CHAPTER 1

INTRODUCTION

1.1 GENERAL

The steady and amazing progress of Information Technology has shown rapid advances in data collection and storage media. This has enabled the organizations and industries to accumulate vast amount of data in the recently emerged database architecture called data warehouse (Han et al. 2006). In addition to On-Line Analytical Processing (OLAP) tools of a data warehouse for data analysis, characteristics such as Data Classification, Clustering, Reduction, Summarization, Aggregation and Characterization are required for in-depth analysis.

Data Classification (Duda et al. 1973) aims at building a classifier model for describing the important data classes and provides a better understanding of data at large. But this data classification is an ill-defined, non-deterministic task because one cannot be sure that the classifier model developed using the past data will work better for the future data.

This chapter details the task of data classification and its importance in Engineering domain. The various approaches available in the literature for design of classifier model are outlined. The challenges and objectives of designing the Fuzzy Logic Based Classification System using Intelligent Optimization Techniques are discussed.
1.2 DATA CLASSIFICATION AND ITS IMPORTANCE

Classification (Freitas 2002) is one of the most common characteristics of Data Mining. It corresponds to a task that occurs frequently in everyday life. For example, a hospital needs to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness, an opinion polling company wishes to classify people interviewed categorically into those who are likely to vote for each of a number of political parties or are undecided, a teacher intends to classify a student’s performance in examination as first class with distinction, first class and second class.

Similarly many of the Engineering domains namely Pattern Recognition (Lei et al. 2011, Fu et al. 2000), Image Processing (Selvan et al. 2007), Geo-statistics (Kamath et al. 2002), Computer Networks (Drucker 1999, GaneshKumar 2010), Fault Detection (Wu 2004), Medical Diagnosis (Akselrod-Ballin, 2009), Bio-informatics (Wang 2005) and so on also involve data classification. Its importance is evident from the following examples:

- Detecting spam email based upon the message header and content.
- Categorizing cells as malignant or benign based on MRI scans.
- Classifying galaxies as spiral or elliptical based upon their shapes.
- Detecting intrusions based upon the action performed by the user.
- Classifying protein structure using the amino acid sequences.
- Identification of different regions of images using the texture values.
- Categorizing faults in mechanical system based on its performance.
- Classification of dialects in a language using its pronunciation.
- Identification of hand written characters based on the writing style.

In general, the objective of a data classification task is to find a decision boundary or a class separating boundary that partition the input data space into one or more classes. This decision boundary (Oyang et al. 2005) could be either a simple single straight line or a more complicated piece-wise linear or non-linear functional form that depends on the nature of the problem domain. Usually, the data used in day to day problems are simpler and groups together in certain regions of the data space. Hence it is easier to provide a class separating boundary when the groups are disjoint or non-overlapping.

However, the data derived from most of the Engineering domains are only representative in nature (i.e) a sample of some larger, possibly unknown distribution. Many data points that exist in the real world are not included in this small sample set. Therefore any class separating boundary (Williams et al. 2007) must appropriately account for unseen data. This kind of data with overlapping classes poses a lot of difficulties when it is analyzed manually. Hence, the demands for automatic classification of data using computer based approaches are increasing in the current scenario.

When the classes in consideration by a classification system overlap with one another, then the key issue is to find an optimal placement of the discriminant function (Ghorai et al. 2010) that minimizes the number of misclassifications on the given data, and also simultaneously minimizes the probability of misclassification on unseen patterns. The first issue is essentially one of the approximations; the second issue is one of the generalizations.
1.3 DESIGN OF DATA CLASSIFICATION SYSTEM

Data classification is a supervised learning task (Oriols-Puig et al. 2009) that takes labeled data samples and generates a classifier model for classification of new data samples in different predefined groups or classes. Mathematically, this is stated as: given a set of data \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) the objective is to produce a classifier \( h: X \rightarrow Y \) which maps an object \( x \in X \) to its classification label \( y \in Y \). Figure 1.1 depicts the steps followed for designing a data classifier model.

The process of designing a data classifier model starts with data collection for any real world problem and then it requires separation of the data into training and test data. This design process is carried out in two major phases as shown in the Figure 1.2.

![Figure 1.1 Data Classifier Design](image-url)
In training phase, a limited amount of training data and a priori knowledge concerning the problem domain is used to adjust parameters and/or learn the structure of the classifier. In test phase, the classifier designed from the training phase is evaluated on new test data by providing a classification decision for each input pattern.

