ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Mehmet ACI

DEVELOPMENT OF A HYBRID CLASSIFICATION METHOD FOR MACHINE LEARNING

DEPARTMENT OF COMPUTER ENGINEERING

ADANA, 2009

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DEVELOPMENT OF A HYBRID CLASSIFICATION METHOD FOR MACHINE LEARNING

Mehmet ACI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİMDALI

Bu tez ..../...../…... tarihinde aşağıdaki jüri üyeleri tarafından oybirliği/oyçokluğu ile kabul edilmiştir.

İmza……..	İmza……..	İmza……..
Yrd.Doç.Dr.Mutlu AVCI DANIŞMAN	Doç.Dr.Mustafa GÖK ÜYE	Yrd.Doç.Dr.Ulus ÇEVİK ÜYE

Bu tez Enstitümüz Bilgisayar Mühendisliği Anabilim Dalında hazırlanmıştır.
Kod No:

Prof. Dr. Aziz ERTUNÇ
Enstitü Müdürü
İmza ve Mühür

Bu Çalışma Çukurova Üniversitesi Bilimsel Araştırma Projeleri Birimi Tarafından Desteklenmiştir.

Proje No: MMF2009YL36

ÖZ

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ İÇİN HİBRİT BİR SINIFLAMA METODU GELİŞTİRİLMESİ

Mehmet ACI

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Danışman: Yrd.Doç.Dr. Mutlu AVCI
Yıl: 2009, Sayfa: 44
Jüri: Yrd.Doç.Dr. Mutlu AVCI
Doç.Dr. Mustafa GÖK
Yrd.Doç.Dr. Ulus ÇEVİK

Bu çalışmada "Bayes, K En Yakın Komşu Metotları ve Genetik Algoritma Kullanılarak Hibrit Sınıflama" ve "Kestirim Eniyileme Tabanlı Sınıflama Metodunda K En Yakın Komşu Metodundan Faydalanılması" adlarında iki araştırma yapılmıştır. İlk araştırmada k en yakın komşu, Bayes metotları ve genetik algoritma kullanılarak birlikte kullanılarak hibrit bir metot oluşturulmuştur. Amaç öğrenmeyi zorlaştıran verileri eleyerek sınıflamada mükemmel sonuçlara ulaşmaktır. Önerilen metot üç ana adımda uygulanmıştır. İlk adımda mevcut verilerle yeni veriler oluşturulmuş ve k en yakın komşu metodu ile iyi olanları seçilmiştir. İkinci adımda ise seçilen veriler genetik algoritma ile işlenmiş ve daha iyi veri kümeleri oluşturulmuştur. Son olarak en iyi veri kümesi belirlenmiş ve sınıflamadaki başarısını belirlemek için Bayes metodu ile işlenmiştir. Ayrıca orijinal ve en iyi veri kümeleri tarafsız bir değerlendirme için yapay sinir ağlarında test edilmiştir. İkinci araştırmada, veri sınıflamasını iyileştirmek için bir veri eleme yaklaşımı önerilmiştir. Bayes ve k en yakın komşu metotlarında yapılan düzenlemelerle hibrit bir metot oluşturulmuştur. Ana fikir k en yakın komşu metodu ile veri sayısını azaltmak ve en benzer eğitim verileri ile sınıfı tahmin etmektir. Sonrasında Bayes metodunun kestirim eniyileme algoritması kullanılmıştır. K en yakın komşu Bayes sınıflayıcısının önişlemcisi olarak belirlenmiş ve sonuçlar gözlemlenmiştir. Test işlemleri, University of California Irvine (UCI) makine öğrenmesi veri kümelerinin en bilinenlerinden beşi olan Iris, Breast Cancer, Glass, Yeast ve Wine ile yapılmıştır.

Anahtar Kelimeler: Bayes metodu, k en yakın komşu metodu, genetik algoritma, yapay sinir ağları, sınıflama.

ABSTRACT

MSc THESIS

DEVELOPMENT OF A HYBRID CLASSIFICATION METHOD FOR MACHINE LEARNING

Mehmet ACI

COMPUTER ENGINEERING
INSTITUTE OF NATURAL AND APPLIED SCIENCES
UNIVERSITY OF CUKUROVA

Supervisor: Assist.Prof.Dr. Mutlu AVCI
Year: 2009, Pages: 44
Jury: Assist.Prof.Dr. Mutlu AVCI
Assoc.Prof.Dr. Mustafa GÖK
Assist.Prof.Dr. Ulus ÇEVİK

In this work two studies are done and they are referred as first study which is named "A Hybrid Classification Method Using Bayesian, K Nearest Neighbor Methods and Genetic Algorithm" and second study which is named "Utilization of K Nearest Neighbor Method for Expectation Maximization Based Classification Method". A hybrid method is formed by using k nearest neighbor (KNN), Bayesian methods and genetic algorithm (GA) together at first study. The aim is to achieve successful results on classifying by eliminating data that make difficult to learn. Suggested method is performed at three main steps. At first step new data is produced according to available data, and then right data is chose with KNN method. At second step chosen data is processed with GA and better data sets are generated. Finally, best data set is determined and processed with Bayesian method to specify the success on classifying. Also the original and best data sets are tested on artificial neural networks (ANN) for an unbiased evaluation. In second study a data elimination approach is proposed to improve data clustering. A hybrid algorithm is formed with modifications on Bayesian and KNN methods. Main idea is to reduce the number of data with KNN method and to guess a class with most similar training data. The rest is same as expectation maximization (EM) algorithm of Bayesian method. KNN method considered as the preprocessor for Bayesian classifier and then the results over the data sets are investigated. Test processes are done with five of well-known University of California Irvine (UCI) machine learning data sets. These are Iris, Breast Cancer, Glass, Yeast and Wine data sets.

Keywords: Bayesian method, k nearest neighbor method, genetic algorithm, artificial neural network, classifying.

ACKNOWLEDGMENT

First of all, I would like to thank my advisor Assist.Prof.Dr. Mutlu AVCI for his supervision guidance, encouragements and his valuable time for this work.

I would also like to thank members of MSc thesis jury Assoc.Prof.Dr. Mustafa GÖK and Assist.Prof.Dr. Ulus ÇEVİK for their suggestions and corrections.

Furthermore, I would like to thank my family; my parents Turan, Kamet, my sister Havva and brother Can for their endless support and encouragements for my life and career.

Last but not the least; I would like to thank my love and wife Çiğdem for her endless support and patience.

CONTENTS PAGE

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

1. INTRODUCTION

Machine learning is a knowledge area starting at the point when data are explained or estimations are produced for the future. It generates functional approximation or classification models for the data. To convert the learning studies to machine learning some paradigms and approaches are used. Symbolic processing like decision trees and version spaces, connectionist systems, statistical pattern recognition, case based learning, evolutionary programming and GAs are some of them (Anonymous). Various methods and algorithms form the base of machine learning. Everyday new ones are added to these methods and algorithms or existing is developed. At the machine learning the aim is to realize the human learning job by computers. Various methods and algorithms are used during this learning. KNN, Bayesian methods and GAs are several of them. One of the goals of these methods and algorithms is to find out the class of new data when the information about the classes of past data is given. This process is named as classifying (Amasyali, 2006).

Data classification is a kind of data analysis form, which can forecast data trend in future. The objective of data classification classified to some events or objects; it mainly analyzes the historical data whose category is known, and then summarizes a classified model (Lu et al, 2008).

There are two studies are explained in this thesis. These studies are referred as first study which is named “A Hybrid Classification Method Using Bayesian, K Nearest Neighbor Methods and Genetic Algorithm” and second study which is named “Utilization of K Nearest Neighbor Method for Expectation Maximization Based Classification Method”.

