Fuzzy Logic - Retrieval of Data from Database

R.V.S.Lalitha M.Tech., (Ph.D)
Asst. Professor,
Sri Sai Aditya Institute of Science And Tech.,Surampalem.
e-mail: rvslalitha@gmail.com

N.Pavani (M.Tech)
Sri Sai Aditya Institute of Science and Technology,Surampalem.
pavaninarala@gmail.com

Received: 2011-10-21

Accepted: 20 11-10-29

Published: 2011-11-04

Abstract There has been a complicated relationship associated with fuzzy logic and probability theory. All the techniques in fuzzy logic discuss possibility theory and probability theory which measure two kinds of uncertainty. In classical probability theory, a probability measure is a number between 0 and 1. Fuzzy rule based system consists of a set of fuzzy rules with partially overlapping conditions. This paper demonstrates new methodologies for predicting an output, when a particular input "triggers" with multiple fuzzy rules. This paper analyzes the behavior of width of an interval to represent imprecision of the probability estimates. We also propose new applications of possibility theory and probability theory as can be applied to Fuzzy logic.

Introduction: Fuzzy logic is developed by American mathematician Lotfi Zadeh. Fuzzy logic is form of multi-valued logic derived from fuzzy set to deal with approximate values rather precise values. In real, there are only two truth-values: "True" and "False". The fuzzy logic introduces infinite number of truth values between "True" and "False". True can be represented as "1" and False by "0" and any truth value between "True" and "False" lies in the interval 0 and 1.(eg. 0.6) are considered as approximate rather precise. Crisp logic uses binary sets have binary logic i.e.1 for "True" and 0 for "False" to deal with precise information. In contrast with this, fuzzy logic is not constrained to 0 and 1, but also with the degree of truth statement that can range between 0 and 1. Degree of truth is related to probability because both range between 0 and 1. The representation of degree of membership in fuzzy set is not same as an event or a condition in probability. Fuzzy logic and probability are different ways of expressing uncertainty. Fuzzy logic uses the concept of fuzzy set membership (i.e. how much a variable in a fuzzy set) where as a probability uses the concept of subjective (how probable do I think that a variable is in a set) or a condition in probability theory. A rough set, first described by Zdzislaw.I.Pawlak, is a formal approximation of crisp set in terms of a pair of sets which give the lower and the upper approximation of the original set. In rough set theory, he lower and upper approximation sets are crisp sets and approximation sets that lie between lower and upper approximation sets are fuzzy sets. Defuzzification is the process of producing a quantifiable result in fuzzy logic. Fuzzy set will have number of rules that transform a number of variables into a resultant fuzzy set. Thus the resultant fuzzy set is the set whose elements have degree of membership.

In this paper, we proceed with our discussion by taking standard example dataset from data mining techniques [1].Secondly, we discuss how to generate fuzzy sets for the data given. Thirdly, we compute possible occurrence of few events. Fourth, we discuss about the skewness factor for decision analysis based on error rate. Fifth, to take effective decision, we divide the data into subsets using CHAID (Chi-squared Automatic Interaction Detector). Finally, we give membership value for each subset, to proceed to pruning.

By observing the leaf node information, we came to know what records fall under which condition. If the test data set is not satisfying the leaf node information, then pruning is performed on the decision tree, to see that, what all nodes that do not satisfy the criterion are. Based on that, that particular branch is removed. This also happens either with posing wrong query, or with the complexity in the query itself, while dealing with large databases. So, for every query, we need to refer, what is data stored in the training data set, what is obtained after execution of query. Literature says, if the results are not processed properly, then pruning is performed based on the error rate. In this paper, pruning is performed based on the membership value.

1.Decision Tree:

Data mining is all about automating the process of searching for patterns in the data. While dealing with large amount of data, users have to check, whether the results produced are same as the data stored in the original database. Decision tree is a classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given data set. The set of records available for classification methods is generally divided into two disjoint subsets – a training set and a test set. The former is used for deriving the classifier, while latter is used to measure the accuracy of the classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified. Attributes are divided into two different types. 1. Numerical 2 Non-numerical or Categorical. There is one distinguishable attribute called class label. The goal of this classification is to build a concise model that can be used to predict the class of the records whose class label is not known.

We proceed with our discussion by taking standard example dataset from data mining concepts[1].
Consider the training data set X

Outlook	Temp	humidity	windy	class
Sunny	79	90	T	No play
Sunny	56	70	F	play
Sunny	79	75	T	play
Sunny	60	90	T	No play
Overcast	88	88	F	No play
Overcast	63	75	T	play
Overcast	88	95	F	play
Rain	78	60	F	play
Rain	66	70	F	No play
Rain	68	60	T	No play

