Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 06, pp.52482-52484, June, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

LITERATURE REVIEW ON PROBABILISTIC THRESHOLD QUERY ON UNCERTAIN DATA

***Mrs. Kavita Santosh Shevale and Prof. Gajanan V. Bhole**

Bharati Vidypeeth College of Engineering, Pune, India

ARTICLE INFO

Article History:
Received 26th March, 2017
Received in revised form
06th April, 2017
Accepted 11th May, 2017
Published online 30th June, 2017

Key words:

Nearest Neighborhood,
Revers Nearest Neighborhood,
Top K, Probability Threshold query

ABSTRACT

Uncertain data is nothing but data with impurity or a data which is not completely correct. This data uncertainty has many reasons like incorrect input, changing environment, incorrect sampling, conversions and calculations. To deal with an Information Retrieval of such a data there must be refinement at both end that is query (input) and result of search (output). This challenge is well tackled in this paper literature review of various ways of inputs and forms of expected output; has been made. To analyze the effective processing of a data Nearest Neighborhood, Revers Nearest Neighborhood, Top K forms are considered for a study. After analysis of an input suitable comparative analysis of various systems is necessary and the effective means to introduce dynamism could be added in later half; to complete the system. The backbone of the system is to set the threshold effectively to fetch the desired data with highest degree of purity.

Citation: Mrs. Kavita Santosh Shevale and Prof. Gajanan V. Bhole, 2017. "Literature review on Probabilistic Threshold Query on uncertain data", *International Journal of Current Research*, 9, (06), 52482-52484.

INTRODUCTION

Result of the query fetched may have some impurity of a data. This impurity could lead to incorrect result even after using correct algorithm to collect the result. Analysis of the gap between actual result and expected result of a query will improve the Information Retrieval process significantly. Uncertainty of a data force to deal with exact value with their dependencies. This type of scenarios was considered with comparatively less accuracy in classification algorithm like Decision Tree. Without exact value, it is difficult to categories the data for almost any classification algorithm. To understand the concept of uncertainty of a data, consider a simple example of rating given to hotel and the associated probability is specified. Based on this data selection of best fitted hotel can be made. In case of Table 1: relational database, consider example of hotel "C" where rating is 4/5 and the probability of existence of result is "1";so, the best fitted hotel (4/5) *1.0=0.8. Extending same example for probabilistic database, consider hotel "C" from table 2: probabilistic database; in this case hotel has two different rating, one is 4/5 with probability 0.4 and other is 5/5 with probability 0.6. Best fitted hotel has uncertainty in the result, that is (4/5)*0.4=0.32 and (5/5)*0.6=0.6 respectively. This ambiguity in the exact decision lead to uncertainty of data. Probability Threshold Query is a limit set to filter the result in such a way so that the results which are missed because of the constraints of a query could be also fetched. For example, if a user is looking for best hotels in a pune, the hotel which is best but just outside of pune cannot be displayed. To consider this type of valuable of data, limit or threshold need to be adjusted. This concept of adjustment of a threshold is nothing but probability threshold Query.

Corresponding author: **Mrs. Kavita Santosh Shevale,**
Bharati Vidypeeth College of Engineering, Pune, India.

Literature Survey

Probability Threshold Query, is influenced by the type of query used that is Skyline Query, Nearest Neighborhood, Revers Nearest Neighborhood, Top K queries.

Skyline Queries

Considering input in the form of skyline query, one can get results by considering the results beyond the constraints provided. Consider the example of finding best hotel in terms of cost and quality. In case of skyline query three cases are considered.

Case 1: Hotel with good quality but cost is higher which is indicated by blue line in the Figure No. 1, intersection point indicates the combination cost and quality.
Case 2: Hotel with moderate cost and quality is indicated by green lines
Case 3: Hotel with high cost and low quality is denoted by red lines, skyline query will return objects which are not worst, both in terms of cost and quality.

Table 1. Relational database for hotels

Sr. No.	Hotel Name	Rating	Probability
1	A	2/5	1
2	B	3/5	1
3	C	4/5	0.6
4	D	0/5	1

Table 2. Probabilistic database for hotels

Sr. No.	Hotel Name	Rating	Probability
1	A	2/5	1
2	B	3/5	1
3	C	4/5	0.4
		5/5	0.6
4	D	0/5	1

According to the requirement, if we restrict to set of hotels with good quality AND cost. In these scenario, we will lose valuable information about hotels which are best in only parameter; that is either cost or quality but not both.

