International
Journal
Of Advanced Research in Engineering & Management (IJAREM)
ISSN: 2456-2033 || PP. 04-09

Uncertain Data Representaion using Fuzzy Database Model

P. E. S. N Krishna Prasad[1], K. Madhavi[2], B D C N Prasad[3]

[1,3]*P V P Siddhartha Institute of Technology, Vijayawada, India.*
[2]*Pydah college of Engineering & Technology, Visakhapatnam, India.*

Abstract: Uncertain data is inherent in many important applications such as Data Integration, Wireless Sensor Networks, Wireless Mesh Networks, Mobile Networks, Path finding in games, Stock Market Data analysis, Biomedical Data analysis and Clinical Data, Social and Economical Research. Analysis of large collections of uncertain data is a primary task in these applications, because data is incomplete inaccurate and inefficient. Ranking queries are the most natural and useful in analyzing uncertain data.
In this paper, we proposes the representation of Uncertain Data model into fuzzy database model and vice versa for the process of uncertain data using various data models such as possibilistic linkage model, possibilistic Graphical models and Data streams. This paper presents about graphical models and representation of fuzzy data model using possibility theory.
Keywords: Uncertain Data, Ranking Queries, Fuzzy graphs, Fuzzy Linkages, Possibilistic Networks;

1. Introduction

Fuzzy logic[5,13] has the aim of studying the problem of representation of imprecision and uncertainty as well as to design and develop of fuzzy reasoning mechanism, which is able to use and take into account the fuzzy and uncertain knowledge, such as the human being able to do. Fuzzy logic is a mathematical tool. It is possibly the best tool for treating uncertain, vague, or subjective information.
The original interpretation of fuzzy sets arises from a generalization of the classic concept of a subject extended to embrace the description vague and imprecise notions. This generalization is considered as a membership of an element to a set becomes a fuzzy or vague concept. The fuzzy logic (Zadah 1992) is the logic we have approximate reasoning instead of exact reasoning. Its importance lies in the fact that may types of human reasoning particularly the reasoning based on common sense, are by nature approximate. Fuzzy reasoning framework can be applied in two cases:

1. The expression's of user preferences and the relative levels of importance.
2. The evaluation of queries in order to rank the answers according to the possibility degree to which they satisfy the user preferences.

Graphical models [1, 4,9,14] provide a general methodology for approaching the problems and many of the models developed by researchers in various fields, are instance of the general graphical models formalism.
Statisticians are increasingly concerned with the computational aspects both theoretical and practical, of models and inference procedures. Computer Experts are increasingly concerned with the systems that interact with the external world and interpret uncertain data in terms of underlying possibility models. One area in which these trends are most evident is that of possibility graphical models.

A graphical model [9] is a family of possibility distributions defined in terms of directed/ undirected graphs. The nodes in the graph are identified with random features, and joint possibility distributions are defined by considering products over functions defined on connected subsets of nodes. By exploiting the graph-theoretic representation, the formalism provides general algorithms for computing possibilities of interest. Moreover, the formalism provides control over the computational complexity associated with the operations.
The graphical approach is its naturalness in formulating possibility models of complex phenomena in applied fields such as sensor networks, Biomedical, Clinical data, Image mining, Geographical data analysis and so on, while maintaining control over the computational costs associated with these models.

Bayesian networks [6,7] and belief models provide a well founded framework for knowledge representation and reasoning with uncertain, but precise data. Extending pure probabilistic strategies to the treatment of imprecise information usually restricts the computational tractability of the corresponding inference

mechanisms. So an alternative mechanism has been considered on uncertain calculi that provide a well defined form of information compression in order to support efficient reasoning in the presence of imprecise and uncertain data without affecting the expressive power and correctness of decision making.

The ideas of graphical models [2,3] can be traced back to three origins (according to [Lauritzen 1996]), namely statistical mechanics [Gibbs 1902], genetics [Wright 1921], and the analysis of contingency tables [Bartlett 1935]. Originally, they were developed as means to build models of a domain of interest.

The rationale underlying such models is that, since high-dimensional domains tend to be unmanageable as a whole (and the more so if imprecision and uncertainty are involved), it is necessary to *decompose* the available information.

In *graphical modelling* [5,8,20] [Whittaker 1990, Kruse *et al.* 1991, Lauritzen 1996] such a decomposition exploits (conditional) dependence and independence relations between the attributes used to describe the domain under consideration. The structure of these relations is represented as a network or graph (hence the names graphical model and inference network), often called a *conditional independence graph*. In such a graph each node stands for an attribute and each edge for a direct dependence between two attributes.

