Chapter 1
Introduction

In the recent decade, there has been an increased interest in the development of fuzzy pattern recognition techniques for medical data sets that contributes to solve the problems in early diagnosis and prognosis. This thesis is focused on the investigation and the development of fuzzy pattern recognition techniques viz. clustering and classification algorithms and fuzzy cluster validity index with an application to medical diagnosis. This chapter introduces the background and motivation for this research work and the details of its aims and objectives. The contributions of this thesis are also summarized. Finally, an overview of the remainder of this thesis is provided.

1.1 Background and motivation

Diagnosis and prognosis is the task of medical science and the biomedical diagnosis systems can be implemented in both hardware and software. For the past several decades, researchers in the field of medical data analysis have focused on improving the performance of the existing systems modalities to clinicians (Base, 2004). Software design can assist doctors in interpreting the acquired non-invasive medical information and one of the most important examples of such software is computer-aided diagnosis (CAD) which describes a variety of artificial intelligence techniques applied to medical data. CAD systems work with physicians either by identifying features that might have been overlooked or by providing a second opinion for physicians, making their efforts more efficient.

The most important problems in medical diagnosis and prognosis are (Adlassing et al., 1986, William et al., 2009):
(i) limited observation and subjectivity of the specialist,
(ii) uncertainties and incompleteness in medical knowledge and
(iii) poor time effect in diagnosis

These difficulties have to be recognized during medical decision. The most important problem of medical diagnosis, in general, is the subjectivity of the specialist.
Introduction

As humans are limited in observation, different doctors with different professional levels and clinical experience may have a case to make different diagnostic results, resulting in misdiagnosis. To understand the disease process at its early stage, it is important to perform quantitative analysis in addition to qualitative evaluation of the medical data (both numeric and image). To reduce deaths due to diseases requires early diagnosis and prognosis which requires an accurate and reliable diagnostic procedure. Thus, the use of computers in medical diagnosis and prognosis has become necessary with the increasing size and number of medical data. A large number of Computer Aided Diagnosis (CAD) has been employed in medical diagnosis and prognosis radiology for early diagnosis. Doi (2007) presents a historical review on Computer Aided Diagnosis (CAD) in medical imaging together with the current status and future potential of CAD environment.

Soft Computing refers to a collection of computational techniques in computer science, artificial intelligence, machine learning and some engineering disciplines, which attempts to study, model, and analyze very complex solutions. Its aim, unlike conventional (hard) computing, is to exploit the tolerance for imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness, low solution cost, and close resemblance with human like decision-making [Mitra Sushmita et al. (2006), Jena Rabindra K. et al. (2009)]. The constituents of soft computing are: Fuzzy Logic, Artificial Neural Network (ANN), Evolutionary Algorithms (EAs) [including genetic algorithms (GAs), genetic programming (GPs), and evolutionary strategies (ES)], Support vector machines (SVM), Wavelets, Rough Sets (RS), Simulated Annealing (SA), Swarm Optimization (SO), Memetic Algorithms (MA), Ant Colony Optimization (ACO) and Tabu Search (TS).

Uncertainty is largely present in the medical data sets because of the noise (during the acquisition process). Formalizing the uncertain information upon medical diagnosis and treatment, fuzzy set theory and fuzzy logic provides a number of properties that makes it suitable. A patient can have a set of symptoms which can attribute to several diseases. In observing these symptoms, different doctors with different professional levels and clinical experience may have a case to make different diagnostic results, resulting in misdiagnosis. Also due to the unknown noise in acquisition process, the uncertainty is largely present in medical images. As in medical imaging system, uncertainties can be present at any point resulting from incomplete or imprecise input information, ambiguity in medical images, ill-defined or overlapping boundaries among the disease classes or regions, and indefiniteness in extracting features and relations.
among them, the techniques based on fuzzy clustering plays a vital role in MRI to analyze the patient's data and was introduced by Bezdek (1981) and for the first time Hall et al. (1992) applied this in brain tissue to give visual representation of the original data. Modern neurosurgery takes the advantages of magnetic resonance imaging (MRI) of the patient before the diagnosis procedure (Descoteaux et al., 2008). Any improvement in segmentation methods can lead to important impacts on MR image processing technique.