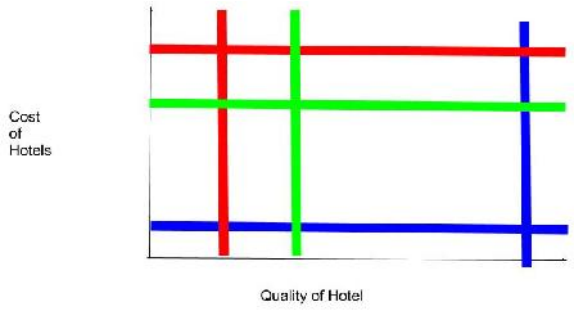


Figure 1. Skyline query to find best hotel (cost, quality)

Nearest Neighbor

Nearest neighbor approach gives the results which gives the result in the descending order of nearness. That is if we consider the same example, in that case Nearest neighbor approach will return results with descending value of quality and ascending value of cost. KNN gives the one more filtration level to the Nearest neighbor approach by allowing to provide, For example, list of 10 best hotels in the world. value k, which is nothing but the positive number. Top K approach works on same phenomenon it delivers all number of tuples, with filtration.

Reverse Nearest Neighbor

Reverse Nearest neighbor is an approach which returns all the values satisfying the result. This means, Reverse Nearest neighbor will return all the hotels considering the condition that values must be part of the system. This analysis will make the task of algorithm working on query, for setting the limit of threshold correctly. Approach to be utilized for processing should be fast to cope up with effective processing of a query dynamically. There are possibly good options like ensemble Algorithm, Adaboost Algorithm, Support Vector Machine Algorithm, Decision Tree algorithm. These approaches might be useful not only for learning but also for classification.

Support Vector Machine

SVM algorithm is used for classification and regression analysis. The approach is used to separate linear data with the help of straight line, the approach also works well to separate non-linear data. Function performing this separation is known as Kernel Function. SVM is based on the supervised learning approach.

Ensemble Learning

Ensemble is process based on the combination of models for performing the activity. The most popular ensemble learning approach is majority voting approach.

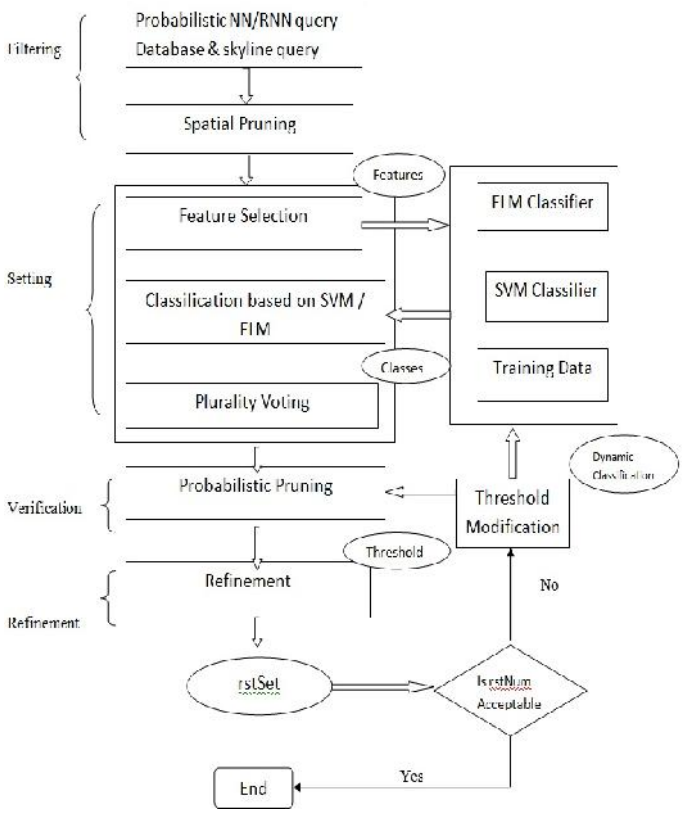
Adaboost

Adaboost algorithm is based on the phenomenon of identifying the week learner and after identification stronger approach will help to boost the performance. As a result of boosting the performance of weak learner, overall performance will be improved.

Extreme Learning Machine

Extreme Learning Machine is feed forward network used for performing the classification and regression. There exists a hidden layer, a random fixed weight is allotted to the nodes. As a layers is hidden and weight allotment is random, this algorithm or approach is one of the fastest approach that are discussed in this paper.

System Architecture



Modules

There are four phases present in this architecture.