However, such a conditional independence graph[8] turns out to be not only a convenient way to represent the content of a model. It can also be used to facilitate reasoning in high-dimensional domains, since it allows us to draw inferences by computations in lower-dimensional subspaces.

Propagating evidence about the values of observed attributes to unobserved ones can be implemented by locally communicating node processors and therefore can be made very efficient. As a consequence, graphical models were quickly adopted for use in expert and decision support systems [Neapolitan 1990, Kruse *et al.* 1991, Cowell 1992, Castillo *et al.* 1997, Jensen 2001]. In such a context, that is, if graphical models are used to draw inferences, we prefer to call them *inference networks* in order to emphasize this objective.

Graphical modelling [5,13,19,22] was also generalized to be usable with uncertainty calculi other than probability theory [Shafer and Shenoy 1988, Shenoy 1992b, Shenoy 1993], for instance in the so-called valuation based networks [Shenoy 1992a], and was implemented, for example, in PULCINELLA [Saffiotti and Umkehrer 1991]. Due to their connection to fuzzy systems, which in the past have successfully been applied to solve control problems and to represent imperfect knowledge, possibilistic networks gained attention too. They can be based on the context model interpretation of a degree of possibility, which focuses on imprecision

Initially the standard approach to construct a graphical model was to let a human domain expert specify the dependences in the domain under consideration. This provided the network structure. Then the human domain expert had to estimate the necessary conditional or marginal distribution functions that represent the quantitative information about the domain. This approach, however, can be tedious and time consuming, especially if the domain under consideration is large. In some situations it may even be impossible to carry out, because no, or only vague, expert knowledge is available about the (conditional) dependence and independence relations that hold in the considered domain, or the needed distribution functions cannot be estimated reliably.

Graphical models [5,19] have several advantages when applied to knowledge discovery and data mining problems.

1. The network representation provides a comprehensible qualitative (network structure) and quantitative description (associated distribution functions) of the domain under consideration, so that the learning result can be checked for plausibility against the intuition of human experts.
2. Secondly, learning algorithms for inference networks can fairly easily be extended to incorporate the background knowledge of human experts. In the simplest case a human domain expert specifies the dependence structure of the domain to be modeled and automatic learning is used only to determine the distribution functions from a database of sample cases. More sophisticated approaches take a prior model of the domain and modify it (add or remove edges, change the distribution functions) w.r.t. the evidence provided by a database of sample cases [Heckerman *et al.* 1995].
3. Finally, although fairly early on the learning task was shown to be NP-complete in the general case [Chickering *et al.* 1994, Chickering 1995], there are several good heuristic approaches that have proven to be successful in practice and that lead to very efficient learning algorithms

2. Possibilistic Models

Possibility theory [2, 14, 16, 20] is one the framework for the process of imprecise and uncertain data. This theory is based on the idea that we can evaluate the possibility of determinant variable ‘x’ belongings to a

determinant set or event A. Here, fuzzy sets are called possibility distributions instead of membership degrees may often called possibility degrees.

When the state of knowledge is expressed by a body of evidence it becomes clear that probability measures address precise but differentiated items of information, whereas possibility measures reflect imprecise but coherent items. So, possibility measures are useful for subjective uncertainty: one expects from information no very precise data; however, one expects the greatest possible coherence among his statements. On the other hand, precise, but variable data are usually the result of carefully observing physical phenomena. As a rule, the state of knowledge is neither precise nor totally coherent, in the general case, the of "*A degree of credibility"*(the degree of confidence).

Possibilistic Networks and Possibilistic logic [2, 4, 6, 9, 10, 11, 18] are two standard frameworks for representing uncertain pieces of knowledge. Possibilistic Networks exhibit relationship between the features where as Possibilistic logic ranks the logical formulas according to their level of certainty, it is well known that the inference process is a hard problem. These two types models representation are semantically equivalent when they lead to same possibility distribution. A possibility distribution can be decomposed using a chain of rule that may be based on two different kinds of conditioning that exist in possibility theory. These two types induce the possibilistic graphs.

The notion of possibilistic graphs [17] for the representation of multidimensional possibility distributions are of two types 1) undirected graphs or hypergraphs 2) Directed graphs.

Possibilistic Networks is an important tool for an efficient representation and analysis of Uncertain Data. The simplicity and capacity of representing and handling of independence relationships are important for an efficient management of uncertain information.

Possibilistic networks are Undirected / Directed graphs where each node (vertex) encodes a variable/feature and every edge represents a casual or an inference relationship between variables. Uncertainty is expressed as conditional possibility distribution for each node in the context.

