Approaches for Handling Uncertainty in Decision Making

K.Soundararajan
Assistant Professor
Vivekanandha College of
Engineering for Women

S.Suresh Kumar, PhD.
Principal
Vivekanandha College of
Technology for Women

ABSTRACT
Data uncertainty is common in real world applications due to various causes, including imprecise measurements, network latency, out dated sources and sampling errors. These kinds of uncertainty have to be handled cautiously, or else the mining results could be unreliable or even wrong. In this paper, we are describing the various ways for managing, mining and handling uncertainty. Uncertain data are inherent in many applications. Recently, considerable research efforts have been put into the field of managing uncertain data. There are many algorithms to handle the uncertainty. Some of them are iterative algorithm, Rule based classification approach, Associative classification model and Density based clustering approach and probabilistic queries and Decision rule based on rough set theory. The algorithm can select the decision rules on the basis of meeting the support and confidence, which can improve the accuracy and reasonableness of the decision rules mining.

Keywords
Uncertain Data, Iterative algorithm, Rule based classification, Associative classification, Density based clustering, Probabilistic queries and Rough set theory.

1. INTRODUCTION
In many applications, data contains inherent uncertainty. A number of factors contribute to the uncertainty, such as random nature of physical data generation and collection process, measurement and decision errors, unreliable data transmission and data stalling. The iterative algorithm is effectively used to estimate the missing attribute values in both training data and test data. The goal is to construct a classifier from a training set, and improve the prediction accuracy on the test set. In Rule based classification and prediction algorithm called uRule for classifying uncertain data. This introduces new measures for generating, pruning and optimizing the rules. The new measures are computed considering uncertain data interval and probabilistic distribution function. It can process both numerical and categorical data. In Associative classification model the transformation of quantitative association rule into linguistic representation can be achieved in discretizing the numerical intervals into rough interval described with respective rough membership values. In Density based clustering approach, the distance between objects is computed based on vague and uncertain data. The distance between these uncertain object descriptions are expressed by numeric distance value and by distance probability functions. The fuzzy distance function assigns a probability value to each possible value. This algorithm works directly on these fuzzy distance functions. The probabilistic query evaluation is based on uncertain data. If the degree of error between the actual value and the database is controlled, we can place more confidence in answers to the queries. The implementation of rough set theory in incomplete data decision and reasoning of basic principles of two-dimensional relational databases, and presents a decision rule mining algorithm of information system based on rough set.

2. ITERATIVE ALGORITHM
The simplest approach is to simply ignore instances with any missing values. This reduces the amount of information available. In Iterative algorithm, the attributes are selected one by one. For each attribute, the unknown values are predicted using a decision tree built using the other attributes from cases with known values of the attribute. The training set filled in this way is used to classify a tuning set whose prediction error rate decides which attribute is selected to be filled in the current iteration. Many strategies to deal with incomplete data have been developed. Another common approach is to replace missing values with global or class-conditional mean/mode.

2.1 Estimating missing values in training data
In iteration approach, the missing values are estimated iteration by iteration. At each iteration, one attribute is chosen, and all missing values of that attribute are filled. It is not always the case that all the missing values lead to the best performance.
To address this problem, we track the tuning error of the selected classifier at each iteration. After all attributes are filled select the classifier that produces the lowest tuning error rate.

2.2 Estimating missing values in test data
The missing values in the test data will also cause the class prediction to degrade even if the training data are filled. The selected attribute at the current iteration is regarded as the target attribute, and the attribute tree used for training data is again applied on the test data. If multiple missing values exist in a test case, replace all them with their modes before the prediction of the target.

2.3 Imputation methods
There are four imputation methods which are able to deal with missing data in both training data and test data. The methods are as follows
(i)*Mode*
It replaces all the missing values in training data and test data with the mode of attribute in the training data and then apply the decision tree method to the filled data set.
(ii) Attribute Tree(AT)
It constructs a decision tree for each attribute to estimate the missing value.
(iii) Ordered Attribute Value(OAT)
A decision tree is constructed to determine the missing values of each attribute by using the information contained in other attributes. Also, an ordering for the construction of the decision trees for the attributes is formulated. The ordering is based on the concept in Information Theory called mutual information, which has been successfully used as a criteria for attribute selection in decision tree learning.
(iv) Dynamically Ordered Attribute Tree (DOAT)
DOAT utilizes the correlation between attributes for estimating missing values and also intends to select the best attribute at each iteration by evaluating the performance of a tuning set which stimulates the result of the test data. The important part of

estimating the utility of acquiring the missing value would be an estimate of how well the value can be filled automatically.

