Preface

Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the users. There are a number of data preprocessing techniques: data cleaning, data integration, data transformation and data reduction. The need to cluster large quantities of multi-dimensional data is widely recognized. Cluster analysis is used to identify homogeneous and well-separated groups of objects in databases. It plays an important role in many fields of business and science. Existing clustering algorithms can be broadly classified into four types: partitioning, hierarchical, grid-based, and density-based algorithms. Partitioning algorithms start with an initial partition and then use an iterative control strategy to optimize the quality of the clustering results by moving objects from one group to another. Hierarchical algorithms create a hierarchical decomposition of the given data set of data objects. Grid-based algorithms quantize the space into a finite number of grids and perform all operations on this quantized space. Density-based approaches are designed to discover clusters of arbitrary shapes. These approaches hold that, for each point within a cluster, the neighborhood of a given radius must exceed a defined threshold. Each of the existing clustering algorithms has both advantages and disadvantages. The most common problem is rapid degeneration of performance with increasing dimensions, particularly with approaches originally designed for low-dimensional data. To solve the high-dimensional clustering problem, dimension reduction methods have been proposed which assume that clusters are located in a low-dimensional subspace. However, this assumption does not hold for many real-world data sets. The difficulty of high-dimensional clustering is primarily due to the following characteristics of high-dimensional data:

- High-dimensional data often contain a large amount of noise (outliers). The existence of noise results in clusters which are not well-separated and degrades the effectiveness of the clustering algorithms.
- Clusters in high-dimensional spaces are commonly of various densities. Grid-based or density-based algorithms therefore have difficulty choosing a proper cell size or neighborhood radius which can find all clusters.
- Clusters in high-dimensional spaces rarely have well-defined shapes, and some algorithms assume clusters of certain shapes.
- The effectiveness of grid-based approaches suffers when data points are clustered around a vertex of the grid and are separated in different cells.

To sum up, the classical techniques lack in one way or other as regards to faithful analysis and clustering of multidimensional data owing to inherent uncertainties in assumptions and heuristic choices. It is in this scenario that the soft computing paradigm can be effectively used to arrive at effective and productive throughputs.

The book would surely come to the benefits of several categories of students and researchers. At the students' level, this book can serve as a treatise/reference book for the special papers at the masters level aimed at inspiring possibly future researchers. Newly inducted PhD aspirants would also find the contents of this book useful as far as their compulsory courseworks are concerned. At the researchers' level, those interested in interdisciplinary research would also be benefited from the book. After all, the enriched interdisciplinary contents of the book would always be a subject of interest to the faculties, existing research communities and new research aspirants from diverse disciplines of the concerned departments of premier institutes across the globe. This is expected to bring different research backgrounds (due to its cross platform characteristics) close to one another to form effective research groups all over the world.

The present edited volume comprises 15 well-versed chapters spanning across varied domains of multidimensional data clustering and analysis.

Since the huge volume of data that is generated today, businesses need to have tools to efficiently manage such data. MonetDB is a column-oriented database management system which has shown to have better query processing time with respect to row-oriented systems. Chapter 1 proposes a physical design strategy that improves query execution times in MonetDB. The proposed physical design strategy was empirically studied for 18 TPC-H queries. The experiments were conducted on the basis of cold cache. Each of the queries were executed first using the proposed physical design strategy in this work and then without any physical design. The reported results show that the runtimes using physical design strategy are better for all queries with a minimum percentage improvement of 29%. Also, they showed that the improvement was statistically significant by means of statistical tests.

In Chapter 2, the basics of data clustering and some kind of its applications are given with examples and a real data set. The examples show data types and help to explain basic clustering algorithms. The real data help to classify countries in terms of their computer and internet proficiency levels. After examining the resulting clusters obtained from four different basic methods according to eight computer and internet proficiencies, it is found that the regional closeness of countries and being outside of a union are the major drivers of those clusters' formation. This application gives some possible future research area extensions to researchers about clustering the same or other countries by different proficiencies and unions.

Clustering creates meaningful and useful clusters to analyze and describe the real world for discovering interesting patterns in areas such as, information retrieval, pattern recognition in biological sequence data, etc. Existing clustering approaches suffer from rapid degeneration of performance with increase in dimensions; particularly those are designed for low-dimensional data and due to ineffective cluster evaluation and analysis of multidimensional data owing to inherent uncertainties. ARM is a useful technique that can be used to extract association rules or sets of frequent patterns. These AR leads to potential knowledge to detect the regularities and path in large databases for designing scalable association-rule mining algorithms that will find the number of records in a cluster but also the number of dimensions in a cluster to focus on the domains. Chapter 3 illustrates the fidelity and properties of the ARM technique with recourse to its applications in detail.