Pattern recognition activities in medical diagnosis do not depend on a systematized solution but on interpretation of the patient's signal, and that the experiments of the professional is closely related to the final diagnosis. The common pattern recognition techniques are clustering and classification. The distinction between the two approaches is that clustering is unsupervised (Jain et al., 1999, Xu et al., 2005) and has no predefined classes and also does not involve training examples. In contrast, classification is a supervised learning process which is trained on a set of pre-labelled in order to predict into which class new patterns should be placed. The aim of clustering is to group the patterns into clusters based on their similarities. In general there are two main approaches to clustering which is crisp and fuzzy clustering and the fuzzy clustering has been the most successful offspring of fuzzy pattern recognition (Kuncheva, 2010). Several methods based on fuzzy approaches in pattern recognition for medical diagnosis have been developed and amongst them, fuzzy c-mean algorithm is widely used to analysis the medical data (both numeric and image). In the computational intelligence community, several fuzzy set based and hybrid approaches to pattern recognition have been made to attract the considerable attention in medical applications (Ruspini, 1998, Tedeescuet et al., 2000, Fan, J. L. et al., 2003, John, 2003, Pal, 2003, Mitra et al., 2005, Herrera, 2005, Hung Wen-Liang et al., 2006, Moussaoui, 2006, Maji, 2007, Maji, 2008, Jabarajan, 2008, Punitha, 2008, Moein, 2008). One of the most visible hybrid approaches is the integration of neural networks and fuzzy set theory commonly known as neuro-fuzzy (Hall, 1992), is increasingly applied to medical diagnosis. Tedeescuet et al., 2000, present a review report of research activities in fuzzy AI and medicine at USF CSE, including the fuzzy and neuro-fuzzy methods for medical diagnosis such as fuzzy clustering, fuzzy filtering, fuzzy image segmentation and fuzzy expert systems. Another hybrid approach, that plays a prominent role in formalizing uncertainty, vagueness, and incompleteness in diagnosis is the combination of rough set and fuzzy set into the c-means framework commonly known as Rough Fuzzy c-mean
(RFCM) and is successfully applied in the literature (Maji, 2007, Maji, 2008). It is found from the literature that integration of the individual Soft Computing tools help in designing hybrid systems which are more versatile and efficient compared to stand alone use of the tools. Various Soft Computing methodologies and various applications in medicine between the years 2000 and 2008 are briefly presented by Yardimci, 2009.

For formalizing the uncertain information in medical diagnosis fuzzy set theory and fuzzy logic provides a number of properties that makes it suitable. Thus, applying fuzzy pattern recognition technique to medical diagnosis can solve the problems that are inherent in traditional pattern recognition methods for medical data with high rate of subjectivity. Towards this perspective, investigating the fuzzy pattern recognition techniques and hybrid approaches on medical information is extremely promising to improve diagnostic and therapeutic quality of future medical practice.

1.2 Objectives

The ultimate goal of this research work is to establish the variants of fuzzy pattern recognition techniques necessary to develop automated diagnostic tools that will be practical and useful across a wide range of medical domains. This thesis focuses on both the investigation and development of fuzzy pattern recognition techniques that are able to extract the interested and meaningful structure on the medical data (both numeric and image). In order to achieve this aim, the following objectives are identified:

(i) To study the various existing pattern recognition techniques viz., clustering and classification algorithms with an application to medical data.

(ii) To investigate the strength and weakness of the clustering and classification algorithms those are useful across a wide range of medical domains with regard to computational speed and optimality of the result.

(iii) To develop fuzzy set based clustering and classification algorithms that is able to cluster and classify different types of medical data, both numeric and image data.

(iv) To study and investigate cluster validity indices for the fuzzy clustering algorithm and to formulate cluster validity index for fuzzy clustering algorithm.
To evaluate the performance of the proposed techniques and to compare its performance with the existing technique with regard to computational speed and optimality of the result.

1.3 Methodology

(i) **Literature Survey**: Pattern recognition techniques viz., clustering and classification with an application to medical data are extensively studied. Different existing fuzzy set based clustering and classification algorithms for medical diagnosis are also extensively studied to highlight the strength and weakness of the various approaches. Various performance metrics for assessing the quality of the cluster found by fuzzy clustering algorithms and confusion matrix for classification are also investigated.

(ii) **Data Collection**: Medical data may be numeric or image in nature and to conduct experiments on the fuzzy pattern recognition algorithms, data sets are collected from public domain digital repositories in the Internet viz., UCI data sets (www.ics.uci.edu/~mlearn), microarray data sets (Kent Ridge Biomedical Data Set Repository (www.levis.tongji.edu.cn/gzli/data/mirror-kentridge.html)) and brain MR images (http://myweb.msoe.edu/~martynsc/images/mri/mri.html).

(iii) **Implementation of existing algorithms**: The existing algorithms of clustering and classification are implemented in MATLAB environment and are applied on datasets collected from public domain digital repositories in the internet to evaluate the performance of the existing techniques with regard to computational speed and optimality of the result.

