CLUSTERING UNCERTAIN DATA BASED ON PROBABLITY DISTRIBUTION SIMILARITY

S.Athirayan[1], K.Rajasekar[2]

[1]*Assistant Professor, Department of Computer Science and Engineering,*
Pandian Saraswathi Yadav Engineering College, Sivagangai, Tamilnadu, India.

[2]*Assistant Professor,Department of Computer Science and Engineering,*
Pandian Saraswathi Yadav Engineering College, Sivagangai, Tamilnadu, India.

Abstract

This paper introduces uncertain data management, data records are typically represented by probability distributions rather than deterministic values. Usually, Clustering can be considered the most important unsupervised learning *problem; so, as every other problem of this kind, it deals with finding a* structure *in a collection of unlabeled data. The problem of clustering uncertain objects according to probability distributions happens. Previous work regarding clustering uncertain data had more complexities to handle the objects. So the proposed system introduced KL divergence for measuring similarity between two kinds of methods values. Here, the main concept of process is to calculate the probability distribution. According to the distribution value which moves on the similarity measurement and also the clustering process. For clustering k-means algorithm is used to find the distances between the objects. Finally the results of proposed clustering are represented in clustering graph.*

Introduction

1.1 Scope

The main scope of this project is to calculate the probability distribution. According to the distribution value it is to be moved on the similarity measurement and also the clustering process. For clustering, have to use k-means clustering to find the distances between the objects. Finally it has to produce the clustering results with the help of clustering graph based view.

1.2 Synopsis

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. In this case, it is easily identify the four clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures. So, the goal of clustering is to determine the intrinsic grouping in a set

of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, it could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

The first process of the project is to load the dataset into the process. In that dataset contains weather related information's and it can be separated by any deli meter. So after selecting the dataset, have to give the deli meter for open the dataset. The opened dataset contains the details of month, region, maximum temperature, minimum temperature and rain fall. After that, have to do the process of feature extraction. In feature extraction, have to know the full details regarding to the weather status. The details may be of date and month, maximum temperature, minimum temperature, rain fall and mean temperature. All those details can be extracted in this process.

After finishing the feature extraction, have to move on to the probability distribution. Here, probability was found for all selected features. According to the features, have to find the probability. Here, the probability can be dividing into two groups. They are maximum temperature and minimum temperature. The value of the probability is a main thing to cluster the temperature. The maximum temperature can have some probability and minimum temperature can have some probability. If it finds the probability distribution of the given data's after it is to be found the similarity between the two data's after that the clustering function is performed. For finding the similarity between the data the KL divergence algorithm is used. The KL divergence algorithm easily finds the similarity between the distributions of the data sets.

In this proposed concept the K-means clustering technique is used to create or generate the cluster. Cluster means grouping of data's. Clustering the uncertain data is more complex technique but in our concept the cluster is generated for uncertain data. In this proposed concept weather report had been taken as an uncertain data set. From that data the cluster is formed. If it made the cluster for the uncertain data set now the graph is constructed between these two data sets. The generation process is used to show the difference between the formations of cluster between the two data sets. The graph is constructed based upon the time and category of the data sets. The graph will show the both maximum temperature and minimum temperature data results. In our implementation side also the graph result is shown.

If it loads the data using delimiter after that calculate the probability between the data's and then find the similarity between the data using the KL divergence algorithm and finally form the cluster and generate the graph for the data set the cluster is tested. To find the efficiency of the clustering method the test function is performed. The Gauss method is used to test the cluster formation. The results show that the cluster formation is more efficient when compared to the other existing techniques.

In probability theory and information theory, the Kullback–Leibler divergence (also information divergence, information gain, relative entropy, or KLIC) is a non-symmetric measure of the difference between two probability distributions P and Q. Specifically, the Kullback–Leibler divergence of Q from P, denoted $D_{\mathrm{KL}}(P||Q)$, is a measure of the information lost when Q is used to approximate P: 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*.

