Clustering Of Uncertain Data

Thasnim A.
M. Tech Scholar
Department of Computer Science and Engineering
Sree Buddha College of Engineering,Pattoor
Alappuzha, India
thasnim2020@gmail.com

Lakshmi S.
Assistant Professor
Department of Computer Science and Engineering
Sree Buddha College of Engineering,Pattoor
Alappuzha, India
lakshmi.rnath@gmail.com

Abstract—Clustering is the process of grouping a set of objects into objects of similar classes. Clustering of certain data had been studied for years. But there is only preliminary research on the field of clustering of uncertain data. Most studies uses geometrical distance. Here geometrical distance along with probability is combined. For the project, camera specification is used as the data set. Different users can make survey regarding the features of the model. Based on this probability calculation is done. Here grouping is done with the help of similarity calculation. For clustering DBSCAN is used.

Keywords- clustering; probability; similarity calculation

I. INTRODUCTION

Clustering is the process of grouping items which comes under datamining. It is similar to classification in which data are arranged. However, the groups are not predefined as in classification. Instead, the grouping is accomplished by discovering similarities between data according to characteristics found in the original data. It is an important branch where many new ideas are emerging. Application comes when working with large data sets , in most scenarios it is useful to be able to separate information by dividing the data into smaller groups, and eventually, to do class identification. Different application of clustering comes in business, biology, statistics, data mining .Common uses of clustering includes an exploratory data analysis tool. In one-dimension, clustering is a good way to quantify real-valued variables into k non-uniform buckets. The technique is used on acoustic data in speech understanding to convert waveforms into one of k categories (known as vector quantization). Also used for choosing color palettes on old fashioned graphical display devices color image segmentation. Clustering can be classified as clustering of certain data as well as clustering of uncertain data. Clustering of certain data has been studied for years which includes many techniques as pattern recognition, machine learning and so on. But there is only a few research that is under clustering of certain data. Different clustering algorithms are used. By clustering those items that shows similar property will be belonging to the same cluster. Thereby size of large data set will be reduced.

II. BACKGROUND

Different techniques for clustering had been used so far. In hierarchical clustering, clusters are created iteratively, using clusters created in previous step. In partitional clustering, single partition is created, there by minimizing cost function. In the paper, efficient clustering of uncertain data[1],the data object is represented by an uncertainty region over which a probability density function (pdf) is defined. The paper uses uk-means algorithm. In uk-means, an object is assigned to the cluster whose representative has the smallest expected distance to the object. This paper explained why expected distance computations are expensive and thus argued that effective pruning techniques are necessary for a computationally feasible clustering algorithm. In subspace clustering for uncertain data analyzing [3], uncertain databases is a challenge in data mining research. Usually, data mining methods rely on precise values. In scenarios where uncertain values arise these algorithms cannot deliver high quality patterns. In this paper,a method for subspace clustering for uncertain data that delivers high quality patterns was used because in uncertain scenarios a strict assignment of objects to single clusters is not appropriate, the model is enriched with the concept of membership degree. Major drawback of subspace clustering for uncertain data is computationally expensive. In clustering of uncertain data objects using improved k-means algorithm [4][2],the uncertainty araises in a information because of the imprecise measurement of the results. Here for clustering indexing techniques are applied to the k-means algorithm then the cluster generation time is drastically concentrated and the clustering will be done more clearly. The paper proposes an approach which will minimizes the computation time for clustering uncertain data. Major drawback of the paper is that cluster elements of one group are overlapped with another group and also the clusters that are generated are not proper. In Dbscan With Obstacle Constraints [5],spatial clustering has been used. The clustering algorithms that are presently used ignores the existence of many constraints in the real world. This paper proposes a new spatial clustering algorithm with obstacles constraints (PSODBSCAN)[6].This method integrates Parti-

IJERT

cle Swarm Optimization (PSO) global optimization ability with local search features of DBSCAN algorithm. The algorithm reduces computational amount to a large extent, thereby improving execution speed of the algorithm. Major drawback is the complexity of doing in higher dimension exist[8].

Motivation

III. MOTIVATION

Clustering is the process of combination of a set of physical objects into program of similar objects, it is similar to organization in that data are grouped. Basically there are two types of clustering ,clustering of certain data and clustering of uncertain data. Clustering of certain data had been studied for years which include machine learning ,pattern recognition, bio informatics and so on. But there is only preliminary research on the field of uncertain data. Data uncertainity brings new challenges to clustering since clustering uncertain data demands a measurement of connection between uncertain data objects. Most studies of uncertain data uses geometric distance based on similarity measure and only exact value is obtained. Here similarity calculation is done which can be used as the basis of clustering. So by this project we are aiming to find uncertainity along with intermediate result so that it can be used for prediction and statistical purpose.