Possibilistic logic is an extension of classical logic. A weight is associated with each propositional formula. This weight represents the priority reading other formulas. The set such weighted formulas is called possibilistic knowledge base.

This paper proposes a new representation of uncertain data using possibilistic models that provides the ranking to queries of uncertain data in the uncertain databases (Knowledge Bases) through possibilistic logic. Such mechanism is often called as "*Possibilistic ranking queries on uncertain data*".

The basic element of possibility theory is the possibility distribution 'π' which maps from Ω to [0, 1]. The degree π(ω) represents the compatibility of ω with the available information about the real world. A possibility distribution π is said to be normal if π(ω)=1, there exists at least one interpretation which is consistent with all the available beliefs.

The possibility distribution associated with the knowledge base $\sum = \{(p, \alpha)\} is\ \forall \omega \in \Omega$

$$\pi_{\{(p,\alpha)\}}(\omega) = \begin{cases} 1 \ \ if\ \omega \in [p_i] \forall (p_i, \alpha_i) \in \Sigma;, \\ 1 - \max\{\alpha_i : (p_i, \alpha_i) \in \Sigma;\ \omega \notin [p_i]\} otherwise \end{cases}$$

Thus using the minimum operator, $\pi(\omega) = \min \{\pi_{\{(p_i,\alpha_i)\}}(\omega) : (p_i, \alpha_i) \in \Sigma;$

2.1 Axioms of Possibilistic model

The list of axioms of the Possiblistic model is given below:

$$\pi(X) = 1$$
$$\pi(\emptyset) = 1$$
$$\pi(E_1 \cup E_2) = \max(\pi(E_1), \pi(E_2))$$
$$\pi(E_1 \cap E_2) \leq \min(\pi(E_1), \pi(E_2))$$
$$\pi(E_1 \cap E_2) = \min(\pi(E_1), \pi(E_2)) \text{ (for non-interactive events)}$$
$$max(\pi(E\_1), \bar{\pi}(E\_2)) = 1$$
$$N(E) = 1 - \pi(E)$$

2.2 Conditional Independence

Let E_1,E_2 and E_3 be the three disjoint subsets of features, then E_1 is called conditionally independent of E_2 given E_3 with respect to π, if $\forall \omega \in \Omega$

$$\pi(\omega_{E_1 \cup E_2} | \omega_{E_2}) = min\{\pi(\omega_{E_1} | \omega_{E_2}), \pi(\omega_{E_2} | \omega_{E_2})\}$$

Whenever $\pi(\omega_{E_2}) > 0$, where $\omega_{E_3} = proj_{W_3}(\omega)$ is the projection of a tuple ω to the features in E_1 and $\pi(.|.)$ is a non-normalized conditional possibility distribution

$$\pi(\omega_{E_1} | \omega_{E_2}) = \max\{\pi(\omega) | \omega \in \Omega \wedge proj_{E_1}(\omega) = \omega_{E_1} \wedge proj_{E_2}(\omega) = \omega_{E_2}$$

2.3 Possibilistic Classifier

A Possibilistic network [12, 21] represents a decomposition of a multi-variate possibility distribution according to $\pi(E_1, \ldots\ldots, E_n) = min_{j=1}^{n} \pi\left(E_j \middle| parents(E_j)\right)$,

where parents(E_j) is the set of parents of features/attributes E_j, is as small as possible by exploiting conditional independence of the type indicated.

Let π be a possibility distribution on the attributes/features $E_1, \ldots\ldots, E_n$ and a Classifier *C*, then possibilistic classifier can be defined as

$$\pi\left(C = c_i \middle| E_1 = a_{i_1}^{(1)}, \ldots\ldots, E_n = a_{i_n}^{(n)}\right)$$
$$= \min\{\pi\left(E_1 = a_{i_1}^{(1)} \middle| C = c_i\right),$$
$$\pi(E_2 = a_{i_2}^{(2)} | C = c_i),$$
$$\ldots\ldots\ldots$$
$$\pi(E_n = a_{i_n}^{(n)} | C = c_i\}$$

Where c_i is the class that predits the highest degree of the possibility.

3. Uncertain Data Model

Modeling and Querying uncertain data [6, 13, 15, 23, 24, 25] has been a fast growing research direction and receives an increasing attention. Various models of uncertain and fuzzy data have been developed. We proposed a novel model for modeling uncertain data in the fuzzy environment using *Possibilistic data model***.** The working model for uncertain data describes the existence possibility of a tuple in an uncertain data set and the constraints on the uncertain tuples.

