Effective and Efficient Data Retrieval System

K.Anis Fathima
Research Scholar, Dept. of Computer Science, VELS University, Pallavaram, Chennai – 117.

T.Kamalakannan
Asst. Prof & Head, Department of BCA / IT, VELS University, Pallavaram, Chennai – 117.

Dr.A.Muthukumaravel
Professor & Head, Department of MCA, BHARATH University, Selaiyur, Chennai – 74,

Abstract - We consider the problem of parameter estimation in statistical models in the case where data are uncertain and represented as belief functions. The proposed method is based on the maximization of a generalized likelihood criterion, which can be interpreted as a degree of agreement between the statistical model and the uncertain observations. We propose a variant of the EM algorithm that iteratively maximizes this criterion. As an illustration, the method is applied to uncertain data clustering using finite mixture models, in the cases of categorical and continuous attributes. In the existing system, User gives the Search input to the Search Engine, which provides all sets of data irrespective of Relevant Results with respect to the Query as well as Redundant Results. In the proposed system, we are using Statistical and Evidence Approach to retrieve the Results. Statistical approach is used in re-ranking the results after obtaining the Feedbacks from the different Users in the corresponding URLs. In the Evidence Approach, we are evaluating resultant URLs are really matched to the query, only then the resultant URLs are displayed to the user. Modification that we propose is to get the Feedback of Rating for both the Key word Matched data as well as Information in the Resultant Data. This Process filters unwanted Resultant and provides Exactly Matched as well as Best Resultant Data to the users.

Keywords: Data Retrieval, EM Algorithm, Evidence Approach, Probability Density Function.

I. INTRODUCTION

Recent years have seen a surge of interest in methods for managing and mining uncertain data. As noted uncertain data arise in many applications due to limitations of the underlying equipment (e.g., unreliable sensors or sensor networks), use of imputation, interpolation or extrapolation techniques (to estimate, e.g., the position of moving objects), partial or uncertain responses in surveys, etc. In recent work on uncertain data mining, probability theory has often been adopted as a formal framework for representing data uncertainty. Typically, an object is represented as a Probability Density Function (PDF) over the attribute space, rather than as a single point as usually assumed when uncertainty is neglected. Mining techniques that have been proposed for such data include clustering algorithms density estimation techniques, outlier detection, support vector classification, decision trees etc. In this paper, we extend the approach introduced in , by allowing uncertainty to be expressed not only on class labels in classification problems, but on any continuous or discrete attribute, in any learning problem based on a parametric statistical model.1 The contribution of this paper is threefold:

We propose an uncertain data model in which data uncertainty is represented by belief functions; this model encompasses probabilistic data, interval valued data, and fuzzy data as special cases;
We introduce an extension of the EM algorithm, called the evidential EM (E2M) algorithm, allowing us to estimate parameters in parametric statistical models based on uncertain data.
We demonstrate the application of this algorithm for handling partially supervised clustering problems with uncertain attributes using finite mixture models.

II. OBJECTIVE OF THE PROJECT

The main aim of this project is to provide the exact result to the users when they are surfing using statistical and evidential algorithm.