However, many applications had done on KNN, Bayesian methods and GAs. A hybridization of these three methods is not taken place in literature. Suggested approach at first study brings a new point of view. A hybrid method is formed by using KNN, Bayesian methods and GA together at this first study. The aim is to achieve successful results on classifying by eliminating data that make difficult to learn. Forming new data set approach is proposed according to good data on the hand. Suggested method is performed at three main steps. At first step new data is

produced according to available data, and then right data is chose with KNN method. At second step chosen data is processed with GA and better data sets are generated. Finally, best data set is determined and processed with Bayesian method to specify the success on classifying. Also the original and best data sets are tested on ANNs for an unbiased evaluation.

In second study a data elimination approach is proposed to improve data clustering. The proposed method is based on hybridization of KNN and Bayesian methods. A hybrid algorithm is formed with modifications on Bayesian and KNN methods. Main idea is to reduce the number of data with KNN method and to guess a class with most similar training data. The rest is same as EM algorithm of Bayesian method. Briefly; KNN considered as the preprocessor for Bayesian classifier and then the results over the data sets are investigated.

Test processes are evaluated with five of well-known UCI machine learning data sets (Yildiz et al, 2008). Those are Iris, Breast Cancer, Glass, Yeast and Wine data sets. Test results are investigated in collaboration with the previous works, and the success of the study is considered.

2. LITERATURE REVIEW

2.1. Previous Studies on K Nearest Neighbor Method

At KNN method a constant k value is chosen. Baoli et al had selected different k values for each class instead of constant k value, and by that way they had done more sensitive measurements. More samples (nearest neighbors) had used for deciding whether a test document should be classified to a category, which has more samples in the training set. Experiments on Chinese text categorization has shown that their method was less sensitive to the parameter k than the traditional one, and it can properly classify documents belonging to smaller classes with a large k (Baoli et al, 2003). Yildiz et al had used KNN algorithm to develop an individual filtering model to determine whether the e-mail is spam or not. The developed classifier was useful to make decision. They had provided parallelism to shorten the spent time (Yildiz et al, 2008). Song et al had used instructive data as criterion when determining k value at their study that was called informative KNN pattern classification. They had introduced a new metric that measures the informativeness of objects to be classified. When applied as a query-based distance metric to measure the closeness between objects, Locally Informative-KNN (LI-KNN) and Globally Informative-KNN (GI-KNN) had proposed. By selecting a subset of most informative objects from neighborhoods, their methods had exhibited stability to the change of input parameters, number of neighbors (K) and informative points (I) (Song et al, 2007).

Cucala et al had proposed a reassessment of KNN procedure as a statistical technique derived from a proper probabilistic model. They had modified the assessment made in a previous analysis of this method. They had established a clear probabilistic basis for the KNN procedure and derived computational tools for conducting Bayesian inference on the parameters of the corresponding model. Their new model then provides a sound setting for Bayesian inference and for evaluating not just the most likely allocations for the test dataset but also the uncertainty that goes with them (Cucala et al, 2009). Kubota et al had proposed a hierarchical KNN

classification method using the feature and observation space information. With this method they had performed a fine classification when a pair of the spatial coordinate of the observation data in the observation space and its corresponding feature vector in the feature space is provided (Kubota et al, 2008). Anbeek et al had developed a method that uses a KNN classification technique with features derived from spatial information and voxel intensities. They had used this method for segmentation of four different structures in the neonatal brain: white matter (WM), central gray matter (CEGM), cortical gray matter (COGM), and cerebrospinal fluid (CSF). The segmentation algorithm was based on information from T2-weighted (T2-w) and inversion recovery (IR) scans. The method had resulted in high sensitivity and specificity for all tissue classes. The probabilistic outcomes had provided a useful tool for accurate volume measurements. The described method was based on routine diagnostic magnetic resonance imaging (MRI) and was suitable for large population studies. (Anbeek et al, 2008).

Blanzieri and Melgani had developed a new method that inherits the attractive properties of both the KNN and the support vector machine classifiers. They had presented a new variant of the KNN classifier based on the maximal margin principle and exposed the advantages of new method. The proposed method had relied on classifying a given unlabeled sample by first finding its k-nearest training samples. A local partition of the input feature space was then carried out by means of local support vector machine (SVM) decision boundaries determined after training a multiclass SVM classifier on the considered k training samples. The labeling of the unknown sample had done by looking at the local decision region to which it belongs. The method had characterized by resulting global decision boundaries of the piecewise linear type (Blanzieri and Melgani, 2008). Weinberger and Saul had studied how to improve nearest neighbor classification by learning a Mahalanobis distance metric. They had extended the original framework for large margin nearest neighbor (LMNN) classification with three contributions. First, they had described a highly efficient solver for the particular instance of semidefinite programming that arises in LMNN classification. Second, they had shown how to reduce both training and testing times using metric ball trees; the speedups from ball trees are further

magnified by learning low dimensional representations of the input space. Third, they had shown how to learn different Mahalanobis distance metrics in different parts of the input space (Weinberger and Saul, 2008). Zeng et al had proposed a new variant of the KNN classification rule, a nonparametric classification method based on the local mean vector and class statistics. In this new classification method not only the local information of the KNNs of the unclassified pattern in each individual class but also the global knowledge of samples in each individual class are taken into account. Then the proposed classification method had compared with the KNN and the local mean-based nonparametric classification in terms of the classification error rate on the unknown patterns (Zeng et al, 2008).

2.2. Previous Studies on Bayesian Method

Bayesian method based works are also popular. Ozcanli et al had modeled the relations statistically between image segments and word set at an explained database by using translation with computer method. They had used the relations to label the given segments. While statistical modeling they had used EM algorithm and Bayesian method forms the base for it. They had concluded that the performance of the proposed method was dependent to used attributes and their importance according to each other (Ozcanli et al, 2003). Kotsiantis et al had improved the performance of the Naive Bayes MultiNomial Classifier. They had combined Naive Bayes MultiNomial with Logitboost. They had modified Naive Bayes MultiNomial classifier in order to run as a regression method. They had performed a large-scale comparison with other algorithms on 10 standard benchmark datasets and taken better accuracy in most cases (Kotsiantis et al, 2006).

Gungor had developed three different models that use Bayesian algorithms and ANNs to filter Turkish spam messages. These models' names are binary model, probabilistic model and advanced probabilistic model. He had tested these models by changing them with each other and their parameters. He had prepared data sets with spam and normal messages to filter the messages and then formed a keywords vector to separate the messages. He had used the roots of the words in the messages to form

the keywords vector (Gungor, 2004). Zheng and Webb had proposed the application of lazy learning techniques to Bayesian tree induction and presented the resulting lazy Bayesian rule learning algorithm, called LBR. For each test example, LBR had built a most appropriate rule with a local naive Bayesian classifier as its consequent. It had demonstrated that the computational requirements of LBR are reasonable in a wide cross-section of natural domains. Experiments with these domains had shown that, on average, that new algorithm had obtained lower error rates significantly more often than the reverse in comparison to a naive Bayesian classifier, a Bayesian tree learning algorithm, a constructive Bayesian classifier that eliminates attributes and constructs new attributes using Cartesian products of existing nominal attributes, and a lazy decision tree learning algorithm (Zheng and Webb, 2000).

