International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

An Effective Approach to Compute Distance of Uncertain Data by Using K-Nearest Neighbor

A. Jayasri[1], Dr. M. B. Mukesh Krishnan[2]

[1, 2]SRM University, Chennai, India

Abstract: ***Generally, an object of uncertain data could be presented through a probability distribution. Now days the Clustering uncertain data have been determined as a very important issue. In existing system, technique of Kullback-Leibler Divergence is mostly used by information theory in order to calculate the similarity between certain data. This research work based over calculation of Probability mass function, whereas uncertain data values of discrete and continuous are calculated. With use of the probability mass function, the cases of continuous and discrete distance value are individually measured. The probabilistic ratio of continuous and discrete distance is used to determine the similarity between certain data. For clustering the uncertain data by performing the techniques of Density based clustering. Therefore, the main drawback in the existing system is to selecting the nearest neighbor. To overcome from this type of problem, we introduced algorithm of K-nearest-neighbor in our proposed work to determine the nearest neighbor. The algorithm of K-nearest-neighbor use to calculate the distance among scenarios set and a query scenario in the data set. Here, distance is measure for both the cases of discrete and continuous through the use of using Probability mass function. After that the algorithm of KNN is used to measure the nearest neighbor. Hence, our proposed works produce an effective result and overcomes the drawback of existing technique.***

Keywords: KNN, K-L Divergence, Uncertain Data

1. Introduction

The clustering uncertain data for similarity among data has been standardized as very important problem [21], [22]. Normally, an object of uncertain data is used to represent through a probability distribution [23], [24]. According to the probability distributions function the clustering uncertain objects problem happens in many scenarios. The algorithm of K-nearest-neighbor is very important for calculating text categorization [8]. Various researchers have determine that the algorithm of K-nearest-neighbor is accomplished a very better performance in the experiments on dissimilar data sets [9][10][11]. The concept behind the algorithm of K-nearest-neighbor is relatively straightforward. To categorize a latest document, the system has to finds the k nearest neighbors between the training documents along with uses the k nearest neighbors' categories in manner to weight the classify candidates [8]. The main disadvantage of K-nearest-neighbor algorithm is its efficiency, since it essential to compare a test document by the entire samples are in the training set. Generally, the performance of KNN algorithm depend over two factors:
1. A proper value for the k parameter.
2. An appropriate similarity function.

One of the most central problems in dealing of information system by text data is classification. According to the contents, text categorization is a supervised knowledge task of transfer natural language text documents to one or more predefined classes or categories. Through the fast increase of electronic text documents over the corporate intranets and Internet, since a potential apparatus for better managing finding and filtering these sources, text categorization has gained very much attention in current years. [23] Text Categorization is very essential component of several big Machine Learning system or Information Retrieval and it is describe as the content- based over one or more predefined assignment categories to texts. Generally, it is conjectured that it is infeasible to manually categorize the entire the latest documents which are jointed to a system in a timely order. Thus, document classification process of automatic technique is essential in the clustering uncertain data for similarity among data process. Information processing requirements have improved by the increase of textual information sources, like as the Worldwide Web and news media. Routing or finding of texts retrieval System, in order to interest profiles or arbitrary queries. Text categorization could be used to execute document filtering, information extraction and also routing to mechanisms of topic-specific processing or to support Information Retrieval.

We proposed a technique by the use of algorithm of K-nearest-neighbor to identify a proper value for the k parameter along with an appropriate similarity function. Along with by the KNN algorithm calculate the distance among scenarios set and a query scenario in the data set and also measure the distance for discrete and continuous using probability mass function. Hence our proposed system, we implemented an approach to finds the k nearest neighbors between data and nodes by the use of these techniques:
1. K-nearest-neighbor techniques.
2. Probability mass function

The rest of the paper is organized as follows. Section II which overviews the related work. Then we provide the detailed description of our proposed scheme in Section III. Section IV gives the results and discussions of our proposed scheme. Finally, Section V concludes the whole research work.