Chapter 4 covers data clustering in detail, which includes; introduction to data clustering with figures, data clustering process, basic classification of clustering and applications of clustering, describing hard partition clustering and fuzzy clustering. Some most commonly used clustering methods are also explained in the chapter with their features, advantages, and disadvantages. A variant of K-Means and extension method of hierarchical clustering method, density-based clustering method and grid-based clustering method are also covered.

Chapter 5 proposes a content based image retrieval method dealing with higher dimensional feature of images. The kernel principal component analysis (KPCA) is done on MPEG-7 Color Structure Descriptor (CSD) (64-bins) to compute low-dimensional nonlinear-subspace. Also the Partitioning Around Medoids (PAM) algorithm is used to squeeze search space again where the number of clusters are counted by optimum average silhouette width. To refine these clusters further, the outliers from query image's belonging cluster are excluded by Support Vector Clustering (SVC). Then One-Class Support Vector Machine (OCSVM) is used for the prediction of relevant images from query image's belonging cluster and the initial retrieval results based on the similarity measurement is feed to OCSVM for training. Images are ranked from the positively labeled images. This method gives more than 95% precision before recall reaches at 0.5 for conceptually meaningful query categories. Also comparative results are obtained from (1) MPEG-7 CSD features directly and (2) other dimensionality reduction techniques.

Data mining has great contributions to the healthcare such as support for effective treatment, healthcare management, customer relation management, fraud and abuse detection and decision making. The common data mining methods used in healthcare are Artificial Neural Network, Decision trees, Genetic Algorithms, Nearest neighbor method, Logistic regression, Fuzzy logic, Fuzzy based Neural Networks, Bayesian Networks and Support Vector Machines. The most used task is classification. Because of the complexity and toughness of medical domain, data mining is not an easy task to accomplish. In addition, privacy and security of patient data is a big issue to deal with because of the sensitivity of healthcare data. There exist additional serious challenges. The objective of Chapter 6 is to provide a descriptive study aimed to provide an acquaintance to data mining and its usage and applications in healthcare domain. The use of data mining in healthcare informatics and challenges is also examined.

CAD is a relatively young interdisciplinary technology, which has a tremendous impact on medical diagnosis, specifically cancer detection. The accuracy of CAD to detect abnormalities on medical image analysis requires a robust segmentation algorithm. To achieve accurate segmentation, an efficient edge-detection algorithm is essential. Medical images like USG, X-Ray, CT and MRI exhibit diverse image characteristics but are essentially collection of intensity variations from which specific abnormalities are needed to be isolated. In Chapter 7, a robust medical image enhancement and edge detection algorithm is proposed, using tree-based adaptive thresholding technique. It has been compared with different classical edge-detection techniques using one sample two tail t-test to examine whether the null hypothesis can be supported. The proposed edge-detection algorithm shows 0.07 p-values and 2.411 t-stat where $a = 0.025$. Moreover, the proposed edge is single pixeled and connected which is very significant for medical edge detection.

Different graph theoretic approaches are prevalent in the field of image analysis. Graphs provide a natural representation of image pixels exploring their pairwise interactions among themselves. Graph theoretic approaches have been used for problem like image segmentation, object representation, matching for different kinds of data. Chapter 8 mainly aims at highlighting the applicability of graph clustering techniques for the purpose of image segmentation. Different spectral clustering techniques like minimum spanning tree based data clustering, Markov Random Field (MRF) model for image segmentation are discussed in this respect.

Data mining has been a popular technique in many applications. In Chapter 9, the authors focus on utilizing architecture features and apply data mining techniques for computer design. Since data mining requires lot of collected data for building models, the relevant data needs to be properly generated. The authors demonstrate the applications and their methodology starting from data set generation, feature extraction, modeling and evaluations. Important characteristics of the architecture are considered for data set generation and feature extractions: particularly, the instruction set, and memory access pattern features. The chapter utilizes these features given with observations for building the models for cache prediction, branch prediction, and malware detection.

In Chapter 10, a saliency and phase congruency based digital image watermarking scheme has been projected. The planned technique implants data at least significant bits (LSBs) by means of adaptive replacement. Here more information is embedded into less perceptive areas within the original image determined by a combination of spectral residual saliency map and phase congruency map. The position of pixels with less perceptibility denotes the most unimportant region for data hiding from the point of visibility within an image. Therefore, any modification within these regions will be less perceptible to one observer. The model gives a concept of the areas which has excellent data hiding capacity within an image. Superiority of the algorithm is tested through imperceptibility, robustness, along with data hiding capacity.