Green and Karp had used simple Bayes classifier to identify missing enzymes in predicted metabolic pathway databases. They had developed a method that efficiently combines homology and pathway-based evidence to identify candidates for filling pathway holes in Pathway/Genome databases. Their program had not only identified potential candidate sequences for pathway holes, but combined data from multiple, heterogeneous sources to assess the likelihood that a candidate has the required function. Their algorithm had emulated the manual sequence annotation process, considering evidence from genomic context and functional context to determine the posterior belief that a candidate has the required function. The program had used a set of sequences encoding the required activity in other genomes to identify candidate proteins in the genome of interest, and then evaluated each candidate by using a simple Bayes classifier to determine the probability that the candidate has the desired function (Green and Karp, 2004). Vannucci et al had presented a wavelet-based method for classification based on functional data that uses probit models with latent variables and Bayesian mixture priors for variable selection. They had applied the method to the classification of three wheat varieties based on 100 near infra-red absorbencies and to ovarian cancer discrimination based on mass-spectra. In the applications they had employed wavelet transforms as a tool for dimension reduction and noise removal, reducing spectra to wavelet components.

Their method had been able to identify small sets of coefficients that capture the discriminatory information of the spectral data (Vannucci et al, 2005).

Chen et al had presented two feature evaluation metrics for the Naive Bayesian classifier applied on multi-class text data sets: Multi-class Odds Ratio (MOR), and Class Discriminating Measure (CDM). They had compared CDM and MOR with three variations of Odds Ratio for multi-class datasets like EOR, WOR and MC-OR. They had also compared them with information gain (IG), which is usually among the best performing metrics for many text datasets. Experimental results on two data sets had shown that CDM and MOR are among the best performing metrics for the Naïve Bayes classifier applied on multi-class text datasets and the computation of CDM metric was simpler than other feature evaluation metrics (Chen et al, 2008). Corso et al had presented a new method for automatic segmentation of heterogeneous image data that takes a step toward bridging the gap between bottom-up affinity-based segmentation methods and top-down generative model based approaches. The main contribution of their paper was a Bayesian formulation for incorporating soft model assignments into the calculation of affinities, which are conventionally model free. They had integrated the resulting model-aware affinities into the multilevel segmentation by weighted aggregation algorithm, and applied the technique to the task of detecting and segmenting brain tumor and edema in multichannel MR volumes. The method had run orders of magnitude faster than current state-of-the-art techniques giving comparable or improved results. Their results had indicated the benefit of incorporating model-aware affinities into the segmentation process for the difficult case of brain tumor (Corso et al, 2008).

Elliot et al had compared the accuracy of a Bayesian approach to combining surname and geocoded information to estimate race/ethnicity to two other indirect methods: a non-Bayesian method that combines surname and geocoded information and geocoded information alone. They had assessed accuracy with respect to estimating individual race/ethnicity and overall racial/ethnic prevalence in a population. They had found that the Bayesian Surname and Geocoding (BSG) method was more efficient than geocoding alone. The Bayesian Surname and

Geocoding (BSG) method presented here had efficiently integrated administrative data, substantially improving upon what is possible with a single source or from other hybrid methods (Elliot et al, 2008). Yu et al had proposed a Bayesian approach to determining the separating hyper plane of a support vector machine (SVM). In the proposed model of b-SVM, all the parameters are estimated by the reversible jump Markov chain Monte Carlo (RJMCMC) strategies, and the location parameter of decision boundary is finally described by a posterior distribution. The method minimizes the Bayes error in some derived direction. Tested by many independent random experiments of 2-fold cross validations, the experimental results on some high-throughput biodata sets had demonstrated the promising performance and robustness of their novel's classification method (Yu et al, 2008).

2.3. Previous Studies on Genetic Algorithm

In addition to KNN and Bayesian methods, GAs are developed and utilized in optimization problems. Aminzadeh had used bioinspired algorithms that inspired from biology for high level synthesis at his study. He mentioned that there were number of heuristic algorithms for digital circuit synthesis, which can solve scheduling and binding problems, but these algorithms were time consuming for large designs and they cannot consider several constraints simultaneously. He had developed three GAs for scheduling, module binding and register allocation problems, and then a co-evolutionary strategy had merged the result of these three solutions, targeting improvement of design parameters (Aminzadeh, 2006). Wang and Shi had proposed a hybrid GA combined with split and merge techniques (SMGA) for two types of polygonal approximation of digital curve. The algorithm's main idea was applying two classical methods, split and merge techniques, to repair infeasible solutions. They had mentioned that an infeasible solution can not only be repaired rapidly, but also be pushed to a local optimal location in the solution space. Their experimental results had demonstrated that SMGA was robust and outperforms other existing GA-based methods (Wang and Shi, 2006). Flom and Robinson had benefited from GA to find out the weighted sums for calculation function that is used

in Tetris game. They had tried several things to improve the efficiency of the search for the weights, including different fitness evaluations and crossover operations. Then they had discovered that GAs seem to work very well for searching for good weights for the evaluation function used to allow an agent to play Tetris. They had concluded that there was a non-obvious relationship between the weights of the features because there were strategic tradeoffs between making different plays (Flom and Robinson).

Mantare had introduced a min-max GA that can naturally be applied to the min-max problems. A min-max GA had originally designed for simultaneous minimization and maximization of the same object function during the same optimization run. The method had been applied with multiple sorting for optimizing constrained functions. The other end of the GA population had minimized the constraint violations and the other end had maximized the values of feasible solutions (Mantere, 2006). Liu et al had presented a parallel GA based coarse-grained module for the optimal design of the flexible multibody model vehicle suspensions and had constituted the skeleton implementing. They had tested the algorithm on the cluster system. Their results had shown that the application of the algorithm presented outperformed equivalent sequential GAs for the optimization and also improved the efficiency of the computing time. They had also compared the coarse-grained GA with the master-slave GA, and had found the result of the GA based coarse-grained was better than the result of the parallel GA based master-slave module (Liu et al, 2008). Zhang and Tong had presented a hybrid GA (LSHGA) for symmetric traveling salesman problem. They had contrived a modified local search method based on new neighborhood model as crossover operation and introduced a MUT3 operator as mutation operation. They had used the strategy that unites stochastic tournament and elite reservation. An idea of reservation ratio was put forward, and the theory of self-adaptive was employed for conforming parameters of LSHGA at the same time (Zhang and Tong, 2008).

Moreira had presented a solution method to the problem of automatic construction timetables for the exams with GA. He had used a model matrix, which is justified by the benefits that this model presents when used in the algorithm

solution. The method of solution was a meta-heuristics that includes a GA. The model was directed to the construction of exam timetables in institutions of higher education. The results achieved in real and complex scenarios were satisfactory; the exam timetabling met the imposed regulations (Moreira, 2008). Goncalves et al had presented a GA for the resource constrained multi-project scheduling problem (RCMPSP). The chromosome representation of the problem was based on random keys. The schedules was constructed using a heuristic that builds parameterized active schedules based on priorities, delay times, and release dates defined by the GA (Goncalves et al, 2008). Zhang et al had presented an approach based on GA for correlation clustering problem, named as GeneticCC. The clustering performance of clustering division constructed by GeneticCC was high and diversity was the important characteristic of clustering divisions of a data set constructed by GeneticCC. They had defined data correlation based clustering precision to estimate the performance of a clustering division and discussed features of clustering precision. Experimental results had shown that the performance of clustering division for UCI document data set constructed by GeneticCC was better than clustering performance of other clustering divisions constructed by SOM neural network with clustering precision as criterion (Zhang et al, 2008).

Hu et al had developed a method of image enhancement based on rough-set and GA. According to the class attribute of rough-set, they had divided the image into the marginal zone and non-marginal zone, and then enhanced them separately. In the process of rough-set classifying, optimizing the threshold value was achieved by applying GA, which can assure the optimization of the classifying. They had combined rough-set classification and GA together to enhance the image, since both of them had its own advantage on image enhancement (Hu et al, 2008). Lee et al had developed an algorithm of GA with ant colony optimization (GA-ACO) for multiple sequence alignment. The proposed GA-ACO algorithm was to enhance the performance of genetic algorithm (GA) by incorporating local search, ant colony optimization (ACO), for multiple sequence alignment. In the proposed GA-ACO algorithm, GA was conducted to provide the diversity of alignments. Thereafter, ant colony optimization was performed to move out of local optima (Lee et al, 2008).