Table 1. Training Data Set X

Decision tree representation for the dataset (Table 1) is :-

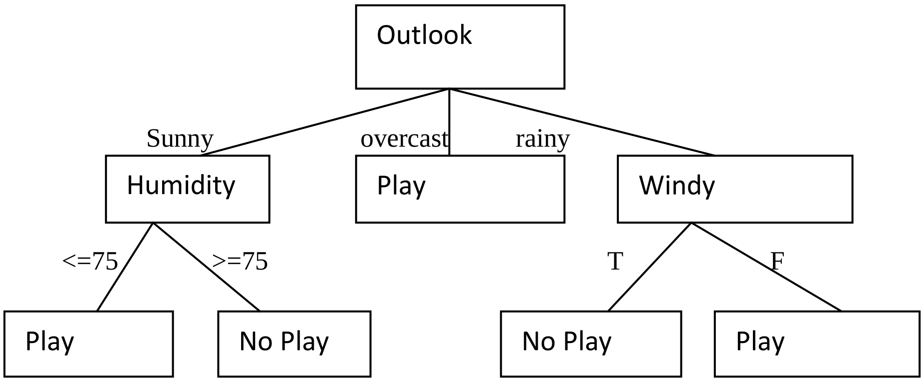


Fig. 1 Decision Tree

From the Fig 1. we can observe that, there five leaf nodes. In decision tree, each node represents a rule. So, the rules corresponding to the decision tree are,

Rule 1: If it is sunny and the humidity is not above 75% then play.

Rule 2: If it is sunny and the humidity is above 75%, then do not play.

Rule 3: If it is overcast, then play.

Rule 4: If it is rainy and not windy, then play.

Rule 5: If it is rainy and windy, then don't play.

2. Fuzzy Referencing:

Now, we apply fuzzy referencing [3], to furnish the details of the dataset in Table 1.

Fuzzy sets provide mathematical meanings to the natural language statements and become an effective solution for dealing with uncertainty. Retrieval of data using fuzzy logic membership functions is well suited to express the intent of a database query when the semantics of the query are rather vague. In this section, we apply a few definitions and concepts from the fuzzy set theory literature.

The attributes in the data set can be represented as objects.

I. Let X={x} set of objects, then

X={"outlook","temp"."humidity","windy","class"}

II. A fuzzy set is characterized by membership function[17]. Eg. membership function for the attribute "outlook" is

$\mu_{outlook}$:X-> [0.0,1.0]

III. Variables in mathematics usually take numerical values. Linguistic variables are non-numeric to represent data in the form of labels.

The attribute outlook ={"Sunny"," Overcast"," Rainy"},which is a subset of Relation R.

IV. Similarly, attribute play={T,F}, a subset of Relation X.

3. Probability Theory and Fuzzy Logic:

The relationship between Probability theory and fuzzy logic has been, and continues to be; an object of controversy [9].Possibility theory is a mathematical theory for dealing with certain types of uncertainty and is an alternative to probability theory. Professor Lotfi Zadeh first introduced possibility theory in 1978 as an extension of his theory of fuzzy sets and fuzzy logic. Possibility can be seen as an upper probability. Any possibility distribution defines a unique set of admissible probability distributions.

Applying the concepts of fuzzy logic and probability theory,

The probability of the attribute *outlook* is

P(outlook)= µoutlook$\int$(x)$\times$ $P_x$dx(x) where x is an element in the set “outlook”.

If x is continuous, or

P(outlook)= $\int$ µoutlook (xi)$\times$ $P_{xi}(x_i)$ dx if x_i is discrete.

Probability of getting humidity >75
P(R)=0.5 obviously probability of getting humidity<=75 is 0.5
An attribute with 0.5 membership is the critical uncertainty point.

Probability of getting windy true is 0.5 and false is 0.5. Again, point of uncertainty is 0.5.

Probability of getting class label play is 0.5 and getting no play is 0.5.Here also, point of uncertainty is 0.5.

If A_1,A_2,…A_K are the attributes of relation X, then decision analysis is based on,
$\sum_{i=1}^{k}$ µAi(xi)

4. Data skew: In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. Consider the tuples that lie to the left and that of tuples that lie to the right of the critical point. The tapering of data on both sides of critical point is not uniform. These tapering sides are called tails. Based on this, data skew can be subdivided into two criterions. 1. Negative skew:-If the left tail is longer, criteria will be based on right side tail, as it has few high values. 2. Positive skew: If the right tail is longer, analysis is done based on left side few high values.

Consider the rule 1 for the illustration of Data skew.

If outlook is sunny and humidity is not above 75 then play
Output:

Outlook	Temp	Humidity	Windy	Class
Sunny	56	70	F	Play
Sunny	79	75	T	Play

Table 2:Tuples that do not fall under the above condition are, rule 2.

OutlooK	Temp	Humidity	Windy	Class
Sunny	79	90	T	No Play

Table 3: If it is sunny and the humidity is above 75%, then do not play.

In our example,(table 3) records that lie to the left of critical point are uniform(nearing to 75) and to that of right are farther from 75. As the critical point is 75, we have to concentrate on left tail tampering rather right.So, skewness factor is required for such analysis.

5. Applying CHAID algorithm for dividing data into subsets: For studying leaf node information, we apply CHAID[1](Chi-squared Automatic Interaction Detector) for dividing attribute into subsets.

