Fuzzy and Probabilistic Object-Oriented Databases

Tru H. Cao
Ho Chi Minh City University of Technology, Vietnam

INTRODUCTION

For modeling real-world problems and constructing intelligent systems, integration of different methodologies and techniques has been the quest and focus of significant interdisciplinary research effort. The advantages of such a hybrid system are that the strengths of its partners are combined and complementary to each other's weakness.

In particular, object orientation provides a hierarchical data abstraction scheme and a mechanism for information hiding and inheritance. However, the classical object-oriented data model cannot deal with uncertainty and imprecision pervasive in real world problems. Meanwhile, probability theory and fuzzy logic provide measures and rules for representing and reasoning with uncertainty and imprecision. That has led to intensive research and development of fuzzy and probabilistic object-oriented databases, as collectively reported in De Caluwe (1997), Ma (2005), and Marín & Vila (2007).

BACKGROUND

The key issues in research on extending the classical object-oriented data models to deal with uncertainty and imprecision are:

1. Modeling partial subclass relationship.
2. Definition of partial class membership.
3. Representation of uncertain and/or imprecise attribute values.
4. Representation and execution of class methods.
5. Expression of partial applicability of class properties.
6. Mechanism for inheritance under uncertainty and imprecision.

In the classical object-oriented data model, a class hierarchy defines the subclass/super-class relation on classes. A class *A* is derived as a subclass of a class *B*, which is then called *A*'s super-class, either by narrowing the crisp value ranges of *B*'s attributes or by adding new properties to *B*'s ones. In the probabilistic and fuzzy case, due to the uncertain applicability of class properties or the imprecision of attribute value ranges, the inclusion between classes naturally becomes graded, which could be computed on the basis of the value ranges of their common attributes (George & Buckles & Petry, 1993, Rossazza & Dubois & Prade, 1997).

As discussed in Baldwin, Cao, Martin, and Rossiter (2000), a set of classes with a graded inclusion or inheritance relation actually forms a network rather than a hierarchy, because if a class *A* has some inclusion degree into a class *B* based on a fuzzy matching of their descriptions, then *B* usually also has some inclusion degree into *A*. Moreover, naturally, a concept is usually classified into sub-concepts that are totally subsumed by it, though the sub-concepts can overlap each other, as assumed in Dubitzky, Büchner, Hughes, and Bell (1999) for instance.

Uncertain and imprecise attribute values lead to partial membership of an object into a class, and there are different measures proposed. Yazici and George (1999), for instance, defined for each class a membership function on a set of objects. Bordogna, Pasi, and Lucarella (1999) used linguistic labels to express the strength of the link of an object to a class. Dubitzky et al. (1999) defined membership as similarity degrees between objects and classes. Blanco, Marín, Pons, and Vila (2001) mentioned different measures, including probabilistic one, to be used for membership degrees. Nevertheless, it is to be answered how measures of different meanings, such as possibility and probability, on various levels of a model are integrated coherently.

Most of the works on fuzzy object-oriented data models, which are referred in this paper, were mainly based on fuzzy set and possibility theories, and used fuzzy sets or possibility distributions to represent imprecise attribute values. Bordogna, Pasi, and Lucarella (1999) and Blanco et al. (2001) also modeled uncertainty about an attribute having a particular value. However, much less concern was given for uncertainty over a set of values of an attribute and a foundation to combine probability degrees and fuzzy sets in the same model.

While class attributes were paid much attention and treatment, class methods, as common in object-oriented systems for modeling object behaviors and parameterized properties, were often neglected. In Dubitzky et al. (1999) and Blanco et al. (2001) methods were not considered.

Bordogna, Pasi, and Lucarella (1999) mentioned about methods but did not provide formal representation and explicit manipulation in their model. In Yazici (1999) and Cao and Rossiter (2003) methods were formally defined as Horn clauses and executed as a reasoning process, which were thus for declarative and deductive in contrast to imperative and procedural models.

In the classical object-oriented data model, the properties that represent a class are necessary and sufficient to define the class. However, there is no commonly agreed set of defining properties for many natural, scientific, artificial, and ontological concepts. Arguing for flexible modeling, Van Gyseghem and De Caluwe (1997) introduced the notion of *fuzzy property* as an intermediate between the two extreme notions of required property and optional property, each of which was associated with a possibility degree of applicability of the property to the class. Meanwhile, Dubitzky et al. (1999) addressed the issue by contrasting the prototype concept model with the classical one, assuming each property of a concept to have a probability degree for it occurring in exemplars of that concept.