3. RULE BASED CLASSIFICATION

This algorithm is used for classifying and predicting uncertain data. Integrate the uncertain data model into the rule based mining algorithm using probabilistic information gain for generating rules. The goal of probabilistic information gain is to identify the optimal split attribute and split point for uncertain training dataset. To build a new rule-based classifier, extract a set of rules that shows the relationship between the attributes of a data set and class label. Each classification rule is of the form R: (Condition) => y. The condition is called rule antecedent, which is conjunction of the attribute test condition. y is called the rule consequent and it is the class label. A rule set consist of multiple rules $R_s=\{R_1,R_2,\ldots,R_n\}$.The Coverage of a rule is the number of instances that satisfy the antecedent of a rule. The accuracy of a rule is the fraction of instances that satisfy both the antecedent and consequent of a rule.

uLearnOneRule () is the key function of the uRule algorithm. It generates best rule for the current class, given the current set of uncertain training tuples. After generating the rule, all the positive and negative examples covered by the rule are eliminated. The rule is then added into the rule set as long as it does not violate the stopping condition. This algorithm will handle both numerical and categorical data.

Algorithm 1: uRule(Dataset D,ClassSet C)

```
begin
   RuleSet=φ;
   for Each class c_i Ɛ C do
        newRuleSet=uLearnOneRule(D,c_i);
        remove tuples covered by newRuleSet from DataSet D;
        RuleSet +=newRuleSet;
end for;
return RuleSet;
end
```

3.1 Uncertain Numerical Data

Table 1. Example of UNA

Row ID	Home Owner	Marital status	Annual Income	Defaulted Borrower
1.	Yes	Single	110-120	No
2.	No	Single	60-85	No
3.	Yes	Married	110-145	No
4.	No	Divorced	110-120	Yes
5.	No	Married	50-80	No
6.	Yes	Divorced	170-250	No
7.	No	Single	85-100	Yes
8.	No	Married	80-100	No

When the value of a numerical attribute is uncertain, the attribute is called an uncertain numerical attribute (UNA).The value of UNA is represented as a range or interval and the Probability Distribution Function (PDF) over the range.

The PDF f(x) can be related to an attribute if all the instances have the same distribution or related to each instances if each instance has different distribution. The data in Table I are used to predict whether borrowers will default on loan payments. Among all the attributes, the annual income is UNA, whose precise value is not available.

3.2 Uncertain Categorical Data

Under the uncertainty categorical model, a dataset can have attributes that are allowed to take uncertain values. The dataset in Table II is used for a medical diagnosis and it contains information for cancer patients with a tumor.

Table 2. Example of UCA

Row ID	Sex	Symptom	Tumor	Class
1.	M	a	(Benign,0.3) (Malignant,0.7)	1
2.	M	b	(Benign,0.2) (Malignant,0.8)	0
3.	M	c	(Benign,0.9) (Malignant,0.1)	1
4.	F	b	(Benign,0.3) (Malignant,0.7)	1
5.	F	a	(Benign,0.8) (Malignant,0.2)	1
6.	F	a	(Benign,0.4) (Malignant,0.6)	1
7.	F	a	(Benign,0.1) (Malignant,0.9)	0

The type of tumor for each patient is a UCA attribute, whose exact value is unobtainable.

4. ASSOCIATIVE CLASSIFICATION (AC) MODEL

This model is used for uncertain data analysis using rough membership function. Transformation of quantitative association rules into linguistic representation can be achieved in discretizing the numerical interval into rough interval described with respective rough membership values. The rough membership values of each linguistic frequent item are composed to form the weighted associative classification rule.

4.1 Capturing uncertainty using linguistic associationmining

Linguistic association mining describes the data attributes in a human understandable language. It conveys the message in linguistic terms instead of numerical values in quantitative attributes. Linguistic associative mining task is divided into two phases.

The first phase is to discretize the continuous numeric data and represent those intervals with a linguistic term. The second phase is of frequent pattern mining. Fuzzy set theory has become the more prominent technique in dealing with linguistic representation on soft boundary interval.

4.2 Phases of Rough Associative Classification Model

The rough associative classification model comprises of three phases.

(i) Generating Rough Membership Value (RMV).
(ii) Association Mining
(iii) Weighted Linguistic Associative Classification Rules Generation.

First phase serves the purpose of inducing rough intervals with the corresponding rough membership values for the linguistic representation and transforms the linguistic decision system.
Second phase transforms the decision system into linguistic Information System.
Third phase aims to generate weighted associative classification rules from the linguistic frequent pattern.

5. DENSITY BASED CLUSTERING

In many applications, the distance between objects has to be computed based on the vague and uncertain data. Commonly, the distance between the uncertain objects descriptions are expressed by one numerical distance value. Based on such single-valued distance functions standard data mining algorithms can work without any change. The key idea of Density Based Clustering is that for each object of a cluster the neighborhood of a given radius Ɛ has to be contain at least a minimum number of μ objects, i.e., the cardinality of the neighborhood has to exceed a given threshold. Some of the basic definitions of Density Based Clustering are as follows.

