ABSTRACT

Advancements in technology produce huge amounts of data in various fields, increasing the need for efficient and effective data mining tools to uncover the information contained implicitly in the data. Such a voluminous store of data of diverse characteristics is mostly stored and made available in digitized form. With such a phenomenal increase in the storage and availability of data, it is imperative that the available data is well-organized to generate useful knowledge. Data mining methodologies were employed to interpret the voluminous data at a faster pace and to classify the same with greater reliability. Machine learning techniques have been widely applied to accurately predict the target class for each case in the data. In this thesis, learning approaches are proposed, integrating several machine learning algorithms with optimization techniques in a synergistic way to maximize the effectiveness of a learning task. The machine learning techniques like Decision Trees, Neural Networks, Support Vector Machines and optimization algorithms namely Biogeography Based Optimization, Particle Swarm Optimization and Artificial Bee Colony Optimization were used for classification and transformation of data into useful information. The underlying problems with large datasets are that they result in low classification accuracy and high computational time. In this research, effective hybrid schemes are proposed to classify data, text, blur, image with increased classification accuracy and network convergence speed in comparison with traditional schemes.

The data classification was addressed with hybrid Error Correction Output Code (ECOC) and hybrid Recursive ECOC generation methods to
decompose the multi-class data to several two-class data. The Biogeography Based Optimization (BBO) algorithm has been proposed with the code generation methods in the present study to improve the binarization process. The generated code matrix is best for a given problem because it is designed taking into account the features of the code matrix such as overall classifier accuracy, minimum hamming distance and margin of classification, and the features of the problem such as attributes, samples and classes. The generated multiple binary classes were classified using standard Support Vector Machine (SVM), C 5.0 Decision Tree algorithm and Radial Basis Function Neural (RBFN) Network. The code matrix generated using hybrid BBO based ECOC and RECOC schemes with base learner improve the classification accuracy and the speed of the convergence. The Ecoli and Pendigits datasets were taken from U.C.Irvine Machine learning repository, University of California, U.S.A for testing the effectiveness of the proposed schemes.

The text classification from a corpus was done using enhanced High Dimensional Growing Self Organization Map with Randomness (HDGSOM(r)) algorithm and Dynamic Adaptive Self Organization Map (DASOM). After removing the stop words, two adjacent words are combined and they are called as technical terms. The features of each technical terms are derived using Term Frequency–Inverse Document Frequency (TF-IDF), Jaccard’s Similarity Coefficient (JSC), Simple Appearance Ratio. In order to improve convergence speed and classification accuracy of the proposed hybrid network to identify emergent trends, the optimum weight vectors are derived using Particle Swarm Optimization (PSO) algorithm. The influence of the PSO on unsupervised machine learning algorithms such as DASOM, HDGSOM(r) in the design of an efficient ETD schemes were tested with
Digital Bibliography and Library Project (DBLP) database which is a computer science bibliography website hosted at University Trier, in Germany.

Restoration of a high quality image from a degraded recording is an important problem in vision processing. System degradations such as blur due to optical system motion in the imaging setup can be classified using neural networks and they are restored using image restoration technique. An accurate estimate of Point Spread Function (PSF) parameter is one of the crucial features which characterize the performance of a blur identification scheme. In blur classification scheme, the identification of various blur and its parameters present in a blurred image are taken into account. The blur classification was performed using proposed hybrid Biogeography Based Optimization (BBO) based Multi-Layer neural network with Multi-Valued Neuron (MLMVN). The multi value neurons are working under the principle of multiple valued threshold logic over the field of the complex numbers. The optimum weight vector for the MLMVN network is derived using BBO algorithm to achieve better performance in terms of classification accuracy and convergence speed. The derivative-free back propagation learning algorithm is used in MLMVN network. The collection of images for training and testing the hybrid schemes have been taken from the Research center, University of California, Berkeley (US).

Image classification from a vast collection has become one of the interesting challenges and the development of image classification approaches based on low-level descriptors, such as color, texture and shape has drawn attention of researchers. The main idea of image classification schemes are used to analyze image information and to set up feature vectors of an image as
its index. In addition, these image classification schemes are designed to reduce the semantic gap between low level feature and Human Visual Perception (HVP). In this research, Artificial Bee Colony optimization (ABC) based Fuzzy Radial Basis Function Neural network (FRBFN) and Multi-feature Similarity Score Fusion (MSSF) methods are proposed for image classification by combining the color and texture features of the image to retrieve the images from the dataset. The proposed hybrid image classification schemes achieved high performance in terms of classification accuracy, precision and recall rate with minimum computational time. In the FRBFN network a Fuzzy c-means (FCM) algorithm is used to locate cluster centers in the high dimensional space of image features. The images from the COREL database with various percentages of instances are taken for training and testing of the proposed hybrid ABC based FRBFN network.

The important performance measures used to evaluate the classification effectiveness of proposed schemes for classifying data, text, blur and images are accuracy, computational time such as training and testing time. The investigations indicate the effectiveness of the proposed approach using BBO based ECOC and RECOC with RBFN for data, BBO based MLMVN network for blur classification, PSO based HDGSOM(r) and DASOM neural network for text classification and ABC based FRBFN network for color and texture based image classification. The proposed hybrid schemes for classification are applicable in many fields of technology such as medical diagnosis, customer segmentation, fraud detection, direct marketing, credit analysis, security, biomedical and drug response modeling to achieve effective results.