Bide at al had developed a method that integrates fuzzy c with IGA is put forward for fault diagnosis of power transformer. The method had converted the problem about minimum for fuzzy c to the problem about maximum for IGA. They had concluded that the new method could diagnose the power transformer's faults effectively (Bide et al, 2008). Wang at al had designed a mutation with cycle probability simulating the evolutionary rule of the earth creature, and a GA based on the cycle mutation had presented the ability in improving search efficiency and overcoming premature to some extent. The selection was mended according to the phenomena that optimum individual always plays a major role, and an improved cycle mutation GA was proposed (Wang et al, 2008).

3. MATERIALS AND METHODS

3.1. K Nearest Neighbor Method

KNN method is one of the oldest and simplest methods for general, non-parametric classification and based on supervised learning (Bay, 1999). KNN is a simple and easy-to-implement method that performs competitive results even compared to the most sophisticated machine learning methods (Song et al, 2007). Despite its simplicity, it can learn from a small set of examples, incrementally add new information at runtime and give competitive performance with more modern methods such as decision trees or ANNs (Bay, 1999). The aim is to find nearest k sample from the existing training data when a new sample appears and classify the appeared sample according to most similar class (Mitchell, 1997). Recent studies have continued to develop and advance the KNN approach, with efforts focused on the investigation of techniques to reduce the time required for finding the nearest neighbors and applying non-uniform weights to the k neighbors in the calculation of the class posterior probability estimates (Remus et al, 2008).

The concept of using a distance metric to weight the k neighbors was first introduced and implemented by Dudani in the distance-weighted KNN method (Dudani, 1976). Dudani suggested that the closest neighboring points provide a better prototype for classification than more distant points and hence should carry greater weight in the class posterior probability estimate, an argument that is consistent with the aim of KNN that is modeling the underlying probability distributions in the region of the test point using only the local neighborhood. The original work by Dudani has motivated the development of various weighted KNN methods that use weights based on the distances between the k neighbors and the unlabeled test point to calculate the class probability estimates. There is not yet a consensus approach to determining the best weights on the k neighbors since many of these methods utilize arbitrary scaling functions or scaling parameters to calculate the weights based on the distances (Remus et al, 2008).

The classifier predicts the class label of a query vector x_0 on the predefined P classes from a set of N labeled instances $\{x_i, y_i\}_1^N$ (Song et al, 2007). Generally closeness is defined with Euclidean distance. Mitchell (1997) had explained Euclidean distance precisely with a formula. An arbitrary instance x be described by the feature vector $\langle a_1(x), a_2(x), \ldots, a_n(x) \rangle$ where $a_r(x)$ denotes the value of rth attribute of instance x. Then the distance between two instances x_i and x_j is defined to be $d(x_i,x_j)$ as follows.

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (1)$$

Afterwards, unknown sample is appointed to most similar class from KNN. Also KNN method is used to guess a real value for an unknown sample (Yildiz et al, 2008).

Primarily choosing appropriate k value and distance measurement determines the performance of a KNN classifier (Song et al, 2007). When the points are not uniformly distributed, determining the k value becomes difficult. Generally larger k values are chosen in the event of noised data sets to make the boundaries smooth between the classes (Song et al, 2007). A good k can be selected by various heuristic techniques like cross-validation. The special case where the class is predicted to be the class of the closest training sample (of course k is equal to 1) is called the nearest neighbor algorithm. It is impossible to choose same k value for all different applications (Song et al, 2007).

The accuracy of the KNN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes.

KNN has some interesting properties. It requires the tuning of only one free parameter. It is nonparametric and does not assume any particular statistical distribution of the training data. It converges to the optimal Bayes bound under certain conditions. Its main drawback is the majority voting decision strategy generally adopted to perform the classification task. This strategy overlooks the geometric configuration of the k nearest training samples in the decision process, thereby resulting in an under exploitation of their discrimination potential (Blanzieri and Melgani, 2008).

Researchers have attempted to propose new approaches to increase the performance of KNN method by using prior knowledge such as the distribution of the data and feature selection. Discriminant Adaptive NN (DANN), Adaptive Metric NN (ADAMENN), Weight Adjusted KNN (WAKNN), Large Margin NN (LMNN) are some of these approaches (Song et al, 2007).

In general the following steps are performed for KNN algorithm (Yildiz et al, 2008):

1. Chosen of k value: K value is completely up to user. Generally after some trials a k value which gives the best result is chosen.
2. Distance calculation: Any distance measurement can be used for this step. Generally most known distance measurements like Euclidean and Manhattan distances are used.
3. Distance sort in ascending order: Chosen k value is also important at that point. Found distances are sorted in ascending order and minimum k distance is taken.
4. Finding k class values: Existing classes of k nearest data are identified.
5. Finding dominant class: In last step identified k classes are formed a ratio and the class which has maximum ratio is taken. This taken class gives us the class of desired data.

3.2. Bayesian Method

Bayes theorem is effective and simple method and for this reason it is used frequently on classifying problems (Gungor, 2004; Kim et al, 2002). In machine

learning determining the best hypothesis from some space *H*, given the observed training data *D* is often interested in. Bayes theorem provides a way to calculate the posterior probability *P(h|D)*, from the prior probability *P(h)*, together with *P(D)* and *P(D|h)* (Mitchell, 1997).

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{2}$$

P(D) and *P(D|h)* denotes the prior and the posterior probability of observed training data *D* respectively.

Generally there are lots of hypothesizes whose aim is to find the maximum probability hypotheses. It is called as maximum a posteriori (MAP) hypothesis (Mitchell, 1997).

$$\begin{aligned} h_{MAP} &\equiv \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)P(h) \end{aligned} \tag{3}$$

P(D|h) is called as likelihood of data *D* for given *h*. The hypothesis that makes *P(D|h)* maximum is named as maximum likelihood (ML) hypothesis.

$$h_{ML} \equiv \arg\max_{h \in H} P(D \mid h) \tag{4}$$

In ANNs maximum likelihood hypothesis is also used for predicting probabilities (Mitchell, 1997).

$$P(D \mid h) = \prod_{i=1}^{m} P(x_i, d_i \mid h) \tag{5}$$

X_i and d_i values are random variables and trainings are done independent of each others. D_i values take 0 and 1 values for *f(x)*.

$$P(D \mid h) = \prod_{i=1}^{m} h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i) \tag{6}$$

After applying the maximum likelihood hypothesis and organizing the equation, it becomes the form as follows.

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i)$$

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i)) \tag{7}$$

Another method is minimum description length (MDL) principle. This method requires the hypothesis that minimize the description length and the description lengths of the data those give this hypothesis. The simple results of Bayes and information theorems are used to form logic for this MDL principle (Mitchell, 1997).

$$h_{MDL} = \arg\min_{h \in H} L_{c_1}(h) + L_{c_2}(D \mid h)$$

$$L_{c_1}(h) = L_{c_H}(h) = -\log_2 P(h)$$

$$L_{c_2}(D \mid h) = L_{c_{D|h}}(D \mid h) = -\log_2 P(D \mid h) \tag{8}$$

Bayes optimal classifier combines all alternative hypothesizes estimates with weighted posterior probabilities to guess the possible class. This process expressed as follows (Mitchell, 1997).