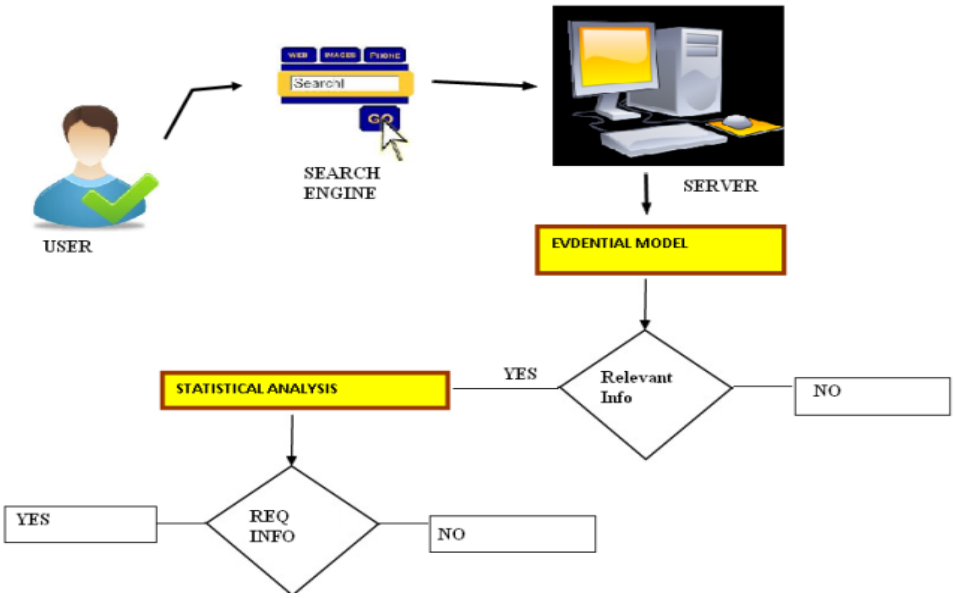
DISADVANTGES

- Uncertain data arise in many applications due to limitations of the underlying equipment (e.g., unreliable sensors or sensor networks), use of imputation, interpolation or extrapolation techniques (to estimate, e.g., the position of moving objects), partial or uncertain responses in surveys, etc.

ADVANTAGES

- Provide exact result to the user.
- We not only allowing uncertainty to be expressed not only on class labels in classification problems, but on any continuous or discrete attribute, in any learning problem based on a parametric statistical model

III. ARCHITECTURE DIAGRAM



IV. FUZZY MULTIDIMENSIONAL SCALING

N Multidimensional scaling (MDS) is a data analysis technique for representing measurements of (dis) similarity among pairs of objects as distances between points in a low-dimensional space. MDS methods differ mainly according to the distance model used to scale the proximities. The most usual model is the Euclidean one, although a spherical model is often preferred to represent correlation measurements. These two distance models are extended to the case where dissimilarities are expressed as intervals or fuzzy numbers. Each object is then no longer represented by a point but by a crisp or a fuzzy region in the chosen space. To determine these regions, two algorithms are proposed and illustrated using typical datasets. Experiments demonstrate the ability of the methods to represent both the structure and the vagueness of dissimilarity measurements.

V. THE TRANSFERABLE BELIEF MODEL

We describe the Transferable Belief Model, a model for representing quantified beliefs based on belief functions. Beliefs can be held at two levels: 1) Creedal level where beliefs are entertained and quantified by belief functions, 2) Pignistic level where beliefs can be used to make decisions and are quantified by probability functions. The relation between the belief function and the probability function when decisions must be made is derived and justified. Four paradigms are analyzed in order to compare Bayesian, upper and lower probability and the transferablebelief approaches.

VI. PRUNING BELIEF DECISION TREE METHODS IN AVERAGING AND CONJUNCTIVE APPROACHES

The belief decision tree (BDT) approach is a decision tree in an uncertain environment where the uncertainty is represented through the Transferable Belief Model (TBM), one interpretation of the belief function theory. The uncertainty can appear either in the actual class of training objects or attribute values of objects to classify. From the procedures of building BDT, we mention the averaging and the conjunctive approaches. In this paper, we develop pruning methods of belief decision trees induced within averaging and conjunctive approaches where the objective is to cope with the problem of over fitting the data in BDT in order to improve its comprehension and to increase its quality of the classification.

VII. MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA THE EM ALGORITHM

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behavior of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyper parameter estimation, iteratively reweighted least squares and factor analysis.

VIII. MODULES

1. User and Query Search
2. Server
3. Evidential Model
4. Statistical Model and Feedback
5. Re-Ranking

IX. USER AND QUERY SEARCH

To access the site, the user must have an account with that site. So that the user has register with that site by providing their user name, password and other details etc. These details are stored in the server for future purpose. Also to get the access details of the users and their feedback regarding data they have searched and viewed. For this reason, the users have to register with the site.

X. SERVER

In this module, we'll store the user information like who accessed the serve, at what time they have surfed and search details. Also the server has to retrieve the data from the database as per users query. Also the server will rank the data as per the feedback of the users they have surfed the server.

XI. EVIDENTIAL MODEL

In this module, we are evaluating resultant URLs are really matched to the query, only then the resultant URLs are displayed to the user. When the user searches the data by entering the query, the request will be send to the server. The server will retrieve the data to the user's browser page by listing the URLs of the data that the user is requested.