A fuzzy database comprises of multiple fuzzy tables. A fuzzy table contains a set of tuples, where each tuple is associated with a fuzzy membership value, which is treated as Degree of Possibility in the Possibilistic Model. A fuzzy table may also come with some generation rules to capture the dependencies among tuples, where a generation rule specifies a set of exclusive tuples, and each tuple is involved in at most one generation rule.

Another useful model is the *Uncertain Object Model*, an uncertain object is conceptually described by a fuzzy membership function, i.e. Possibility Distribution in the data space. In this scenario a possibility degree of an uncertain object is unknown, a set of samples (instances) are collected to approximate the fuzzy distribution, which is a possibility distribution.

Definition: An *Uncertain Object* is a set of instances $X = \{x_1, x_2, \ldots\ldots, x_m\}$ such that each instance $x_i\ (1 \leq i \leq m)$ takes a a Possibility degree $\pi(x_i) and\ \sum_{i=1}^{m} \pi(x_i) \leq 1$. The cardinality of an uncertain object $X = \{x_1, x_2, \ldots\ldots, x_m\}$ denoted by |X| is the number of instances contained in *X*. The set of all uncertain objects denoted by *U*, U= $\{X_1, X_2, \ldots\ldots, X_n\}$.

Possible worlds of Uncertain Objects: Let $U = \{X_1, X_2, \ldots\ldots, X_n\}$ be a set of Uncertain Objects. A possible world $W = \{x_1, x_2, \ldots\ldots, x_m\}$ is a set of instances such that one instance is taken from each uncertain Object. The existence membership of ω is $\mu(W) = \pi(W) = \prod_{i=1}^{n} \pi(x_i)$, where W denotes the set of all

possible worlds. Let $|X_i|$ be the cardinality of Object X_i ($1 \le i \le m$), the number of possible worlds is $|W| = \prod_{i=1}^{n} |X_{i|}$. Thus, $\pi(W) = \sum_{w \in W} \pi(w) = 1$.

A fuzzy database model is used to represent uncertain data, which is a finite set of fuzzy tables, A fuzzy table contains a set of *uncertain tuples* T and a set of *generation rules* $\Re$. Each uncertain tuple $t \in T$ is associated with a *possibility degree* $\pi(1) > 0$. Each generation rule $R \in \Re$ specifies a set of exclusive tuples in the form $R: t_{r_1} \oplus t_{r_2} \oplus \ldots\ldots \oplus t_{r_m}$, where $t_{r_i} \in T(1 \le i \le m)$, $\pi\left(t_{r_i} \wedge t_{r_j}\right) = 0 \ (1 \le i, j \le m, i \ne j)$ and $\sum_{i=1}^{m} \pi\left(t_{r_i}\right) = 1$.

The cardinality of a generation rule R, denoted by |R|, is the number of tuples involved in R. The generation rule R is the set of all tuples $t_{r_1}, t_{r_2}, \ldots\ldots\ldots t_{r_m}$ involved in the rule, atmost one tuple can appear in a possible world. R is a singleton rule if there is only one tuple involved in the rule, otherwise R is a multiple rule, thus the fuzzy database follows a possible world.

Given a fuzzy Table, $\tilde{T}$ a possible world W is a subset of $\tilde{T}$ such that for each generation rule $\in \Re_{\tilde{T}}, |R \cap W| = 1, if\ \pi(R) \le 1 and\ |R \cap W| \le 1, if\ \pi(R) < 1.$ Thus, the existing membership of W is

$$\pi(W) = \prod_{R \in \Re_{\tilde{T}}, R \cap W = 1} \pi(R \cap W). \prod_{R \in \Re_{\tilde{T}}, R \cap W = \emptyset} \pi(1 - \pi(R)).$$

For an uncertain table $\tilde{T}$ with a set of generation rules $\Re_{\tilde{T}}$, the number of all possible worlds is

$$|W| = \prod_{R \in \Re_{\tilde{T}}, \pi(R) = 1} |R|. \prod_{R \in \Re_{\tilde{T}}, \pi(R) < 1} |R| + 1.$$

We can convert the Uncertain Object model into Fuzzy database model and vice versa. When, it is concerned that both are equivalent. The conversion process will be presented below.