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) \tag{9}$$

When attributes values are independent from target values, Naive Bayes classifier is used (Mitchell, 1997).

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \tag{10}$$

In this study classifying process is optimized by using EM algorithm. EM algorithm is a method that is used to guess units which has missing data and includes maximum similarity probabilities (Friedman, 1998). EM method is a repeated method and it has two stages. Expectation stage gives expectation for the data. Maximization stage gives expectation about mean, standard deviation or correlation when a missing data is appointed. This process continues until the change on expected values decreases to a negligible value (Ozcanli et al, 2003).

Bayesian methods are based on probability calculus. EM algorithm is one of them. It is utilized for unknown values with known probability distributions. Radial basis function is a popular function for explaining probability distributions (Mitchell, 1997; Neal, 2004). The expression of this function is given as:

$$y = e^{-\left(\frac{x-\mu}{\sigma}\right)} \tag{11}$$

In Equation (11), x denotes training data, μ denotes mean of data and σ refers variance.

One way to form an EM algorithm is to guess the average value of Gauss functions. Let us have a sample data set which has k different classes. It means that data set is formed from a probability distribution which is a mixture of k different normal distribution. Each sample is formed with two steps process. At first step a random normal distribution is chosen from k normal distribution as seen in Figure 3.1. At second step a sample data is formed according to this distribution. These steps are repeated for each point in the set.

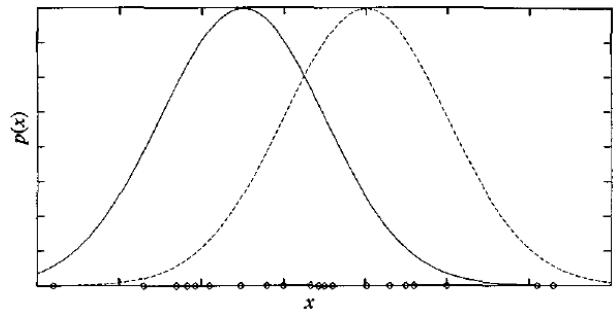

Figure 3.1. The Normal distributions when k=2 and sample data are on the x-axis (Mitchell, 1997)

3.3. Genetic Algorithm

GAs are search and maximization methods that work similar to the evolutionary continuum at nature. It searches the best solution at multi dimensional search space according to "best lives" principle. GAs produce a solution set that includes different solutions instead of producing only one solution. Hereby; many points at search space are evaluated at the same time and probability of reaching a total solution is increasing. The solutions in solution set are completely independent from each other. Each one is a vector on the multi dimensional space vector. GAs simulates evolutionary continuum on computer environment to solve problems. They do not develop only one structure for solution like other maximization methods, but develop a set that formed with these structures (Beasley et al, 1993a; Beasley et al, 1993b).

GAs operate on a population of potential solutions applying the principle of survival of the fittest to produce better and better approximations to a solution. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation (Chipperfield et al).

Individuals, or current approximations, are encoded as strings, chromosomes, composed over some alphabet(s), so that the genotypes (chromosome values) are

uniquely mapped onto the decision variable (phenotypic) domain. The most commonly used representation in GAs is the binary alphabet {0, 1} although other representations can be used, e.g. ternary, integer, real-valued etc.

Having decoded the chromosome representation into the decision variable domain, it is possible to assess the performance, or fitness, of individual members of a population. This is done through an objective function that characterizes an individual's performance in the problem domain. In the natural world, this would be an individual's ability to survive in its present environment. Thus, the objective function establishes the basis for selection of pairs of individuals that will be mated together during reproduction. During the reproduction phase, each individual is assigned a fitness value derived from its raw performance measure given by the objective function. This value is used in the selection to bias towards more fit individuals. Highly fit individuals, relative to the whole population, have a high probability of being selected for mating whereas less fit individuals have a correspondingly low probability of being selected (Chipperfield et al).

Once the individuals have been assigned a fitness value, they can be chosen from the population, with a probability according to their relative fitness, and recombined to produce the next generation. Genetic operators manipulate the characters (genes) of the chromosomes directly, using the assumption that certain individual's gene codes, on average, produce fitter individuals. The recombination operator is used to exchange genetic information between pairs, or larger groups, of individuals. The simplest recombination operator is that of single-point crossover. This crossover operation is not necessarily performed on all strings in the population. Instead, it is applied with a probability *Px* when the pairs are chosen for breeding. A further genetic operator, called mutation, is then applied to the new chromosomes, again with a set probability *Pm*. Mutation causes the individual genetic representation to be changed according to some probabilistic rule. In the binary string representation, mutation will cause a single bit to change its state, 0 to 1 or 1 to 0 (Chipperfield et al).

Mutation is generally considered to be a background operator that ensures that the probability of searching a particular subspace of the problem space is never

zero. This has the effect of tending to inhibit the possibility of converging to a local optimum, rather than the global optimum.

After recombination and mutation, the individual strings are then, if necessary, decoded, the objective function evaluated, a fitness value assigned to each individual and individuals selected for mating according to their fitness, and so the process continues through subsequent generations. In this way, the average performance of individuals in a population is expected to increase, as good individuals are preserved and bred with one another and the less fit individuals die out. The GA is terminated when some criteria are satisfied, like a certain number of generations, a mean deviation in the population, or when a particular point in the search space is encountered.

GA differs substantially from more traditional search and optimization methods. The four most significant differences are:

• GAs search a population of points in parallel, not a single point.

• GAs do not require derivative information or other auxiliary knowledge; only the objective function and corresponding fitness levels influence the directions of search.

• GAs use probabilistic transition rules, not deterministic ones.

• GAs work on an encoding of the parameter set rather than the parameter set itself (except in where real-valued individuals are used) (Chipperfield et al).

GAs have two important features that underlie their success. The first is their employment of an algorithmic equivalent of natural selection. When chromosomes are chosen as parents during the reproduction process, the probability that a given chromosome will be chosen is biased in accord with its fitness. Thus, the fittest chromosomes those that solve the problem best will tend to have more children than the less fit ones. The use of fitness-based reproduction generally leads to an improvement in the population as a GA runs. The second feature is the use of mutation and crossover operators during reproduction. Mutation operators cause children to differ from their parents through the introduction of localized change. Crossover operators create children that combine chromosomal matter from two parents. The production of high-performance chromosomes can be greatly speeded

up with crossover working to combine subparts of good solutions from multiple parents on a single child (Kelly and Davis, 1991).

The process of a GA usually begins with a randomly selected population of chromosomes. These chromosomes are representations of the solutions of the problem. According to the attributes of the problem, different positions of each chromosome are encoded as bits, characters, or numbers. These positions are sometimes referred to as genes and are changed randomly within a range during evolution. The set of chromosomes during a stage of evolution are called a population. An evaluation function is used to calculate the "goodness" of each chromosome. During evaluation, two basic operators, crossover and mutation, are used to simulate the natural reproduction and mutation of species. The selection of chromosomes for survival and combination is biased towards the fittest chromosomes (Li and Wei, 2004).

Figure 3.2 shows the structure of a simple GA. It starts with a randomly generated population, evolves through selection, recombination (crossover), and mutation. Finally, the best individual (chromosome) is picked out as the final result once the optimization criterion is met (Pohlheim and Hartmut, 2003).