Although it is often intuited as a metric or distance, the KL divergence is not a true metric — for example, it is not symmetric: the KL from *P* to *Q* is generally not the same as the KL from *Q* to *P*. However, its infinitesimal form, specifically its Hessian, is a metric tensor: it is the Fisher information metric.

For discrete probability distributions *P* and *Q*, the K–L divergence of *Q* from *P* is defined to be

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i).$$

In words, it is the expectation of the logarithmic difference between the probabilities *P* and *Q*, where the expectation is taken using the probabilities *P*. The K–L divergence is only defined if *P* and *Q* both sum to 1 and if $Q(i) = 0$ implies $P(i) = 0$ for all i (absolute continuity). If the quantity $0 \ln 0$ appears in the formula, it is interpreted as zero because $\lim_{x \to 0} x \ln(x) = 0$.

For distributions *P* and *Q* of a continuous random variable, KL-divergence is defined to be the integral:

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x,$$

Where *p* and *q* denote the densities of *P* and *Q*.

More generally, if *P* and *Q* are probability measures over a set *X*, and *P* is absolutely continuous with respect to *Q*, then the Kullback–Leibler divergence from *P* to *Q* is defined as

$$D_{\mathrm{KL}}(P\|Q) = \int_X \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P,$$

Where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon–Nikodym derivative of *P* with respect to *Q*, and provided the expression on the right-hand side exists. Equivalently, this can be written as

$$D_{\mathrm{KL}}(P\|Q) = \int_X \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q,$$

which recognizes as the entropy of *P* relative to *Q*. Continuing in this case, if μ is any measure on *X* for which $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ exist, then th e Kullback–Leibler divergence from *P* to *Q* is given as

$$D_{\mathrm{KL}}(P\|Q) = \int_X p \ln \frac{p}{q}\,\mathrm{d}\mu.$$

The logarithms in these formulae are taken to base 2 if information is measured in units of bits, or to base *e* if information is measured in nats. Most formulas involving the KL divergence hold irrespective of log base.

2 Problem Statement

2.1 Feature Extraction

The first process of our project is to load the dataset into the process. In that dataset contains weather related information's and it can be separated by any deli meter. So after selecting the dataset, have to give the deli meter for open the dataset. The opened dataset contains the details of month, region, maximum temperature, minimum temperature and rain fall. After that, have to do the process of feature extraction. In feature extraction, have to know the full details regarding to the weather status. The details may be of date and month, maximum temperature, minimum temperature, rain fall and mean temperature. All those details can be extracted in this process.

2.2 Probability Distribution

After finishing the feature extraction, have to move on to the probability distribution. Here, probability was found for all selected features. According to the feature, have to find the probability. Here, the probability can be dividing into two groups. They are maximum temperature and minimum temperature. The value of the probability is a main thing to cluster the temperature. The maximum temperature can have some probability and minimum temperature can have some probability.

2.3 Clustering

If it finds the probability distribution of the given data's after that it finds the similarity between the two data's after that the clustering functions performed. For finding the similarity between the data the KL divergence algorithm is used. The KL divergence algorithm easily finds the similarity between the distributions of the data sets. In this proposed concept the K-means clustering technique is used to create or generate the cluster. Cluster means grouping of data's. Clustering the uncertain data is more complex technique but in this concept the cluster for uncertain data is generated. In this proposed concept the weather report had been taken as an uncertain data set. From that data the cluster is formed.

2.4 Graph Generation

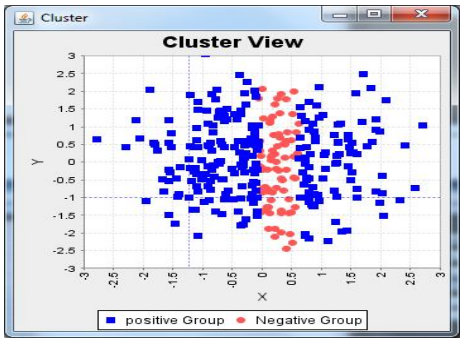
If it made the cluster for the uncertain data set now it is constructed the graph between these two data sets. The generation process is used to show the difference between the formations of cluster between the two data sets. It is constructed the graph based upon the time and category of the data sets. The graph will show the both maximum temperature and minimum temperature data results.