1. *Conversion between Uncertain Object Model into Fuzzy Database Model*: A set of uncertain objects can be represented by a fuzzy table. For each instance '*x*', of an uncertain Object 'X', to create a tuple t_x, whose membership or possibility degree value is $\mu(x) = \pi(x)$. for each uncertain object $X = \{x_1, x_2, \ldots\ldots, x_m\}$, to create one generation rule $R_X: t_{x_1} \oplus t_{x_2} \oplus \ldots\ldots \oplus t_{x_m}$.
2. *Conversion between Fuzzy Database models to Uncertain Object Model*: A fuzzy table can be represented by a set of uncertain objects with discrete instances. For each tuple *t* in a fuzzy table, to create an instance x_t, whose membership or possibility degree is $f(x_t) = \pi(t)$. For a generation rule $R: t_{r_1} \oplus t_{r_2} \oplus \ldots\ldots \oplus t_{r_m}$, to create an uncertain object X_R, which includes instances $x_{t_{x_1}}, \ldots\ldots x_{t_{xm}}$ corresponding to $t_{r_1}, \ldots\ldots\ldots t_{r_m}$, respectively. Moreever, $\sum_{i=1}^{m} \pi(t_{r_i}) < 1$. We create another instance x_Q whose fuzzy membership function is $f(x_Q) = 1 - \sum_{i=1}^{m} \pi(t_{r_i})$ and u_Q to the uncertain object X_R.

4. Conclusions and Future Work

The main aim of this paper is to represent the uncertain data in various forms of Object models for processing and evaluating Query and also give the ranking to the evaluated query. Here, an uncertain object model is represented as Fuzzy Database Model and vice versa so that the uncertain data model can be evaluated through the query evolution mechanism using Fuzzy Database model. The future scope of this task is to represent uncertain data as Possibilistic Linkage model and Possibilistic Graphical Models that process the data objects to evaluate through query evaluation mechanism using Possibility theory.

References

[1]. Christian Borgelt, Possibilistic Graphical Models: How to Learn them from Data.
[2]. Salem Benferhat, Didier Dubois Laurent Garcia and Henri Prade, Possibilistic logic bases and Possibilistic Graphs.
[3]. Angelos Vasilakopoulos andVerena Kantere, Efficient Query Computing for Uncertain Possibilistic Databases with Provenance.
[4]. Jorg Gebhardt and Rodulf Kruse , Learning Possibilistic Graphical Networks from Data, IEEE Transaction on Fuzzy Systems
[5]. M. S. Sunitha and A. Vijay Kumar, Complement of a Fuzzy Graph, Indian J. Pure Applied Mathematics, 33(9), 1451-1464, September 2002.
[6]. Damiani. E, Tanca L.,Building Queries to XML data, Database and Expert Systems Applications, 2000, pp 266-279.
[7]. Christian Borgelt, J¨org Gebhardt, and Rudolf Kruse, Possibilistic Graphical Models,
[8]. Pascale FONCK, Conditional Independence in Possibility Theory
[9]. Christian Borgelt, and Rudolf Kruse, Data Mining with Possibilistic Graphical Models
[10]. Salem Benferhat and Salma Smaoui, A Hybrid Possibilistic Networks
[11]. Salem Benferhat and Salma Smaoui, on the use of Possibistic bases for local computations in product based possibilistic Networks
[12]. Christian Borgelt and J¨org Gebhardt, A Naïve Bayes style Possibilistc Classifier.
[13]. Anish Das Sarma, Omar Benjelloun, Alon Halevy,Jennifer Widom, Working Models for Uncertain data.
[14]. Christian Borgelt and Rudolf Kruse, Operations and Evoluation measures for Learning Possibilistic Graphical Models
[15]. Jiang Chen and Ke Yi, Dynamic structures for Top-kqueries on Uncertain Data
[16]. AbdelKader Heni, M Nazih Omri and Adel M Alimi, Knowledge Representation with Possibilstic and Certain Bayesian Networks
[17]. Christian Borgelt, J¨org Gebhardt, and Rudolf Kruse, Learning from Imprecise data: Possibilstic Graphical Models
[18]. C.C. Agarwal, Managing and mining Uncertain Data
[19]. Brian Babcock Shivnath Babu Mayur Datar Rajeev Motwani Jennifer Widom, Models and Issues in Data Stream Systesms
[20]. Anish Das Sarma, Omar Benjelloun, Alon Halevy, Shubha Nabar, and JenniferWidom, Representing Uncertain Data: models,Properties and Algorithms
[21]. Salem Benferhat and Salma Smaoui,Possibilistic Casual Networks for handling Interventions: A New Propogation Algorithm
[22]. Salem Benferhat, Didier Dubois Laurent Garcia and Henri Prade, On the Transformation between Possibilistic Logic bases and Possibilistic Casual Networks
[23]. Jiang Chen and Ke Yi , Ranking Distributed Probabilistic Data
[24]. Goetz Graefe, Query Evaluation Techniques for Large Databases.
[25]. Graham Cormode, Divesh Srivastava, Entong Shen and Ting Yu, Aggregating Query Answering on Possibilistic Data with Cardinality Constraints.