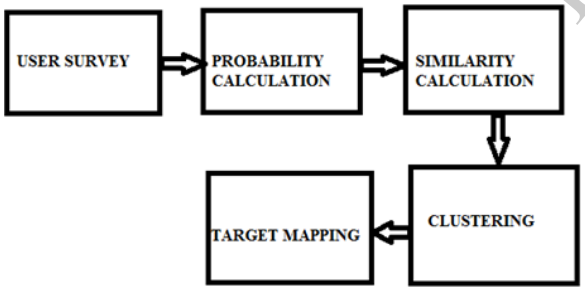

Figure 1. Overall Design

The project can be divided into 3 different modules. Similarity calculation module, clustering module and target mapping module. For similarity calculation first user survey is done, where user can rate different camera models that are presently available based on their knowledge. Probability calculation acts as the basis for similarity calculation where probability is found by comparing score given by the user to each camera models. Based on this similarity calculation is done. This is a significant method for clustering. For clustering Dbscan is used.

A. Similarity Calculation

In similarity calculation there are different sub modules. In admin module, admin can add all new cameras presently available. Features of the camera specified by the owner can be displayed in this page. User can search for their preferences in search module. All those models with in the restriction given by the user will be displayed. Next is the User Survey where users can make their survey regarding the models that they use. Different grades can be assigned by the user.

Figure 2. User Survey

In this project ,based on the survey of each feature of the camera grades will be assigned. Then total score for particular model is calculated.

$$\sum_{i=1}^{n}(xi)/total\_score$$

Probability is found by comparing each user's score with respect to total score for that particular model. Based on this probability, similarity calculation is done.In probability theory and information theory, the kullback-leibler (KL) divergence is a non-symmetric measure of the divergence between two probability distributions p and q. Specifically, the kullback-leibler divergence of q from p, denoted dkl(p), is a measure of the information lost when q is used to approximate KL measures the expected number of extra bits required to code samples from p when using a code based on q, rather than using a code based on p. Typically P represents the true distribution of data, observations, or a precisely calculated theoretical distribution. The measure Q typically represents a theory, model, description, or approximation of P. KL divergence is a special case of a broader class of divergences called f-divergences. It was originally introduced by Solomon Kullback and Richard Leibler in 1951 as the directed deviation between two distributions. The Kullback Leibler distance (KL-distance) is a natural distance function from a probability distribution, p, to a target probability distribution,q. It can be interpreted as the expected extra message-length per data due to using a code based on

IJERT

the incorrect (target) distribution compared to using a code based on the true allocation[7]. For discrete probability distributions, p=p1, ..., pn and q=q1, ..., qn, the KL-distance is defined to be

$$\mathrm{K\,L}(p;\,q) = \sum p(i) \log \frac{p(i)}{q(i)}$$

For continuous probability densities, the sum is replaced by an integral.

$$\mathrm{K\,L}(p;\,p) = 0$$

$$\mathrm{K\,L}(p;\,q) \geq 0$$

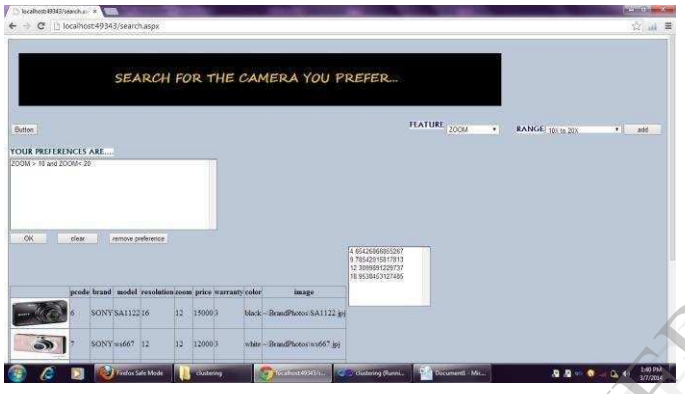

Figure 3. Search Module

This divergence is finite whenever p is absolutely continuous with respect to q and it is only zero if p = q. The KL divergence is central to information theory and statistics. Mutual information measures the information one random variable contains about a related random variable and it can be computed as a special case of the KL divergence. Now based on this similarity calculation, groups are to be formed. This group can be clustered based on any technique for clustering. Finally target mapping have to be done ,where the data is mapped to the data set to which it belongs to.