CHAID algorithm: -

1. Construct the decision tree, by partitioning the data set into two or more subsets, based on the value of non-class attributes.(eg. humidity).
2. Apply same algorithm, to further partitioning.
3. Each subset is partitioned without regard to any other subset.
4. This process is repeated for each subset until some stopping criterion is met.

Consider the dataset for *humidity* and *class*, the class labels given for *play* are incorrectly classified. This type of errors can be identified in test data set.

Humidity	class
90	No play
70	play
75	play
90	No play

Table 4:Training data set Table

Humidity	class
90	play
70	play
75	No play
90	No play

Table 5:Test data set Table

By observing this, it is understood that there are some inequalities(table 4 and table 5) in either query are labels given. Labeling is based on the attribute values. But there is confusion in query and the data in the database. To coupe up with this, we need to opt for elaborate way of data representation in database. This helps in answering queries in a straight forward way. As we cannot change the query, which is a user need, so, it is advisable to change the representation in one or other way without modifying data in the database. Here is the one solution for representation of data.

1. Label each subset

2. Give class labels based on subset labels

Step 1: Partition the *humidity* attribute into two subsets (A,B).The subset A contains records whose humidity<=75 and subset B contains records whose humidity >75.

In the above training data set table 2 two records are wrongly classified in test data set table 3. We are put in trouble to take a decision whether it is to consider as play or no play. When decision tree is constructed, if a branch is does not satisfy the criterion, then it is simply removed for consideration. When a branch is removed, there will be loss of data. To avoid this, fuzzy rough set approach helps us to decide whether to prune or not to prune.

Step 2: Apply fuzzy rough set approach to classify which records fall under which subset(A, B)

Rule-based systems have sharp cutoffs for continuous attributes [1].For example consider the rule 1. Besides having straight cutoffs fuzzy set allows truth values between 0.0 and 1.0 to represent the degree of membership that a certain value has in a given subset.

Each subset then represents a fuzzy set. Based on the membership value, we can decide, whether to discard or not.

Let set A represents, the attribute values, that satisfy the criterion and set B represents the attributes that does not satisfy the criterion.

A={ 60,70,75} B={88.90,95}

Compute the distances from the decision line(<=75) for the elements that are to the left, and to the right for >75. Represent them in terms of membership value.

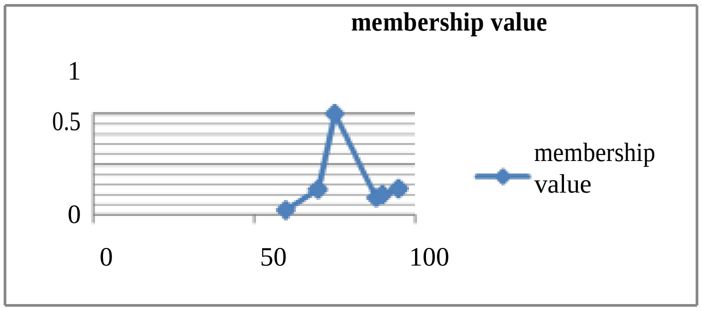
Now membership values for the attribute values are

$M_A(60)=0.05$ $M_B(88)=0.17$

$M_A(70)=0.25$ $M_B(90)=0.2$

$M_A(75)=1.0$ $M_B(95)=0.26$

The graph 1 below represents the membership values corresponding to the attribute values.



Graph 1.

6.Pruning: At each node in a tree it is possible to see the number of instances that are misclassified on a testing set by propagating errors upwards from leaf nodes. This can be compared to the error-rate if the node was replaced by the most common class resulting from that node. If the difference is a reduction in error, then the sub tree at the node can be considered for pruning. This calculation is performed for all nodes in the tree and whichever one has the highest reduced-error rate is pruned. The procedure is then recurred over the freshly pruned tree until there is no possible reduction in error rate at any node.

The queries, when apply on small table of data, may execute correctly. But when applied to large databases, may not work properly because of the following reasons.

1. Complexity of the query.
2. Partial satisfaction of the query
3. Some tuples may not fall under the class labels given
4. Inappropriate formation of the query

While testing, before removing a branch that does not satisfy the criteria, we have to recheck the above mentioned reasons. If the entire above are correct, if still, it is required pruning, pruning rate at which it is to deleted, is decided based on the membership values.

7. Conclusion: Membership values in Fuzzy logic takes a predominant role in the decision analysis. This discussion can be extended very well, whenever, there is a abnormal growth in the database, like population, pay revision etc.

8. References:

1. Arun K Pujari, “Data Mining Techniques”,
2. José Galindo “Introduction and Trends to Fuzzy Logic and Fuzzy Databases” ,University of Málaga, Spain.
3. JOHN YEN and REZA LANGARI ,“Fuzzy Logic”
4.Ramkrishnan ,Gehrke “Database Management Systems”
5.Jiawei Han and Micheline Kamber “Data mining Concepts and Techniques”
6.Andrew W Moore ,“Decision Trees”, Professor
7.“Fuzzy Classification Query Language(fcql)”
8.K.V.S.V.N.Raju A.U. and Arun K.Majumdar ,“ Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems” , University of Guelph,
9.Lotfi A Zadeh,”Probability theory and Fuzzy Logic”.