(iv) **Development of new algorithms**: A set of clustering and classification algorithms are proposed. The existing algorithms viz., Rough Fuzzy c-Mean (RFCM), Fuzzy Soft Kohonen's Competitive Learning (FSKCL), Fuzzy Soft Learning Vector Quantization (FSLVQ) for clustering and Fuzzy $K$-Nearest Neighbor (FKNN) for classification are enhanced by extending the type-1 fuzzy membership function to the type-2 fuzzy membership function and are called Rough Type-2 Fuzzy c-Mean (RT2FCM), Type-2 Fuzzy Soft Kohonen's Competitive Learning (T2FSKCL), Type-2 Fuzzy Soft Learning Vector Quantization (T2FSLVQ), and Type-2 Fuzzy $K$-Nearest Neighbor (T2FKNN).
Vector Quantization (T2FSLVQ) and Type-2 Fuzzy K-Nearest Neighbor (T2FKNN) respectively.

**Development of the validity index:** A comparative study is made on various validity indices to assess the quality of the cluster found by fuzzy clustering algorithm and a new cluster validity index, SM-index, is proposed for type-2 fuzzy clustering algorithm.

### 1.4 Research contributions

The main contribution of this thesis is in providing a set of generalized fuzzy pattern recognition techniques with an application to medical data. To achieve this goal a detailed investigation of the strength and weakness of the various existing fuzzy pattern recognition techniques viz., clustering and classification algorithm with an application to medical data has been carried out.

Motivated by the accomplishments in the development of fuzzy set theory the existing RFCM, FSKCL and FSLVQ clustering algorithms are enhanced by generalizing the type-1 membership function to type-2 membership function and are called RT2FCM, T2FSKCL and T2FSLVQ respectively.

For the classification of medical data sets, the existing FKNN classification algorithm is enhanced by generalizing the type-1 membership function to the type-2 membership function and is called as T2FKNN.

The thesis also addresses the issue of assessing the quality of the cluster found by fuzzy clustering algorithm and a new cluster validity index, SM-index, is formulated for type-2 fuzzy clustering algorithm.

### 1.5 Overview of the thesis

This thesis continues with the following chapters:

**Chapter 2: Literature review**, includes a review of various fuzzy pattern recognition techniques and the algorithmic approaches that have been applied previously in medical domain (both numeric and MR image) for early diagnosis and also introduces the performance measures such as cluster validity index, image quality index and confusion
matrix to evaluate the results of the fuzzy pattern recognition algorithms. It also presents an overview of the pattern recognition system and the relevance of fuzzy set theory in pattern recognition. Since this thesis focuses on medical applications, an overview of the data taxonomy used in medical is also included in this chapter.

Chapter 3: Fuzzy clustering and classification algorithms, presents a detailed investigation of the strength and weakness of the existing fuzzy pattern recognition algorithms viz., Suppressed Fuzzy c-Means (S-FCM), Modified S-FCM (MS-FCM), RFCM, Fuzzy Kohonen’s Competitive Learning (FKCL), FSKCL, Fuzzy Learning Vector Quantization (FLVQ), FSLVQ and FKNN. A comparative performance analysis of the fuzzy pattern recognition technique is included in this chapter. The results show that the hybrid approaches such as RFCM and fuzzy competitive learning algorithms are more suitable than the individual approaches such as S-FCM and MS-FCM clustering algorithms.

Chapter 4: Type-2 fuzzy clustering and classification algorithms, presents the preliminary theoretical aspects of fuzzy set type-2 and rough fuzzy type-2 set. A set of generalized fuzzy pattern recognition technique viz., type-2 fuzzy c-mean, type-2 fuzzy soft kohonen’s competitive learning and type-2 fuzzy soft learning vector quantization for clustering with an application to Magnetic Resonance Image (MRI) segmentation is proposed. Each of the proposed algorithms extends the membership value of each pattern from type-1 to type-2 membership function of the corresponding rough fuzzy c-mean and fuzzy competitive learning algorithms viz., fuzzy soft Kohonen’s competitive learning and fuzzy soft learning vector quantization. For classification type-2 fuzzy k-nearest neighbor is proposed by extending the membership function of the existing fuzzy k-nearest neighbor algorithm to type-2 membership for classification of the numeric data sets. Results of the experiments conducted with the benchmark datasets to evaluate the performance of the proposed algorithms are presented and the results indicate that the proposed fuzzy type-2 based algorithms outperform the existing algorithms.

Chapter 5: A new cluster validity index for type-2 fuzzy clustering, presents a new fuzzy cluster validity index that is suitable for the type-2 FCM. The experimental results included in this chapter demonstrate the effectiveness of the proposed cluster validity index in determining the appropriate number of clusters.
Chapter 6: Conclusions and future work, presents the summary of the work and draws conclusions. Further, some interesting potential avenues for future research arising from the work presented in this thesis are put forward.