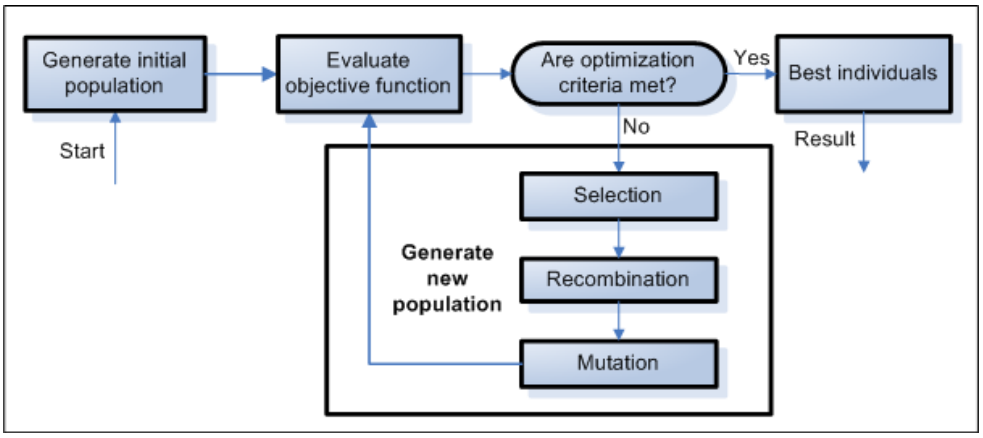

Figure 3.2. Structure of a simple GA (Pohlheim and Hartmut, 2003)

3.4. Artificial Neural Networks

ANNs are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is

determined largely by the connections between elements. We can train a ANN to perform a particular function by adjusting the values of the connections (weights) between elements (Demuth and Beale, 2002).

Commonly ANNs are adjusted, or trained, so that a particular input leads to a specific target output. Such a situation is shown in Figure 3.3. There, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target. Typically many such input/target pairs are used, in this supervised learning, to train a network (Demuth and Beale, 2002).

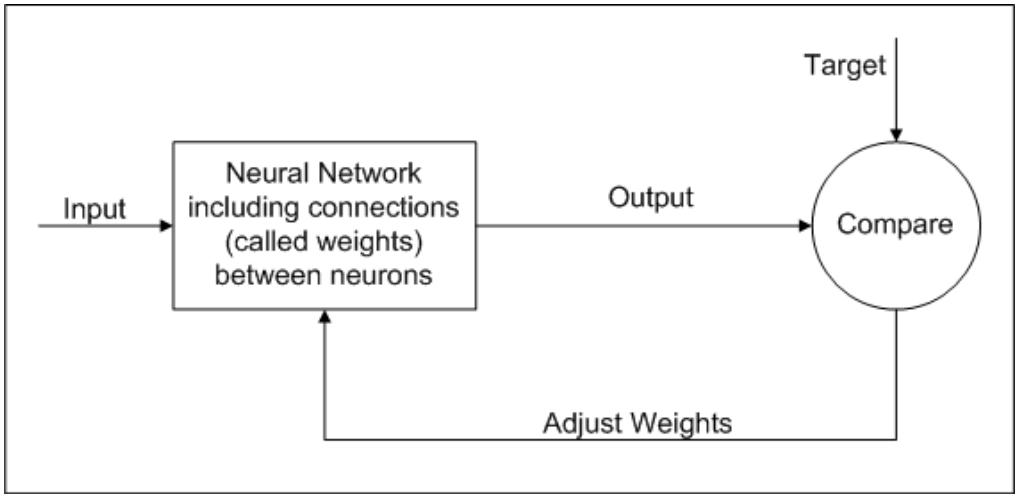

Figure 3.3. Structure of a simple ANN (Demuth and Beale, 2002)

ANNs have been trained to perform complex functions in various fields of application including pattern recognition, identification, classification, speech, control systems and vision. Today ANNs can be trained to solve problems that are difficult for conventional computers or human beings (Demuth and Beale, 2002).

The fundamental building block in an ANN is the mathematical model of a neuron as shown in Figure 3.4. The three basic components of the (artificial) neuron are:

1. The synapses or connecting links that provide weights, w_j, to the input values, x_j for $j = 1, ...m$;

2. An adder that sums the weighted input values to compute the input to the activation function v.

$$v = w_0 + \sum_{j=1}^{m} w_j x_j \tag{12}$$

w_0 is called the bias is a numerical value associated with the neuron. It is convenient to think of the bias as the weight for an input x_0 whose value is always equal to one, so that

$$v = \sum_{j=0}^{m} w_j x_j \tag{13}$$

3. An activation function g (also called a squashing function) that maps v to $g(v)$ the output value of the neuron. This function is a monotone function (Patel, 2003).

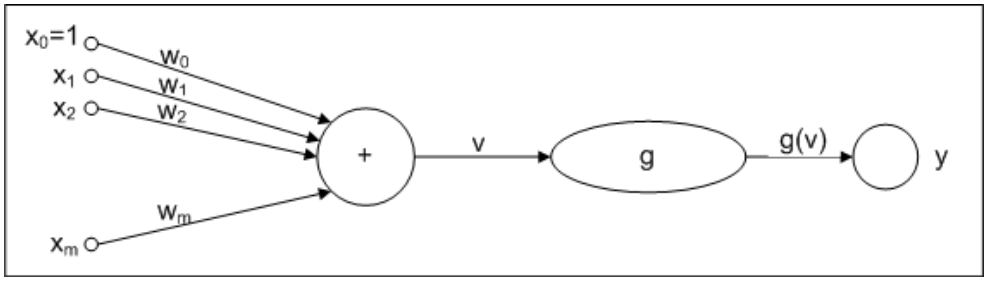

Figure 3.4. Mathematical model of a neuron (Patel, 2003)

While there are numerous different ANN architectures that have been studied by researchers, the most successful applications in data mining of ANNs have been multilayer feed forward networks. These are networks in which there is an input layer consisting of nodes that simply accept the input values and successive layers of nodes that are neurons as depicted in Figure 3.4. The outputs of neurons in a layer are inputs to neurons in the next layer. The last layer is called the output layer. Layers between the input and output layers are known as hidden layers. Figure 3.5 is a diagram for this architecture (Patel, 2003).

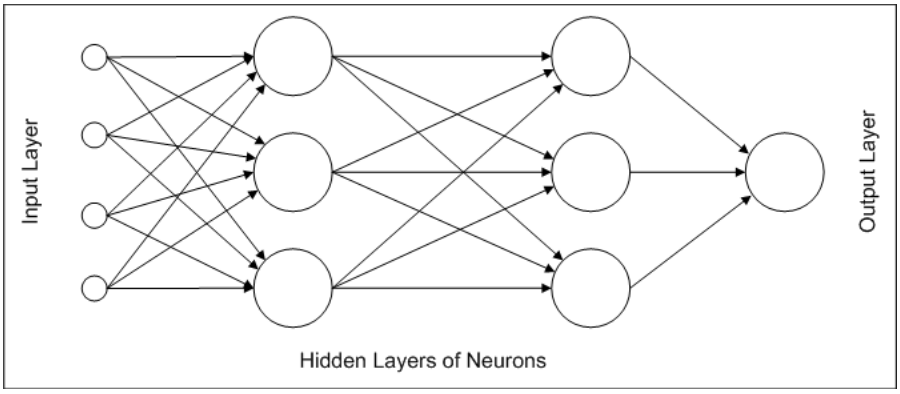

Figure 3.5. Architecture of an ANN (Patel, 2003)

In a supervised setting where a neural net is used to predict a numerical quantity there is one neuron in the output layer and its output is the prediction. When the network is used for classification, the output layer typically has as many nodes as the number of classes and the output layer node with the largest output value gives the network's estimate of the class for a given input. In the special case of two classes it is common to have just one node in the output layer, the classification between the two classes being made by applying a cut-off to the output value at the node (Patel, 2003).

There are two types of networks for ANNs. Single layer networks are the simplest networks consists of just one neuron with the function g chosen to be the identity function, $g(v) = v$ for all v. In this case the output of the network is $\sum_{j=0}^{m} w_j x_j$, a linear function of the input vector x with components x_j .