Definition 1: Core Object
Object o is called a core object with respect to Ɛ and μ in a set of objects D, if $|N_\varepsilon(o)| \geq \mu$,where $N_\varepsilon(o)$ denotes the subset of D contained in the Ɛ-neighborhood of o.
Definition 2: Directly Density-Reachable
Object p is directly density-reachable from object o with respect to Ɛ and μ in a set of objects D, if o is a core object and $p\varepsilon N_\varepsilon(o)$,where again $N_\varepsilon(o)$ denotes the subset of D contained in the Ɛ-neighborhood of o.
Definition 3: Density-Reachable, Density-Connected
An object p is density-reachable from an object o with respect to Ɛ and μ in the set of objects $p_1, p_2, \ldots, pn$, $p1=o$, $p_n=p$ such that $p_i \varepsilon D$ and p_{i+1} is directly density-reachable from p_i with respect to Ɛ and μ.

Object p is density-connected to object q with respect to Ɛ and μ in the set of objects D, if there is an object $o \varepsilon D$ such that both p and q are density-reachable from o with respect to Ɛ and μ in D.

5.1 CIR-DBSCAN

A flat density cluster is defined as a set of density-connected objects which is maximal w.r.t. density-reachability. Then the noise is the set of objects not contained in any cluster. Thus a cluster contains not only the core objects but also the objects that do not satisfy the core object conditions. These border objects are directly density-reachable from at least one core object of the cluster.
CIRDBSCAN, an algorithm based on a representation model of distance distribution between uncertain objects, which uses the Core Influence Rate (CIR) to extend the traditional DBSCAN algorithm in uncertain data.

Algorithm 2: The Procedure of CIR-DBSCAN

1: k ←0
2: for all $O_i \in S$ and is UNVISITED do
3: mark O_i as VISITED
4: L← DDR(O_i)
5: if O_i is not Core Object then
6: mark O_i as NOISE
7: else
8: k←k+1
9: $c_k \in C$←new cluster
10: for all j $O \in L$ do
11: if O_j is UNVISITED then
12: mark O_j as VISITED
13: L′← DDR (O_j)
14: if O_j is not Core Object then
15: L←LUL′
16: end if
17: end if
18: if $O_j \notin C$ then
19: c_k ←add O_j
20: end if
21: end for
22: end if
23: end for

6. PROBABILISTIC QUERIES

In many applications such as sensors for monitoring entities such as temperature and wind speed. A centralized database tracks these entities to enable query processing. Due to continuous change in these values and limited resourses, it is often infeasible to store the exact value all the times. If the degree of the error (uncertainty) between the actual value and the database value is controlled, more confidence in the answers to queries can be placed. To address the issue of measuring the quality of the answers to the queries, and provide algorithm for efficiently pulling the data from relevant sensors or moving objects in order to improve the quality of the executing queries. Answering to the aggregate queries is more challenging than range queries, especially in the presence of uncertainty. The answers to the probabilistic query consists of a set of objects lies in query language. Each object's probability is determined by uncertainty of the object's value and the query range. A probabilistic answer also reflects a certain level of uncertainty that results from the uncertainty of the queries object values. If the uncertainty of all of the objects was reduced, the uncertainty of the result improves

6.1 Classification:Probablistic Queries

There are two ways for classifying the database queries. First, queries can be classified according to the nature of the answers. The second property for classifying queries is whether aggregation is involved.
There are four types of probabilistic queries. They are as follows.

(i) Value-based Non-Aggregate Class.
(ii) Entity-based Non-Aggregate Class.
(iii) Entity-based Aggregate Class.
(iv) Value-based Aggregate Class.

(i) Value-Based Non-Aggregate Class
This class returns the attribute value of an object as the only answer, and involves no aggregate operators.
Example: VSingleQ
(ii) Entity-based Non-Aggregate Class
This type of query returns a set of objects, each of which satisfies the conditions of the query.
Example: ERQ
(iii) *Entity-based Aggregate Class*

This query returns a set of objects which satisfy an aggregate condition.
Example: (EMinQ (EMaxQ)), ENNQ
(iv) Value-based Aggregate Class.
This type of query involves aggregate operators that return a single value.
Example: (VAvgQ(VSumQ)) A VAvgQ(VSumQ)

7. DECISION RULE BASED ON ROUGH SET THEORY

There are many uncertainties of the data in the information systems, which affect the extraction of decision rules. Using rough set to deal with these non-precise data can achieve effective decision-making rules by deleting redundant attributes. The rough set theory algorithm can select the decision rules on the basis of meeting the support and confidence, which can improve the accuracy and reasonableness of the decision rule mining. The rough set theory was developed to deal with vagueness and uncertainty.

