Data Mining - Miscellaneous Classification Methods

Here in this tutorial we will discuss about the other classification methods such as Genetic Algorithms, Rough Set Approach and Fuzzy Set Approaches.

Genetic Algorithms

The idea of Genetic Algorithm is derived from natural evolution. In Genetic Algorithm first of all initial population is created. This initial population consist of randomly generated rules. we can represent each rule by a string of bits.

For example , suppose that in a given training set the samples are described by two boolean attributes such as A1 and A2. And this given training set contains two classes such as C1 and C2.

We can encode the rule IF **A1 AND NOT A2 THEN C2** into bit string **100**. In this bit representation the two leftmost bit represent the attribute A1 and A2, respectively.
Likewise the rule **IF NOT A1 AND NOT A2 THEN C1** can be encoded as **001**.
Note: If the attribute has K values where K>2, then we can use the K bits to encode the attribute values . The classes are also encoded in the same manner.

Points to remember:

- Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules in the current population and offspring values of these rules as well.
- The fitness of the rule is assessed by its classification accuracy on a set of training samples.
- The genetic operators such as crossover and mutation are applied to create offsprings.
- In crossover the substring from pair of rules are swapped to from a new pair of rules.
- In mutation, randomly selected bits in a rule's string are inverted.

Rough Set Approach

To discover structural relationship within imprecise and noisy data we can use the rough set.

Note: This approach can only be applied on discrete-valued attributes.Therefore, continuous-valued attributes must be discretized before its use.

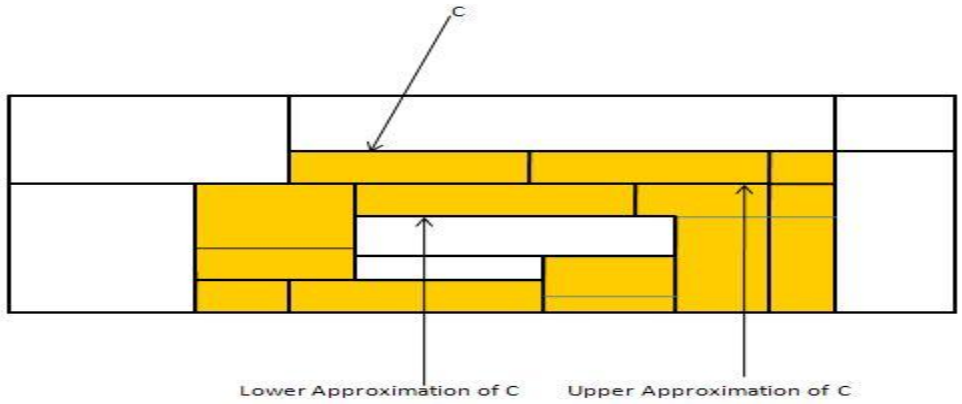
The Rough Set Theory is base on establishment of equivalence classes within the given training data. The tuples that forms the equivalence class are indiscernible. It means the samples are identical wrt to the attributes describing the data.

There are some classes in given real world data, which can not be distinguished in terms of available attributes. We can use the rough sets to **roughly** define such classes.

For a given class, C, the rough set definition is approximated by two sets as follows:

- **Lower Approximation of C** - The lower approximation of C consist of all the data tuples, that bases on knowledge of attribute. These attribute are certain to belong to class C.
- **Upper Approximation of C** - The upper approximation of C consist of all the tuples that based on knowledge of attributes, can not be described as not belonging to C.

The following diagram shows the Upper and Lower Approximation of class C:



Fuzzy Set Approaches

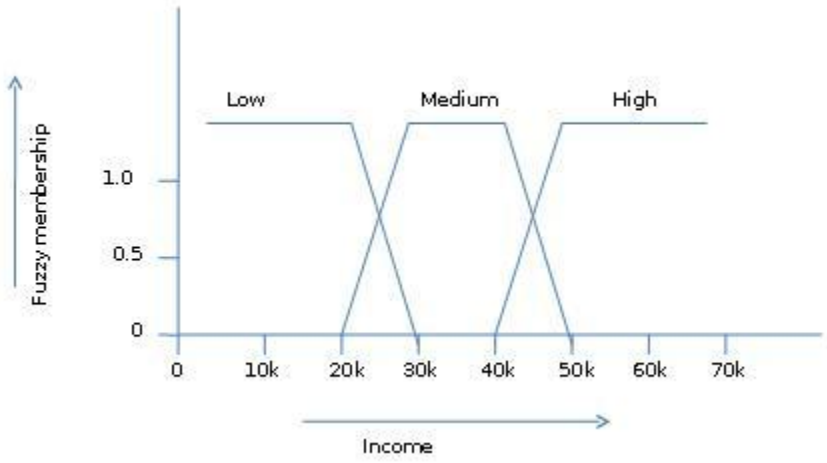
Fuzzy Set Theory is also called Possibility Theory. This theory was proposed by Lotfi Zadeh in 1965. This approach is an alternative **Two-value logic**.This theory allows us to work at high level of abstraction. This theory also provide us means for dealing with imprecise measurement of data.

The fuzzy set theory also allow to deal with vague or inexact facts. For example being a member of a set of high incomes is inexact (eg. if $50,000 is high then what about $49,000 and $48,000). Unlike the traditional CRISP set where the element either belong to S or its complement but in fuzzy set theory the element can belong to more than one fuzzy set.

For example, the income value $49,000 belong to both the medium and high fuzzy sets but to differing degrees. Fuzzy set notation for this income value is as follows:

$$m_{medium\_income}(\$49k)=0.15 \text{ and } m_{high\_income}(\$49k)=0.96$$

where m is membership function that operates on fuzzy set of medium_income and high_income respectively. This notation can be shown diagrammatically as follows:



Source:

http://www.tutorialspoint.com/data_mining/dm_classification_methods.htm