2. Related Work

Systems of DB for uncertain data include the focus of managing, querying and storing data annotated by uncertainty. Various dissimilar techniques have been implemented to optimize the processing of query and storage

of data. Uncertainties of data bring latest challenges in the field of clustering, but clustering of uncertain data requires a similarity measurement among object of uncertain data. The most important idea of the various systems is suggested by [1], [2]. Even as the common concepts is the storing of query technique along with uncertain data on this data, but the complete concepts dissimilar very significantly. Basically, the most important dissimilar between the approaches lies, is the ability of storing various kind of uncertainty and the storing of uncertainty definition. For instance, in the researched work of [1] discussed that the system of Trio is the single approach for nominal uncertainty into several forms, whereas in the survey of [1], [2], [28] describes the Orion system and the Mystic use many complete models, while the storing of uncertainty data is more flexible. In the research work of [10], [4], [15] discussed that mainly focused over several models of data for perfectly capturing the scene of moving objects. In this research work, aim of query algorithms is to decreasing the quantity of data transmission for make sure about the data values accuracy. In the survey of Cheng et al. [4] discussed that they are the first to invent retrieval of uncertain data in common domains and together they represent query types taxonomy by the strategies of corresponding processing. In the research of [2], [5] proposed an algorithm of I/O efficient for nearest neighbor data search. Where in the above research no one is considers the retrieval of prob-range.

In the research of Cheng et al. [6] proposed various solution for queries of prob-range but only for target 1D space. Cheng et al. argue that search range of uncertain DB is essentially very dissimilar comparing to objects of traditional precise and maintain their claims with providing the two methods of theoretical with the intention of achieving the performance of asymptotically optimal. However, the practicability of the two methods of theoretical is incomplete since (i) It can incur huge actual implementation because of the secreted constants value in their difficulty assurance. (ii) It can't support objects by arbitrary pdfs (e.g., one technique goal only identical pdfs). In the research of Suciu and Dalvi [26] explained the "probabilistic databases", in which every set is similar as a tuple into a conservative DB, apart from that it is related by an "existential" probability. For example, mean of 60% existential probability i.e. a tuple cannot exist into the DB with the chance of 40%, but if a tuple does then its value are accurate. Querying and modeling data have concerned significant attention from research community of DB [4], [22], [12]. In research work of [4], [19] discussed that currently, most of works have been occupied by uncertain data in sensor databases in query processing and management particularly in the databases of spatial-temporal [8], [11], [26], [17]. In the research of Cheng et al. [9] develop a wide probabilistic queries classification in order to uncertain data along with developed a methods for assessing queries of probabilistic. The queries of probabilistic range are first investigated by Cheng et al. [8] and developed a structure of two auxiliary indexes for supporting interval efficiency of uncertain querying. In the survey of Tao et al. [4], [26] researched queries of probabilistic range over space of multi-dimensional by the function of arbitrary probability density and recognized along with formulated various pruning rules, as well as developed a latest access technique for optimize both CPU time and I/O cost. Thus, probabilistic DB is dissimilar from uncertain DB, where every object certainly exists other than its real values follow function of probabilistic distribution.

3. Proposed Work

3.1. Overview

In our proposed work, we implemented an approach to determine a clustered uncertain data. Firstly, the probability mass function use to measured the uncertainty of data for discrete and continuous through the use of using and determine the compute the similarity. After that algorithm of KNN is used to measure the nearest neighbor and use to calculate the distance among scenarios set and a query scenario in the data set. Hereafter it computes the distance of nearest neighbor. Finally, user obtains the clustered uncertain data.

3.2 Architecture

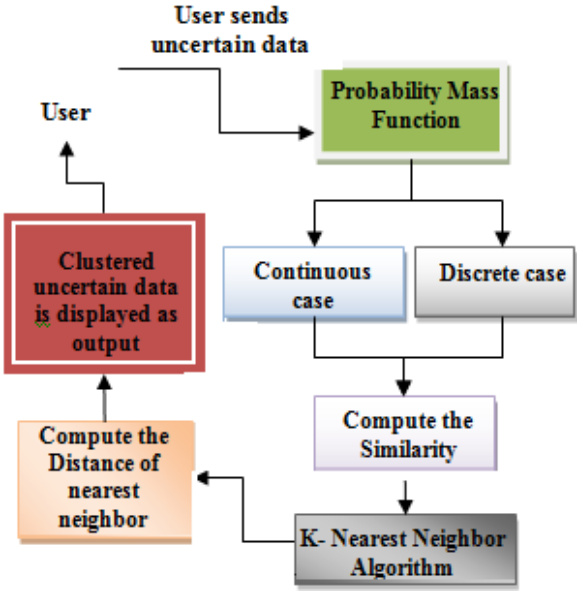

Figure 1: Overall Architecture

3.3. KNN Approach

3.3.1. Improved KNN Classifier

Algorithm of K-nearest-neighbor use neighbors node for determining their class. K-nearest-neighbor labels for the node are selected among the entire labels. And lastly, the class which having highest neighbors number is wins. Unbalanced training nodes distribution in various classifiers affects this category. Classes having most number of training nodes have most number of probabilities to win. By the use the subsequent coefficient for adjust the probability among categories by various samples numbers:

$$G_j = \frac{N}{S_j * C}$$

Where, S_j = samples number in j^{th} class, N, (d_1, d_2, d_3,…., d_N) is total of the training samples, C= classes number. Finally, the uses following formulas to determine sample probability X belong to every class:

$$P(D,S_j) = G_j \sum_{i=1}^{N} SIM(D, d_i) \cdot y(d_i, S_j)$$

Where,

$$Y(d_i,L_j) = \begin{cases} 1, d_i \in S_j \\ 0, Otherwise \end{cases}$$

$$SIM\ (D,\ d_i) = Dice\ (D,\ d_i) = \frac{2\ D\ |D \cap d_i|}{|D| + |d_i|}$$

We calculate the test sample D, SIM (D, d_i) and same node d_i using Dice similarity measure.

Where, $|D \cap d_i|$ = the common number N-grams among training sample and test sample D. Sample D belongs to the class of the biggest P(D, S_j).

3.3.2. Relative Entropy or Kullback-Leibler Divergence

Here, the Kullback-Leibler (KL) divergence estimator [1] is represented as a similar method. Given two random variables m and n by the functions of probability density t(m) along with f(n), the Kullback-Leibler divergence is describes following:

$$D\ (t||f) = -E_p\ [\log\ d_t/d_f]$$

If f is not completely continuous regarding t, then $D(t||f) = \infty$.

Divergence is finite, if t(m) is complete continuous regarding f(m), and if, zero than t(m) = f(m). The property of KL divergence is scale invariant. Particularly, RVs u = αm and v = αn (α>0) by PDFs a(m) and b(m), after that the KL divergence is given by D(a||b) = D(m||n).

For our approach, it is valuable to write the KL divergence as

$$D(a||b) = -Ha(m) - \int a(nm)\ \log 2\ b(m)dx$$

Finally, the both Nodes support is controlled to the interval [−1/2, 1/2] and taking in account.

3.3.3. Computing the distance functions

The Distance accurate computation is inflexible, as it calculating the distance among nodes and executing a point-to-point shortest-path. An ordinary way to rise above the computing intractability the Distance is to estimate it using sampling. The concept is to (e) sample r likely according to P, along with (ii) calculate the shortest-part distances on the nodes.

3.3.4. Clustering Using KL-Divergence as a Similarity Measure

An uncertain object of Clustering according to the correspondence among their probability distributions generate in several scenarios. In the theory of information, the similarity among two distributions could be determined through the Kullback-Leibler divergence. The distribution dissimilarity cannot be identifying through geometric distances. Uncertain objects are measured by random variables through certain distributions and both the continuous cases as well as the discrete case are considered. Consideration of an uncertain object is by a random variable subsequent a probability distribution into domain D.

Uncertain objects are capable of any continuous or discrete distribution. In the case of discrete, the domains have a value of finite number, such as the camera rating can take a value into {1, 2, 3, 4, and 5}. In the case of continuous, the domain has values of continuous range. Kullback-Leibler divergence is determined to calculate the similarity among the two distributions [11].

3.3.5. Algorithm

Input: Probabilistic graph G = (V,E, P,W), node s ∈ V , number of samples r, number k, distance increment γ

Ouput: Tk, a result set of k nodes for the k-NN query

1: Tk ← ∅; D ← 0
2: Initiate r executions of K NN from s
3: while |Tk| < k do
4: D ← D + γ
5: for i ← 1 : r do
6: Continue visiting nodes in the i-th execution of KNN until reaching distance D
7: For each node t ∈ V visited update the distribution $\tilde{p}D_{,s,t}$ {Create the distribution $\tilde{p}D_{,s,t}$ if t has never been visited before}
8: end for
9: for all nodes t ∉ Tk for which $\tilde{p}D_{,s,t}$ exists do
10: if median($\tilde{p}D_{,s,t}$) < D then
11: $T_k \leftarrow T_k \cup \{t\}$
12: end if
13: end for
14: end while

4. Result and Discussion

In this section, the result and discussion are presented that represents the evaluation of the proposed work. It consist the performance parameters of the K-L Divergence with KNN approach which provide the better result in compare to the alone K-L Divergence.