7.1 Rough Set Theory

Rough set is a new mathematical tool of dealing with incompleteness and uncertainty knowledge. To determine the knowledge reduction, deduce decision or classification rules classify the given problem or by indiscernibility relationship and classes. In rough set theory, the extent of concept definition can be described by three similar domains including positive domain, negative domain and boundary domain. The domain of the given problem can be divided based on the knowledge. An Information System can be presented as a pair S=<U,A>, or a function f:U*A→V, where U is the non-empty attribute set and V_a is the value set of attribute a such that a:U→ V_a for every aεA. A decision system is any information system of the form S=<U, A ∪ {d}>. The elements of A are called conditional attributes. The universe, U can be partitioned into disjoint subsets known as equivalent classes which are discernible among one another. All the elements in one particular subset are considered equal or indiscernible based on the selected attribute subset.

8. CONCLUSION

The missing attribute value in both the training data and test data performs consistently better in different domains of datasets. The rule sets are easy for people to understand. The rule based algorithm is to classify and derive optimal rule out of highly uncertain data. This algorithm will directly mine the uncertain data. Associative classification model is used for analyzing the uncertain data using rough membership function which is the special case of fuzzy membership function. Rough membership value is for capturing uncertainty in linguistic representation of discretized numerical intervals. The generated weighted linguistic associative classification rule set is used to perform classification operation. The Density based clustering can be carried out based on the vague and uncertain information which occurs in modern application ranges like sensor databases. It benefits from a direct integration of fuzzy distance functions. Probabilistic queries solve the problem of augmenting the probabilistic information to queries over uncertain data. The queries can be used to measure and improve the quality of answers. The decision analysis based rough set theory does not require the completeness of the data, through the division, reduction and membership function. The decision analysis based rough set theory is used for subjective measure of interest for improving the accuracy and robustness of uncertain data.

9. REFERENCES

[1] Oscar Ortega Lobo and Masayuki Numao, "Ordered Estimation of Missing Values," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1999, pp.499-503.

[2] R.Cheng, D.KalaShnikov and Sunil Prabhakar,"Evaluating probabilistic queries over imprecise data," in ACM SiGMOD International Conference on Management of Data, 2003, pp.551-562.

[3] H.Kriegel and M.Pfeifle,"Density- Based Clustering of uncertain data," in IEEE International Conference on Knowledge Discovery and Data Mining (KDD), 2005, pp.672-677.

[4] JingWang, "Estimating Missing Attribute Values using Dynamically-Ordered Attribute Trees," in IEEE International Conference, 2005, pp.164-168.

[5] Biao Qin, Yuni Xia, Sunil Prabhakar and Yicheng Tu, "A Rule-Based Classification Algorithm for Uncertain Data", in IEEE International Conference on Data Engineering,2009,pp.1633-1640.

[6] YunHuoyChoo, AzuralizaAbuBakar, AzahKamilahMuda, "Capturing Uncertainty in Associative Classification Model," in IEEE International Conference on Data Mining and Optimization,2009,pp.84-89.

[7] F.Thabtah, "A Review of Associative Classification Mining," The Knowledge Engineering Review, vol.22:1, pp.37-65, 2007.

[8] Donghua Pan, Lilei Zhao,"Uncertain Data Cluster Based on DBSCAN," in IEEE International Conference on Knowledge Data Engineering, 2011,pp.3781-3784.

[9] Hong Xin Wan, Yun Peng, "The Decision Rule Mining Algorithm of Information System based on Rough Set," in IEEE International Conference,2011,pp.1-3.

[10] H Potamias, G; V. Moustakis; G. Charissis, 1997, "Interactive knowledge based construction and maintenance", Applied Artificial Intelligence, 11, pp. 697-717

[11] Wang Aihua, Guo Wenge, Xu Guoxiong, Jia Jiyou, Wen Dongmao," GIS-Based Educational Decision- Making System" Proceedings of 2009 IEEE International Conference on Grey Systemss and Intelligent Services, November 10-12, 2009, Nanjing, China., 2009 IEEE, pp 1198-1202.

[12] Qiusheng Liu, Guofang Liu," Research on the Framework of Decision Support System Based on ERP Systems", 2010 Second International Workshop on Education Technology and Computer Science, 2010 IEEE.

[13] D. J. Power, "Supporting Decision-Makers: An Expanded Framework", In Harriger, A.(Editor), e- Proceedings Informing Science Conference, Krakow, Poland, June 19-22, 2001, 431-436.