1. Filtering
2. Setting Phase
3. Verification
4. Refinement

1. Filtering

In the filtering phase, all the objects which do not have chance to being a result of query are removed here. Generally, for improving the process of filtering, the spatial locations are used by the spatial pruning algorithm. The objects which are not pruned insert the cnd Set (candidate object set) and then transfer to the next phase. Filtering process shortlists those tuples with lowest probability of being member of a set of selected tuples. Spatial pruning algorithm helps to improve the performance of filtering process. Tuples which are not pruned are forwarded for next phase.

2. The setting phase

The setting phase is the set of s-threshold λ for query q which is based on classification threshold, which is the main work of this proposed paper. In this phase, first the feature values of q which are useful are selected and calculated here. After that ELM classifier and SVM Classifier predicts the threshold class of q, it should be corresponding class, where method of plurality voting is applied. At the last of this phase, threshold value of the q predicted in this phase is transfer to the next phase.

3. Verification phase

The verification phases confirm that which object will satisfies or fails the PTQ's. Here, for calculating the lower or upper probability bounds several probabilistic pruning algorithms might be proposed. Those objects which are not rejected or accepted are stored in the rfnSet and then transfer to the next phase.

4. Refinement Phase

This is the last phase of the architecture. In this phase the correct probability of every object in the rfn Set is calculated to whether the final result set or not rst Set. If the number of rstSet is suitable, and the query is ended; orelse, the threshold must be modified and recall the phase three. In addition, during the process if dynamic environment is detected which will be based on a sliding window methods, the ELM classifier and SVM classifier must be retrained. There is need to select the approach to efficiently deliver the result. This is simple harmonic mean of precision and recall. Number 2 denotes balancing factor.

$$\text{Efficiency} = \frac{2*\text{precision}*\text{recall}}{(\text{precision} + \text{recall})}$$

Where,

$$\text{Precision} = \frac{\text{desired result} \wedge \text{fetched result}}{\text{Fetched Results}}$$

Precision is the probability of finding desired results in among fetched results

$$\text{Recall} = \frac{\text{desired result} \wedge \text{fetched result}}{\text{Desired Results}}$$

Recall is probability of finding, how many fetched results are desired.

Conclusion

Literature survey on various approaches to deal with probabilistic threshold is examined thoroughly. Important approaches like Nearest Neighborhood, Revers Nearest Neighborhood, Top K were considered with respect of example of selection of best hotel with Hight quality of services and low cost. Also, need of the dynamic approach to adjust the threshold to cover missing result is mentioned. Exact algorithm to proceed further is yet to be selected.

REFERENCES

Bernecker, T., T. Emrich, H. Kriegel, M. Renz, S. Zankl, A. Züfle, 2011. Efficient probabilistic reverse nearest neighbor query processing on uncertain data, Very Large Data Bases, 4(10) 669–680.

Cheema, M., X. Lin, W. Wang, W. Zhang, J. Pei, 2010. Probabilistic reverse nearest neighbor queries on uncertain data, *Trans. Knowl. Data Eng.*, 22(4) 550–564.

Cheng, R., J. Chen, M. Mokbel, C. Chow, 2008. Probabilistic verifiers: evaluating constrained nearest-neighbor queries over uncertain data, In: ICDE, IEEE, pp.973–982.

Cheng, R., X. Xie, M.L. Yiu, J. Chen, L. Sun, 2010. Uv-diagram: avoronoi diagram for uncertain data, in: ICDE, IEEE, pp.796–807.

Huang, G.B., D.H. Wang, Y. Lan, 2011. Extreme learning machines: asurvey, *Int.J. Mach. Learn. Cybern.*, 2(2) 107–122. [17] R. Cheng, D. Kalashnikov, S. Prabhakar, Querying imprecise attain moving object environments, *Trans. Knowl. Data Eng.*, 16(9)(2004)1112–1127.

Huang, G.B., Q.Y. Zhu, C.K. Siew, 2004. Extreme learning machine: a new learning scheme offeed forward neural net works, in: 2004 IEEE International Joint Conference on Neural Networks,2004. Proceedings, Vol.2, IEEE, pp. 985–990.

Huang. G.B., Q.Y. Zhu, C.K. Siew, 2006. Extreme learning machine: theory and applications, Neuro computing, 70(1) 489–501.

Li, J., B. Wang, G. Wang, 2013. Efficient probabilistic reverse k- nearest neighbors query processing on uncertain data, in: Database Systems for Advanced Applications, Springer, pp.456–471.

Lian, X., L. Chen, 2009. Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data ,*Int.J. VeryL argeD ata Bases*, 18(3) 787–808.

Yang, B., H. Lu, C. S .Jensen, 2010. Probabilistic threshold k nearest neighbor queries over moving objects in symbolic in door space, In: EDBT, ACM, pp.335– 346.

****