Multilayer ANNs are undoubtedly the most popular networks used in applications. While it is possible to consider many activation functions, in practice it has been found that the logistic (also called the sigmoid) function $g(v) = (\frac{e^v}{1+e^v})$ as the activation function (or minor variants such as the tanh function) works best. In fact the revival of interest in neural nets was sparked by successes in training ANNs using this function in place of the historically (biologically inspired) step function (perceptron) (Patel, 2003).

There is a minor adjustment for prediction problems where we are trying to predict a continuous numerical value. In that situation we change the activation function for output layer neurons to the identity function that has output value=input value. The backward propagation algorithm cycles through two distinct passes, a forward pass followed by a backward pass through the layers of the network. The algorithm alternates between these passes several times as it scans the training data. Typically, the training data has to be scanned several times before the networks learn to make good classifications. Computation of outputs of all the neurons in the network is done at forward pass stage. At backward pass propagation of error and adjustment of weights is done (Patel, 2003).

The backward propagation algorithm is a version of the steepest descent optimization method applied to the problem of finding the weights that minimize the error function of the network output. Due to the complexity of the function and the large numbers of weights that are being trained as the network learns, there is no assurance that the backward propagation algorithm will find the optimum weights that minimize error. The procedure can get stuck at a local minimum. It has been found useful to randomize the order of presentation of the cases in a training set between different scans (Patel, 2003).

A single scan of all cases in the training data is called an epoch. Most applications of feed forward networks and backward propagation require several epochs before errors are reasonably small. A number of modifications have been proposed to reduce the epochs needed to train a neural net (Patel, 2003).

4. THE PROPOSED ALGORITHMS

4.1. First Study: A Hybrid Classification Method Using Bayesian, K Nearest Neighbor Methods and Genetic Algorithm

In first study's proposed hybrid method KNN, Bayesian methods and GA are used together. Main idea is to improve the existing data sets with the method. Steps of this hybrid method can be summarized as given below:

Step 1: Firstly, Bayesian method based EM algorithm is applied on chosen data set. All data from chosen data set are used as test data. Then number of wrong classified data is obtained. Assume this number as x.

Step 2: In this step a new data set is generated randomly between maximum and minimum values of each class of the original data set. The number of elements of new data set is two times more according to original one.

Step 3: KNN method is applied in this step. Distances to the mean of each class are calculated for each data. The data are sorted according to distances and closer half of them are chosen. So; k value is chosen as fifty percentage of generated number of data.

Step 4: First step is repeated in this step. Test data are again the original data set however train data are the generated data set. EM algorithm is applied and number of wrong classified data is obtained. This number is named as y.

Step 5: In this step x and y values are compared and if x is less than or equal to y, the new generated data set will not be better than the original one. So; step 2, 3 and 4 should be repeated again. This loop will continue till y becomes less than x. When this happens, it means generated data set is better for training than the original one.

Step 6: From this step a new loop starts and the number of loop is up to user according to characteristics of the data set values. GA is applied on last generated data set. Data are sorted again according to their distances and defined crossover ratio. The worst data are put into crossover process with each other and a new data set is formed.

Step 7: EM algorithm is applied again on the new data set and number of wrong classified data named as z is obtained. Z is compared with y and if z is less than y, the new data set will be better than the old one. The loop continues with the new data set.

Step 8: This is the last step and after the completion of the loop, found best data set and minimum number of wrong classified data is saved.

Flow chart of the proposed hybrid method is given in Figure 4.1.

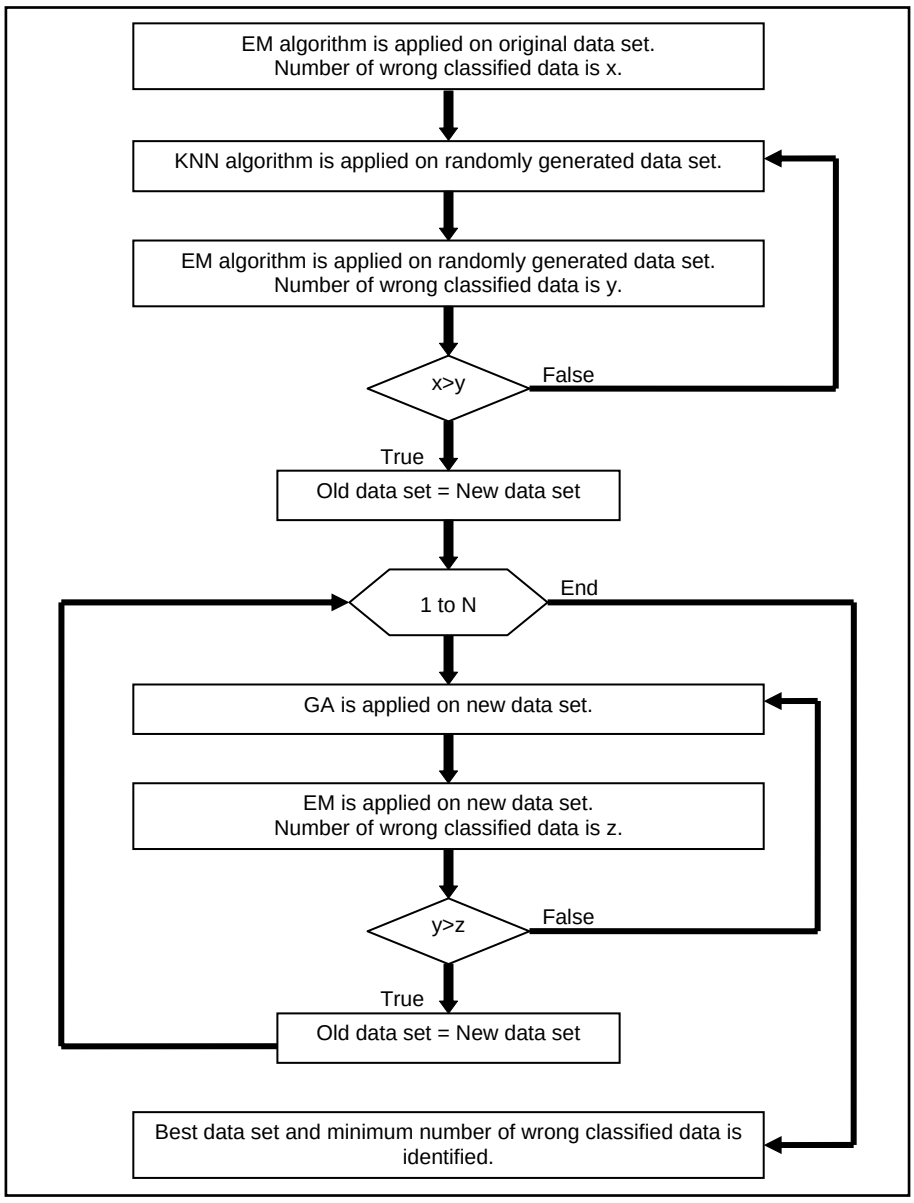

Figure 4.1. Flow chart of the suggested method

The original and generated data sets, also the best one, are used in the algorithm and numbers of wrong classified data are brought out according to the above steps. To determine the reliability of the study the original and best data sets are tested on ANNs. All of the data sets are divided into 5 folds and the results are taken from multi layer ANNs.

4.2. Second Study: Utilization of KNN Method for Expectation Maximization Based Classification Method

KNN and Bayesian methods are used together at this second hybrid method. KNN method considered as the preprocessor for EM algorithm of Bayesian method. Main idea is to reduce the number of data with KNN method and guess the class using most similar training data with EM algorithm. Steps of this algorithm can be summarized as given below:

Step 1: All data of the train and the test sets are read.