In this implementation side also the graph result is showed.

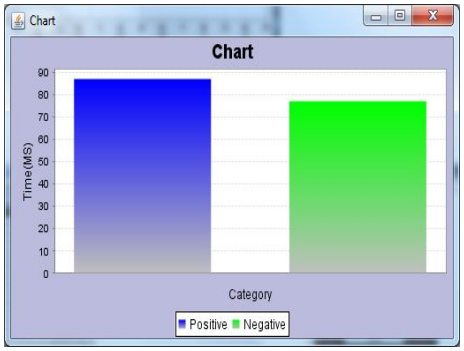
2.5 Cluster Test

This is our final stage, if it is load the data using delimiter after that calculate the probability between the data's and then find the similarity between the data using the KL divergence algorithm and finally form the cluster and generate the graph for the data set the cluster is tested. To find the efficiency of the clustering

method the test function is performed. The Gauss method is used to test the cluster formation. The results show that this cluster formation is more efficient when compared to the other existing techniques



(i) Cluster view



(ii) Graph Construction

3 Results and Discussion

In this project k-means is used for clustering the minimum and maximum temperature of the particular regions .To improve the efficiency of clustering process; have to introduce a new algorithm named as agglomerative clustering. Using that algorithm, have to improve the efficiency of clustering process.

4 Conclusions

To overcome the problems of handling the uncertain amount of data the concept has been implemented. In this proposed concept first, load the data into the system using the delimiters. In this implementation side the weather report was considered the input data. The weather report contains the temperature, rain fall, minimum and maximum temperature information's. These are all taken as an input. After that the probability distribution function is performed for each data's. The similarity is found between these data's using the KL divergence method in this proposed concept used. Once the similarity is found between this data set the data's are separated in very easy manner. Using the KL divergence algorithm to find the similarity and also find the difference between the each data's in the data set. The experimental result shows that our method is providing best result when compared to the existing system for clustering the uncertain data sets.

5 References

1. Banerjee.A, Mergu.S and Dhillon.I.S "An Efficient Image Similarity Measure based on Approximations of KL-Divergence between Two Gaussian Mixtures", Proceeding Intel Conference Very Large Data Bases (VLDB), 2007.

2. Kriegel.H.P and Pfeifle.M "Turning big data into tiny data: Constant-size core sets for k-means, PCA and projective clustering." Proceeding Intel Conference Data Engineering (ICDE), 2006.

3. Dalvi.N.N and Suciu.D, "Management of Probabilistic Data: Foundations and Challenges," Proceeding ACM SIGMOD-SIGACTSIGART Symposium Principles of Database Systems (PODS), 2007.

4. Cheng.R, Kalashnikov.D.V, and Prabhakar.S, "Evaluating Probabilistic Queries over Imprecise Data," Proceeding ACM SIGMOD Intel Conference.Management of Data (SIGMOD), 2003.

5. Ngai.W.K, Kao.B, Chui.C.K, Cheng.R, Chau.M, and Yip.K.Y, "Efficient Clustering of Uncertain Data," Proceeding Sixth Intel Conference Data Mining (ICDM), 2006.

6. Tao.Y, Cheng.R, Xiao.X, Ngai.W.K, Kao.B, and Prabhakar.S, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions," Proceeding Intel Conference Very Large Data Bases (VLDB), 2005.

7. Volk.P.B, Rosenthal.F, Hahmann.M, Habich.D, and Lehner.W, "Clustering Uncertain Data with Possible Worlds," Proceeding IEEE Intel Conference Data Engineering (ICDE), 2009.

8. Ackermann.M.R, Blo¨mer.J, and Sohler.C, "Clustering for Metric and Non-Metric Distance Measures," Proceeding ACM-SIAM Symposium Discrete Algorithms (SODA), 2008.