The designed classifier model (Butchtala et al. 2005) is then evaluated for the following properties:

- **Predictive accuracy** which refers to the ability of the model to correctly predict the class label of new or previously unseen data.
- **Speed** refers to the computation cost involved in generating and using the model.
- **Robustness** refers to the ability of the model to make correct predictions given noisy data or data with missing values.
- **Scalability** refers to the ability to construct the model efficiently given large amounts of data.
• **Interpretability** refers to the level of understanding and insight that is provided by the model.

If these criteria of the classifier model are considered acceptable, then the model can be used to classify future data samples for which the class label is not known.

### 1.4 APPROACHES FOR DATA CLASSIFICATION

During the past history, a number of algorithms have been proposed for solving the classification problem. All those algorithms are grouped under four major categories as below:

- Statistical Approach
- Machine Learning Approach
- Rule Based Approach
- Fuzzy Logic Approach

Each one of these approaches contributes a distinct methodology for addressing the problems in its domain and has its own merits and demerits.

Statistical approaches like naive Bayesian Classifier (Domingos, 1997), weighted voting scheme (Golub, 1999), hidden markov model (Li, 2000), nearest neighbor classification (Domeniconi, 2002), discrimination methods (McLachlan 2004), least square and logistic regression (Fort et al. 2005) are used to generate the data classifier.

These statistical approaches usually result in an inflexible classification system and are unable to classify a data, if it is slightly different from the predefined profile. Moreover, the classifier models produced by them are memory resistant and are not scalable.
Machine learning approaches like Artificial Neural Networks (ANN) (Zhang, 2000) and Support Vector Machine (SVM) (Hsu et al. 2003) have been successfully applied for data classification. Even though these approaches produce good classification accuracy, the results produced by them are hard to interpret. They are popularly called as “Black Box” method as they focus only on the classification performance and do not provide any measure on in-depth understanding of the problem.

Classifiers developed using Case Based Reasoning (CBR) (Arshadi et al. 2005) is instance based. The major problem with this approach is that when data of very large size is provided, it takes a long time for searching and results in high computation cost.

Rule based approaches produce an understandable classifier model with knowledge expressed in the form of rules. Decision tree (Johnson 2002) is used to construct a rule based classifier model. Even though the rules produced by them contain meaningful terms, it is a sensitive type of classifier. Small disturbances in the training sample may lead to large differences in the tree structure. Random Forest (Liaw and Wiener, 2002) that uses ensemble of classification tree is also as sensible as decision tree.

A new symbolic machine learning approach (Wu et al. 2007) has been proposed to extract human understandable rules from decision trees. This approach manipulates symbols on the assumption that such a behavior can be stored in symbolically structured knowledge bases. In practice, the symbolic manipulations limit the situations to which the conventional AI theories can be applied. Because knowledge acquisition and representation are easier by no means, but are arduous tasks.
Although the rule based classifier systems produce simple and interpretable rules, they are not able to completely bring out the hidden information in the data. Moreover they are lacking in robustness with respect to the noisy and missing data. Data Classification followed by knowledge extraction has lot of uncertainties since most of the real world data are vague and uncertain in nature, and are very hard to predict because the boundaries between the output class labels are not well defined.

Recently Fuzzy Logic (FL) (Ross 1995) has been successfully applied in solving classification problems where borders among the data samples pertaining to an output classes are not distinct. The approach considered in FL is to create so-called “fuzzy category memberships functions”, which converts an objectively measurable parameter into a subjective “category memberships”, which are then used for classification.

With the ability of FL that deals with uncertain situation and vagueness (Sivanandam 2006), it seems to be an appropriate approach for classification by extracting knowledge from data.

1.5 CHALLENGES OF FUZZY LOGIC BASED DATA CLASSIFICATION

Fuzzy Logic uses fuzzy-set theory, in which a variable is a member of one or more sets, with a specified degree of membership. When FL is applied for classification of data with overlapping classes, then it will quantify the imprecise information and is helpful to make decisions based on vague and incomplete data. Thus FL allows the data classification system to emulate the human reasoning process.
There are two main categories of Fuzzy Logic Classifiers (Klose et al. 2007) namely Fuzzy classifiers using fuzzy pattern matching (Hamilton-Wright et al. 2007) and Fuzzy classifiers using fuzzy if-then rules (Mansoori, 2009). The former is poorly suited for classification problems because of their lack of normalization and have unacceptable performance where as the latter is more successful because of their ability to incorporate human expert knowledge in decision making.