We note the distinction between the notion of uncertain property values and that of uncertain property applicability. In the former case, an object surely has a particular property but it is not sure which one among a given set of values the property takes. Meanwhile, in the latter, it is even not sure if the object has that property. For example, "John owns a car whose brand is probably BMW" and "It is likely that John owns a car" express different levels of uncertainty. In Bordogna, Pasi, and Lucarella (1999), Blanco et al. (2001), and Cao and Rossiter (2003), the two levels were mixed.

Uncertain class membership and uncertain property applicability naturally result in *uncertain inheritance* of class properties. This was not considered in Bordogna, Pasi, and Lucarella (1999), Dubitzky et al. (1999), and Yazici and George (1999). In Blanco et al. (2001), class membership degrees were used as thresholds to determine what part of the properties in a class would be inherited. In Cao and Rossiter (2003), both membership of an object into a class and applicability of a property to the class were represented by support pairs (Baldwin, Lawry & Martin, 1996) and combined into the support pair for the object to inherit the property.

Recently, Cross (2003) reviewed existing proposals and presented recommendations for the application of fuzzy set theory in a flexible generalized object model. Furthermore, De Tré and De Caluwe (2005) focused on representing data as constraints on object attributes and query answering as constraint satisfaction. For realization of fuzzy object-oriented data models, Berzal et al. (2005) were concerned with implementation of their model on an existing platform. Meanwhile, Fril++, a fuzzy object-oriented logic programming language, was also developed in Rossiter and Cao (2005).

While the fuzzy object-oriented data models referred in this paper were mainly based on fuzzy set and possibility theories, Eiter, Lu, Lukasiewicz, and Subrahmanian (2001) introduced a probabilistic model to handle object bases with uncertainty, called POB (Probabilistic Object Base). For a POB class hierarchy, although a class was assumed to be fully included in its super-classes, the model specified the conditional probability for an object of a class belonging to each of its subclasses. Intuitively, it specified how likely an object of a class belonged to a subclass of that class. Accordingly, the partial class membership was measured by probability degrees. For each attribute of an object, uncertainty about its value was represented by lower bound and upper bound probability distributions on a set of values. The authors also developed a full-fledged algebra to query and operate on object bases.

However, the two major shortcomings of the POB model are: (1) it does not allow imprecise attribute values; and (2) it does not consider class methods. For instance, in the Plant example therein, the values of the attribute *sun light* are chosen to be only enumerated symbols such as *mild*, *medium*, and *heavy* without any interpretation. Meanwhile, in practice, those linguistic values are inherently vague and imprecise over degrees of sun light. Moreover, without an interpretation, they cannot be measured and their probability distributions cannot be calculated.

A HYBRID MODEL

In Cao and Nguyen (2007) and Nguyen and Cao (2007), POB is extended with fuzzy attribute values, class methods, and uncertain applicability of class properties. Here, the term *property* is used to subsume both the terms *attribute* and *method*. This hybrid model is called FPOB (Fuzzy-Probabilistic Object Base).

Figure 1 is an FPOB hierarchy of patients, who are classified as being children, teenagers or adults and, alternatively, as being out-patients, or in-patients. Those subclasses of a class that are connected to a **d** node are mutually disjoint, and they form a cluster of that class. The value in [0, 1] associated with the link between a class and one of its immediate subclasses represents the probability for an object of the class belonging to that subclass. For instance, the hierarchy says 80% of patients are non-resident, while the rest 20% are resident. As such, each object could be a member of a class with some probability.

Basically, imprecise and uncertain values of an attribute is expressed by a *fuzzy-probabilistic triple* of the form $\langle V, \alpha, \beta \rangle$, where V is a set of *fuzzy values*, that is,, those defined by fuzzy sets, and α and β are lower and upper bound probability distributions on V. For example, $\langle \{young, middle\_aged\}, .8u, 1.2u \rangle$, represents that the probability for the age of a patient is *young* or *middle-aged* is between 0.4 and 0.6, where *young* and *middle-aged* are linguistic labels of fuzzy sets, and u denotes the uniform distribution.