ACM SIGIR,

1999.
[21] B.W. Silverman,"Density Estimation for Statistics and Data Analysis", Chapman and Hall, 1986.

Veerpal is student of Lovely Professional University, Jalandhar (INDIA). She has received B.Tech Degree from Malout Institute of Management and Information Technology. Her main research interest includes data mining.

Sanyam Anand is Assistant Professor in Lovely Professional University Jalandhar (INDIA) and he is pursuing P.hd from NITTTR. His Research interest includes Image Processing, Data mining and data Bases.