The fuzzy classification system based on if-then rules, partitions the input data space into regions represented by fuzzy sets in the antecedents (if parts) and class label in the consequents (then parts). Collection of rules (Rule Base) and membership function for each input variable (Data Base) are used as Knowledge Base (Wu 2007) by the fuzzy classifier upon which qualitative reasoning is performed using the fuzzy inference system to derive the conclusion.

Although significant progress has been made in data classification using FL, a number of challenges in applying FL still remains and have not been solved successfully or completely. One of the important challenges in the design of Fuzzy Logic Based Classification System (FLBCS) is knowledge acquisition (Cordon, 2001, Prado et al. 2010) in the form of if-then rules and membership function. In general, the rules and the membership functions used by the FLBCS are formed from the experience of the human experts which is a difficult and time consuming task.

Data-driven approaches (Cordon et al. 2001) have been proposed for developing the FLBCS from numerical data without the knowledge of domain experts. But this approach faces a lack in self-learning and determining the required number of fuzzy if-then rules. Moreover, for classification problems with large number of input variables, the possible
number of rules increases exponentially, which makes it difficult to define a complete rule set. Another way to solve this issue is the incorporation of optimization techniques, which make the FLBCS as a self-learning system.

This research work proposes population based stochastic optimization techniques for extracting near optimal rules and tuning membership functions of a FLBCS.

1.6 OBJECTIVES OF THE RESEARCH WORK

In this research work, the design of a fuzzy classifier system is formulated as a search problem in the solution space where each point represents a rule set, membership function and the respective system behavior. Subsequently, an optimization algorithm is applied to search for an optimal location of this solution space which hopefully represents the near optimal rule set and membership function.

The following population based stochastic optimization techniques are proposed in this research work to develop the classifier for effective data classification.

- Real Coded Genetic Algorithm
- Particle Swarm Optimization
- Hybrid GA-PSO approach
- Water Swirl Algorithm

Each one of the proposed approach contributes a distinct methodology for data classification and this is performed in a cooperative manner rather than a competitive manner. The result is a more intelligent and robust classifier system providing an accurate and human-interpretable result as compared to traditional classification techniques.
Detailed experiments are conducted to analyze the performance of all the proposed approaches. The results obtained by them are compared with the existing previous literatures to prove their level of accuracy.

1.7 ORGANIZATION OF THE THESIS

The thesis is organized into seven chapters including the discussed chapter 1. A brief outline of the forthcoming chapters is given below:

Chapter 2 discusses the design issues of fuzzy logic based classifier system considered in this work. It also provides a detailed summary of the literature review undertaken for this research.

Chapter 3 presents the fuzzy logic based classifier system developed using Genetic Algorithm and the issues to be addressed in developing a Genetic Fuzzy classifier model. The performance of GA approach is tested using the Wisconsin Breast Cancer, Pima Indians Diabetes, Iris, E-coli, Yeast, Magic Comma Telescope, Wine, Glass Identification, Credit Approval and Page Blocks Classification data sets available in the UCI machine learning repository.

Chapter 4 summarizes the weakness of GA in Genetic Fuzzy classifier model and proposes Particle Swarm Optimization (PSO) algorithm as an alternative. The details of development of Swarm based fuzzy classifier model are presented. The performance of the PSO approach is tested using the same data sets and its performance on the designed classifier is noted.

Chapter 5 presents the problem of PSO and possibilities of avoiding those by combining the strength of PSO with the power of GA. The proposed two different forms of hybrid approaches are explained in detail in this chapter. Simulation results are presented to illustrate the usefulness of the proposed hybrid approaches in designing a fuzzy classifier model.
Chapter 6 presents a proposed novel Water Swirl Algorithm (WSA) inspired by the behavior of water that searches the drain inside a sink for designing a fuzzy classifier. Nature of Water and Vortex Particle Theory are reviewed and the details of proposed update equations for improving the performance of fuzzy classifier design are presented. Simulation results show the effectiveness of the proposed approach in producing a compact and interpretable fuzzy rule based system with the highest classification accuracy.

Chapter 7 provides the summary of specific contributions of this research as well as suggestions for future work.