The measurement attribute for performing our proposed approach, we conducted some of the experiment and extract the dataset which was done on the following configuration and our work implementation will be on followed configuration:

1) Windows 7,
2) Intel Pentium(R),
3) CPU G2020 and
4) Processer speed 2.90 GHz.

4.1 Measurement of the Value for number of Cluster for Uncertain Data

Table 1: Measurement of the Value for number of Cluster for Uncertain Data

Number of Cluster	K-L Divergence	K-L Divergence with KNN
5	0.59	0.98
6	0.62	0.92
7	0.64	0.75
11	0.59	0.91
15	0.61	0.88
16	0.58	0.78
18	0.54	0.79

In the above mentioned Table.1, it is representing the measurement of value for number of cluster for the uncertain data. This result is showing the cluster value that having the

more accuracy in compare to the K-L divergence, our proposed technique is better than the existing technique.

4.2 Measurement of the Value for number of Features for Uncertain Data

Table 2: Measurement of the Value for number of Features for Uncertain Data

Number of Features	K-L Divergence	K-L Divergence with KNN
1	0.61	0.81
2	0.59	0.82
3	0.62	0.86
4	0.63	0.83
5	0.61	0.78

In the above Table.2, it is representing the value of feature for the uncertain data; here our proposed technique K-L Divergence with KNN approach is better than the simple K-L Divergence. The performance of the proposed approach is producing a good result.

4.3 Performance Metrics of the K-L Divergence with KNN and K-L Divergence

Table 3: Performance Metrics of the K-L Divergence with KNN and K-L Divergence

Techniques	Cluster Performance	Time	Number of Dataset
K-L Divergence	81.6	71 ms	800
K-L Divergence with KNN	93.2	77 ms	800

The Table.3 is representing the Cluster performance, time and the Number of dataset. It showing the comparison between the K-L divergence and K-L divergence with KNN based on the metrics Time, Number of Dataset and the cluster performance.

4.4 Time Efficiency with Dataset

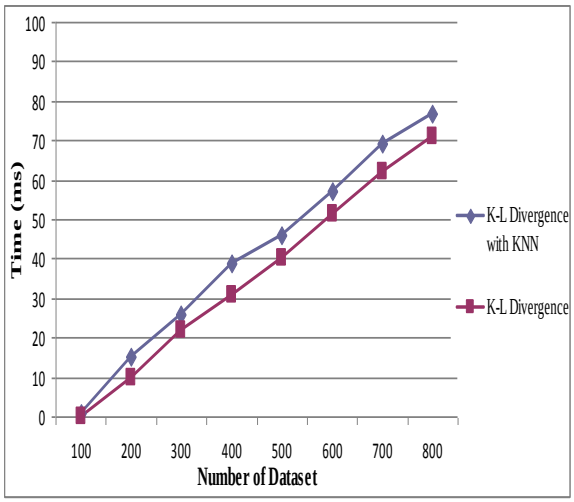

Figure 2: Time Efficiency with Dataset

The above fig.2 is representing the time efficiency with several dataset, where it showing the Time taken is more in our proposed work K-L Divergence with KNN but still producing more efficiency to producing the expected result.

4.5 Cluster Performance

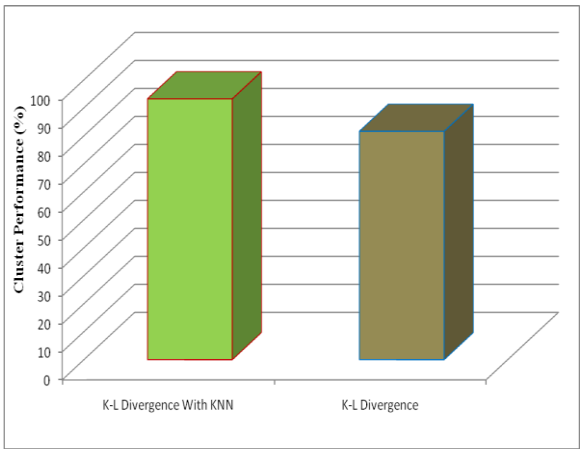

Figure 3: Cluster Performance

The fig.3, represent the cluster performance which is better than the existing technique K-L Divergence. Our proposed technique K-L Divergence with KNN is producing more accuracy and the efficiency in clustering.