B. Clustering

Cluster analysis is a principal method for database mining. It is either used as a stand-alone tool to get insight into the division of a data set. Density-based approaches is applied to a confined cluster criterion. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by noise. These regions may have an random shape and the points within a region may be arbitrarily spread. However, using clustering algorithms involves many problems. It can often be difficult to know which input parameters that should be used for a specific database, if the user does not have sufficient knowledge of the domain. Furthermore, spatial data sets can contain large amounts of data, and trying to find cluster patterns in many dimensions is computationally costly. Always short computing time is favourable. The shapes of the clusters can be arbitrary and in most cases it will be very complex. Finding these shapes can be very burdensome. The DBSCAN algorithm can identify clusters in large spatial data sets by looking at the local density of database elements, using only one input parameter. Furthermore, the user gets a suggestion on which parameter value that would be suitable. Therefore, minimal knowledge of the domain is required. The DBSCAN can also determine what information should be classified as noise. In spite of this, its working process is quick and scales very well with the size of the database approximately linearly. By using the density distribution of nodes in the database, DBSCAN can categorize these nodes into separate clusters that define the different classes. DBSCAN can find clusters of arbitrary shape. However, clusters that lie close to each other tend to belong to the same class. Density-based clustering algorithms are algorithms that aim to discover areas of high density that are separated from each other by regions of low-density. In this approach density is anticipated for a particular point in a data set by counting the number of points within a certain radius. The size of the radius is crucial because if the radius is too large then all points in the data set will have identical density and if the radius is too small then the density of each data point will be 1. Data points are classified as core points, that have more than a specified number of points (MinPts) within the chosen radius (E). Graphically speaking, these are points that are in the interior of a cluster. Border point is a point that has fewer than MinPts within the radius (E) but is still in the neighborhood of a core point.Noise point is any point that is neither a core point nor a border point. Basically there are two types as density reachable and density connected. In Density-reachable, A point p is density-reachable from a point q if there is a chain of points p1, , pn, p1 = q, pn = p such that pi+1 is directly density-reachable from pi. In Density-connected:A point p is density-connected to a point q if there is a point o such that both, p and q are density-reachable from o.Clustering can done by selecting minimum number of points needed to form a cluster. Radius is set so that all those points with in that radius acts as a cluster. Each point in the cluster is evaluated to see whether there is minimum number of points to form a cluster ,if so another cluster is formed [5]. A threshold will be set to find the clustered data set. If the threshold is low then the accuracy will be high. Graphical representation can be used to represent the result. Finally data item corresponding to the actual cluster will be printed.

C. Target Mapping

In target mapping, data items will be mapped to the cluster to which it belongs to.

IJERT

IV. Expected Result

Finally as the output models with in the restriction given by the user will be displayed. That is a cluster will be formed. The output is based on the similarity value that is obtained. Graphical representation can be used to show different clusters with similar property.

V. Conclusion

Our work mainly concentrate on clustering. Basically there are two types of clustering ,clustering of certain data and clustering of uncertain data. There is only preliminary research on the field of clustering of uncertain data. Here for clustering of uncertain data geometrical distance along with probability is combined. Based on this score given by the user probability of each data set is found. After this similarity is found by comparing each users evaluation of a particular model with respect to other models. Thereby intermediate result is generated which can be used for prediction or statistical purpose. DBSCAN is the clustering technique that is used. Some measures must be taken so that it work properly in higher dimensions.

References

[1] Wang Kay Ngai, Ben Kao and Chun Kit Chu," Efficient Clustering Of Uncertain Data",Proceedings of sixth International Conference on Data mining 2006.

[2] Andrew McGregor, Graham Cormode,"Clustering of Uncertain data objects using K-Means Partitional Clustering",ACM 2008.

[3] Stephan Gunnemann ,Hardy Kremer ,Thomas Seid,"Subspace Clustering for Uncertain Data",ACM 2010.

[4] Prof. Mangesh Wanjari, Samir N. Ajani,"Clustering of Uncertain Data Objects using Improved K-means Algorithm", The international Journal of Advanced Research in Computer Science and Software Engineering 2013.

[5] Ying Wang,Guisheng Yin,"Dbscan With Obstacle Constraints",Journal of Theoretical and Applied Information Technology 2005.

[6] Pan,Donghua,Zhao,Lilei,"Uncertain data cluster based on DBSCAN",The international Conference on Multimedia Technology,pp.3781-3784,2011

[7] Rani Nelken and Stuart M. Shieber,"Computing The Kullback-Leibler Divergence Between Probabilistic Automata Using Rational Kernels",Engineering and Applied Sciences Harvard University Cambridge, MA 02138 ,2006

[8] H. Xu and G. Li, "Density-based probabilistic clustering of uncertain data",CSSE, 4:474477,2008

IJERT