XII. STATISTICAL MODEL AND FEEDBACK

This module used in re-ranking the results after obtaining the Feedbacks from the different Users in the corresponding URLs. Once the data is viewed by the user, they'll provide the feedback regarding the data that they've viewed. Based on the feedback, the URLs will be displayed to the new users. We also get the data feedback of the resultant data and as well as the required data.

XIII. RE-RANKING

Once the user viewed the data they will provide the feedback. Based on the feedback, the server will re-rank the data. So that the new users may able to get the exact data that they're surfing. To implement this module, we'll have some rated keywords; so that the users are allowed to provide the feedbacks and based on the feedbacks the URLS are displayed.

XIV. CONCLUSIONS

A method for estimating parameters in statistical models in the case of uncertain observations has been introduced. The proposed formalism combines aleatory uncertainty captured by a parametric statistical model with epistemic uncertainty induced by an imperfect observation process and represented by belief functions. Our method then seeks the value of the unknown parameter that maximizes a generalized likelihood criterion, which can be interpreted as a degree of agreement between the parametric model and the uncertain data. This is achieved using the evidential EM algorithm, which is a simple extension of the classical EM algorithm with proved convergence properties.

XV. FUTURE ENHANCEMENT

As an illustration, the method has been applied to clustering problems with partial knowledge of class labels and attributes, based on latent class and Gaussian mixture models. In these problems, our approach has been shown to successfully exploit the additional information about data Uncertainty, resulting in improved performances in the clustering task. More generally, the approach introduced in this paper is applicable to any uncertain data mining problem in which a parametric statistical model can be postulated and data uncertainty arises form an imperfect observation process. This includes a wide range of problems such as classification, regression, feature extraction, and time series prediction.

REFERENCES

[1] C.C. Aggarwal and P.S. Yu, "A Survey of Uncertain Data Algorithms and Applications," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 5, pp. 609-623, May 2009.

[2] C.C. Aggarwal, Managing and Mining Uncertain Data, series Advances in Data Base Systems, vol. 35. Springer, 2009.

[3] R. Cheng, M. Chau, M. Garofalakis, and J.X. Yu, "Guest Editors' Introduction: Special Section on Mining Large Uncertain and Probabilistic Databases," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 9, pp. 1201-1202, Sept. 2010.

[4] M.A. Cheema, X. Lin, W. Wang, W. Zhang, and J. Pei, "Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 4, pp. 550- 564, Apr. 2010.

[5] Stefan Buttc her and Charles L. A. Clarke. Memory Management Strategies for Single Pass Index Construction in Text Retrieval Systems. University of Waterloo Technical Report CS-2005-32, October 2005.

[6] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and Issues in Data Stream Systems," Proc. Twenty-First ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS '02), pp. 1-16, 2002.

[7] J. Zobel and A. Moffat, "Inverted Files for Text Search Engines," ACM Computing Surveys, vol. 38, no. 2, pp. 1-55, July 2006.

[8] Y. Zhang and J. Callan, "Maximum Likelihood Estimation for Filtering Thresholds," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '01), pp. 294-302, 2001.

[9] K. Mouratidis, S. Bakiras, and D. Papadias, "Continuous Monitoring of Top-k Queries over Sliding Windows," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '06), pp. 635- 646, 2006.

[10] M. Persin, J. Zobel, and R. Sacks-Davis, "Filtered Document Retrieval with Frequency-Sorted Indexes," J. Am. Soc. for Information Science, vol. 47, no. 10, pp. 749-764, 1996.

[11] H.R. Turtle and J. Flood, "Query Evaluation: Strategies and Optimizations," Information Processing Management, vol. 31, no. 6, pp. 831-850, 1995.

[12] M. Kaszkiel, J. Zobel, and R. Sacks-Davis, "Efficient Passage Ranking for Document Databases," ACM Trans. Information Systems, vol. 17, no. 4, pp. 406-439, 1999.

[13] T. Strohman, H. Turtle, and W.B. Croft, "Optimization Strategies for Complex Queries," Proc. Research and Development in Information Retrieval (SIGIR '05), pp. 219-225, 2005.

[14] YK et al.: Knowing a tree from the forest: art image retrieval using a society of profiles. In: Proceedings of ACM Multimedia, Berkeley, CA (2003)

[15] Zhang, R., Zhang, Z.: Stretching bayesian learning in the relevance feedback of image retrieval. In: Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic (2004)