In the medical field diagnosis of a disease at an early stage is very important. Nowadays soft computing techniques such as fuzzy logic, artificial neural network and Neuro- fuzzy networks are widely used for the diagnosis of various diseases at different levels. In Chapter 11, a hybrid neural network is designed to classify the heart disease data set the hybrid neural network consist of two types of neural network multilayer perceptron (MLP) and fuzzy min max (FMM) neural network arranged in a hierarchical manner. The hybrid system is designed for the dataset which contain the combination of continuous and non-continuous attribute values. In the system the attributes with continuous values are classified using the FMM neural networks and attributes with non-continuous value are classified by using the MLP neural network and to synthesize the result the output of both the network is fed into the second MLP neural network to generate the final result.

Due to microarray experiment imperfection, spots with various artifacts are often found in microarray image. A more rigorous spot recognition approach in ensuring successful image analysis is crucial. Chapter 12 proposes a novel hybrid algorithm for this purpose. A wavelet approach is applied, along with an intensity-based shape detection simultaneously to locate the contour of the microarray spots. The proposed algorithm is able to segment all the imperfect spots accurately. Performance assessment with the classical methods, i.e., the fixed circle, adaptive circle, adaptive shape and histogram segmentation shows that the proposed hybrid approach outperforms these methods.

Enhancing the energy efficiency and maximizing the networking lifetime are the major challenges in Wireless Sensor Networks (WSN). Swarm Intelligence based algorithms are very efficient in solving nonlinear design problems with real-world applications. In Chapter 13, a Swarm based Fruit Fly Optimization Algorithm (FFOA) with the concept of K-Medoid clustering and swapping is implemented to increase the energy efficiency and lifetime of WSN. A comparative analysis is performed in terms of cluster compactness, cluster error and convergence. MATLAB Simulation results show that K-Medoid Swapping and Bunching Fruit Fly optimization (KMSB-FFOA) outperforms FFOA and K-Medoid Fruit Fly Optimization Algorithm (KM-FFOA).

Chapter 14 aims to study the use of Hybridization of intelligent techniques in the areas of bioinformatics and computational molecular biology. These areas have risen from the needs of biologists to utilize and help interpret the vast amounts of data that are constantly being gathered in genomic research. It also describes the kind of methods which were developed by the research community in order to search, classify and mine different available biological databases and simulate biological experiments. This chapter also presents the hybridization of intelligent systems involving neural networks, fuzzy systems, neuro-fuzzy system, rough set theory, swam intelligence and genetic algorithm. The key idea was to demonstrate the evolution of intelligence in bioinformatics. The developed hybridization of intelligent techniques was applied to the real world applications. The hybridization of intelligent systems performs better than the individual approaches. Hence these approaches might be extremely useful for hardware implementations.

With the increasing volume of data, developing techniques to handle it has become the need of the hour. One such efficient technique is clustering. Data clustering is under vigorous development. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. Several data clustering algorithms have been developed in this regard. Data is uncertain and vague. Hence uncertain and hybrid based clustering algorithms like fuzzy c means, intuitionistic fuzzy c means, rough c means, rough intuitionistic fuzzy c means are being used. However, with the application and nature of data, clustering algorithms which adapt to the need are being used. These are nothing but the variations in existing techniques to match a particular scenario. The area of adaptive clustering algorithms is unexplored to a very large extent and hence has a large scope of research. Adaptive clustering algorithms are useful in areas where the situations keep on changing. Chapter 15 details some of the adaptive fuzzy c means clustering algorithms which are widely used in a variety of applications especially image processing.

The primary objective of the present endeavor is to bring a broad spectrum of multidimensional data clustering and data analysis applications under the purview of hybrid intelligence so that it is able to trigger further inspiration among various research communities to contribute in their respective fields of applications thereby orienting these application fields towards intelligence. Once the purpose, as stated above, is achieved a larger number of research communities may be brought under one umbrella to ventilate their ideas in a more structured manner. In that case, the present endeavor may be seen as the beginning of such an effort in bringing various research applications close to one another. By academically coming closer to one another, research communities working in diversified application areas involving multidimensional data viz. true color images, videos, big data, would be more encouraged to form groups among themselves paving way for interdisciplinary research. Speaking from the scholastic view, this is a formidable achievement in which the present endeavor may be thought of as the maiden

facilitator. It may however be noted that there are good amounts of contributions of the application of hybrid soft computing in various fields. However, any such previous effort has remained application specific, that is, aimed at identifying a specific application domain where the ingredients of hybrid soft computing have been applied quite effectively. But, to the best of our knowledge, efforts to bring in multiple domains of multidimensional data within one framework are not very frequent. In that sense, this appears to be the first such effort to accommodate cross platform applications of hybrid soft computing. Moreover, efforts of hybridization are very meager in the literature. Once successful, this will become an encouragement towards further research of interdisciplinary nature by providing scope to various research communities to come together through such an effort.

Siddhartha Bhattacharyya
RCC Institute of Information Technology, India

Sourav De
Cooch Behar Government Engineering College, India

Indrajit Pan
RCC Institute of Information Technology, India

Paramartha Dutta
Visva-Bharati University, India