5. Conclusion

In this paper, we explore clustering uncertain data based on the similarity between their distributions. We advocate using the Kullback-Leibler divergence as the similarity measurement, and systematically define the KL divergence between objects in both the continuous and discrete cases. We integrated KL divergence into the partitioning and KNN clustering methods to demonstrate the effectiveness of clustering using KL divergence. To tackle the computational challenge in the continuous case, we estimate KL divergence by kernel density estimation and employ the fast Gauss transform technique to further speed up the computation. The extensive experiments confirm that our methods are effective and efficient. The most important contribution of this paper is to introduce distribution difference as the similarity measure for uncertain data. Besides clustering, similarity is also of fundamental significance to many other applications, such as nearest neighbor search. In the future, we will study those problems on uncertain data based on distribution similarity.

References

[1] P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. U. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, editors, VLDB, pages 1151–1154. ACM, 2006.

[2] J. Boulos, N. Dalvi, B. Mandhani, S. Mathur, C. Re, and D. Suciu. Mystiq: a system for finding more answers by using probabilities. In SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 891–893, New York, NY, USA, 2005. ACM.

[3] D. Pfoser and C. S. Jensen. Capturing the uncertainty of moving-object representations. In *SSD*, pages 111–132, 1999.

[4] O. Wolfson, S. Chamberlain, S. Dao, L. Jiang, and G. Mendez. Cost and imprecision in modeling the position of moving objects. In *ICDE*, pages 588–596, 1998.
[5] R. Cheng, D. V. Kalashnikov, and S. Prabhakar. Querying imprecise data in moving object environments. *TKDE*, 16(9):1112–1127, 2004.
[6] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *VLDB*, pages 876–887, 2004.
[7] Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of ACM Conference on Research and Development in Information Retrieval (1999) 42–49 306
[8] Manning C. D. and Schutze H., 1999. Foundations of Statistical Natural Language Processing [M]. Cambridge: MIT Press.
[9] Yang Y. and Liu X., 1999. A Re-examination of Text Categorization Methods [A]. In: Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 42-49.
[10] Joachims T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features [A]. In: Proceedings of the European Conference on Machine Learning [C].
[11] Li Baoli, Chen Yuzhong, and Yu Shiwen, 2002. A Comparative Study on Automatic Categorization Methods for Chinese Search Engine [A]. In: Proceedings of the Eighth Joint International Computer Conference [C]. Hangzhou: Zhejiang University Press, 117-120.
[12] N. N. Dalvi et al. Efficient query evaluation on probabilistic databases. In VLDB'04
[13] V. Koltun et al. Approximately dominating representatives. In ICDT'05.
[14] R. Cheng et al. Efficient indexing methods for probabilistic threshold queries over uncertain data. In VLDB'04.
[15] O. Wolfson, A. P. Sistla, S. Chamberlain, and Y. Yesha. Updating and querying databases that track mobile units. *Distributed and Parallel Databases*, 7(3):257–387, 1999.
[16] R. Cheng, D. V. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *SIGMOD*, pages 551–562, 2003.
[17] H.-P. Kriegel et al. Probabilistic similarity join on uncertain data. In DASFAA'06.
[18] X. Dai et al. Probabilistic spatial queries on existentially uncertain data. In SSTD'05.
[19] R. Cheng et al. Evaluating probabilistic queries over imprecise data. In SIGMOD'03.
[20] Y. Tao et al. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In VLDB'05.
[21] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.
[22] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering Uncertain Data with Possible Worlds," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009.
[23] J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic Skylines on Uncertain Data," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB), 2007.
[24] Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2005.
[25] S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, and R. Shah. Orion 2.0: native support for uncertain data. In SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1239–1242, New York, NY, USA, 2008. ACM.
[26] N. N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, pages 864–875, 2004.
[27] Mohammad Hossein Elahimanesh, BehrouzMinaei-Bidgoli,Hossein Maleki nezhad "Improving K-Nearest Neighbor Efficacy for FarsiText Classification"