Step 2: All data of the data set are grouped according to their classes. Distances between each train data and the test data are calculated.

Step 3: Distances are sorted in ascending order for each class.

Step 4: K data that have minimum distance to the test data are chosen from each class to form a new data set. After that point data set is updated with the new one.

Step 5: At this step, KNN algorithm is finished, and hybrid method continues with EM algorithm. The aim is to form gauss distributions for the classes. Maximum and minimum data are determined for each class.

Step 6: Step values are calculated to fit the results in the same scale between maximum and minimum values for each class. Then radial basis functions are calculated for each class.

Step 7: Gauss distributions are formed with these k data of each class.

Step 8: Test data is also located on each gauss distribution and then probabilities are found on each gauss distribution.

Step 9: Step 7 and step 8 are applied for each attribute and all found probabilities are multiplied for each class. Therefore there is only one probability value for each class.

Step 10: At last step the maximum probability is chosen among all class probabilities. The maximum class probability gives the class of test data. For the implementation of the method all data are used as both train and test data.

Flowchart of the suggested hybrid method is given in Figure 4.2.

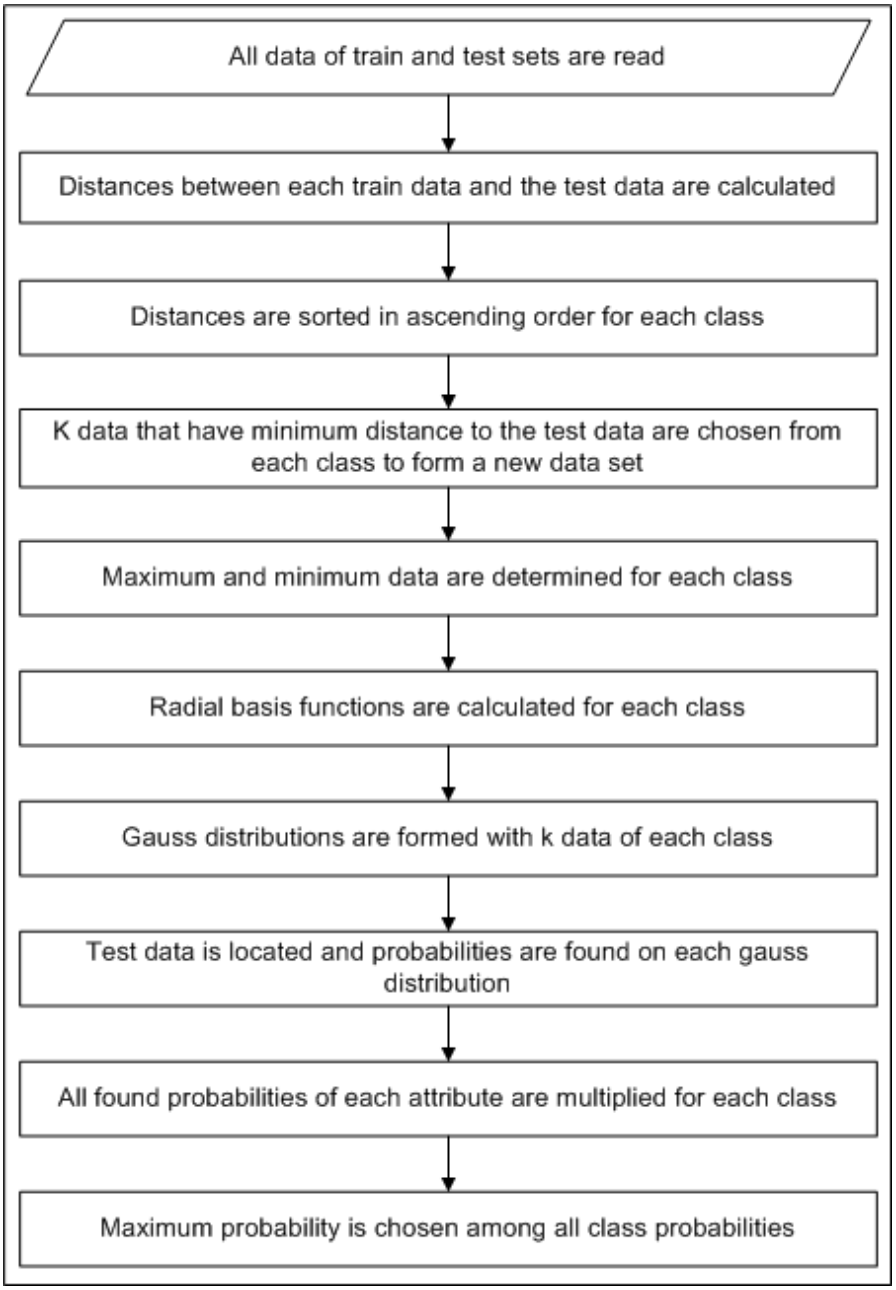

Figure 4.2. Flowchart of the suggested hybrid method

Different distance measurements can be used for distance calculation mentioned at Step 2. Minkowski distance is utilized in this study.

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{x} \left| x_{ik} - x_{jk} \right|^{\lambda}} \tag{14}$$

When the value of λ (lambda) at Minkowski distance is chose as one, Manhattan distance is obtained. Generally λ is chose as two that means Euclidean distance. If λ is equal to three or more, the general name Minkowski is used for that distance. In this work λ value is chose as one, two and three respectively.

There is only one limitation for the k value. K value does not exceed the data number of the class that contains minimum data. At this study k values are swept between 1 and 40, and the k value that gives the best clustering result is determined.

5. TEST RESULTS AND PERFORMANCE EVALUATION

Suggested method is tested on well-known UCI machine learning data sets Iris, Breast Cancer, Glass, Yeast and Wine (Yildiz et al, 2008). The main properties of data sets are given in Table 5.1.

Table 5.1. Data sets and their properties

Properties	Iris	Breast Cancer	Glass	Yeast	Wine
Class	3	2	6	10	3
Sample	150	699	214	1484	178
Distribution	50-50-50	458-241	70-17-76-13-9-29	463-429-244-163-51-44-37-30-20-5	59-71-48

5.1. Test and Performance Results of First Study: A Hybrid Classification Method Using Bayesian, K Nearest Neighbor Methods and Genetic Algorithm

The hybrid method is applied on the above data sets and several results are achieved. The results are compared with the results of EM algorithm.

Iris data set has 150 data at three classes. With the EM algorithm 9 data are wrong classified and it means that 6% of the data are wrong classified. Suggested hybrid method is applied 20 times and all of the trials finished with better results. At fifth and fifteenth trials only five data are wrong classified with the hybrid method. It also shows that 3.33% of the data are wrong classified. All the results belonging to Iris data set are shown in Table 5.2, Table 5.3 and Figure 5.1.

Table 5.2. Achieved number and percentage values of wrong classified data on Iris data set with EM algorithm

Number of wrong classified data with EM algorithm	Percentage values of wrong classified data with EM algorithm (%)
9	6.00

Table 5.3. Achieved number and percentage values of wrong classified data after the trials on Iris data set with hybrid method

Number of trials	Number of wrong classified data	Percentage values of wrong classified data (%)
1	8	5.33
2	8	5.33
3	7	4.67
4	8	5.33
5	**5**	**3.33**
6	6	4.00
7	7	4.67
8	8	5.33
9	8	5.33
10	7	4.67
11	7	4.67
12	6	4.00
13	8	5.33
14	8	5.33
15	**5**	**3.33**
16	7	4.67
17	7	4.67
18	8	5.33
19	7	4.67
20	7	4.67

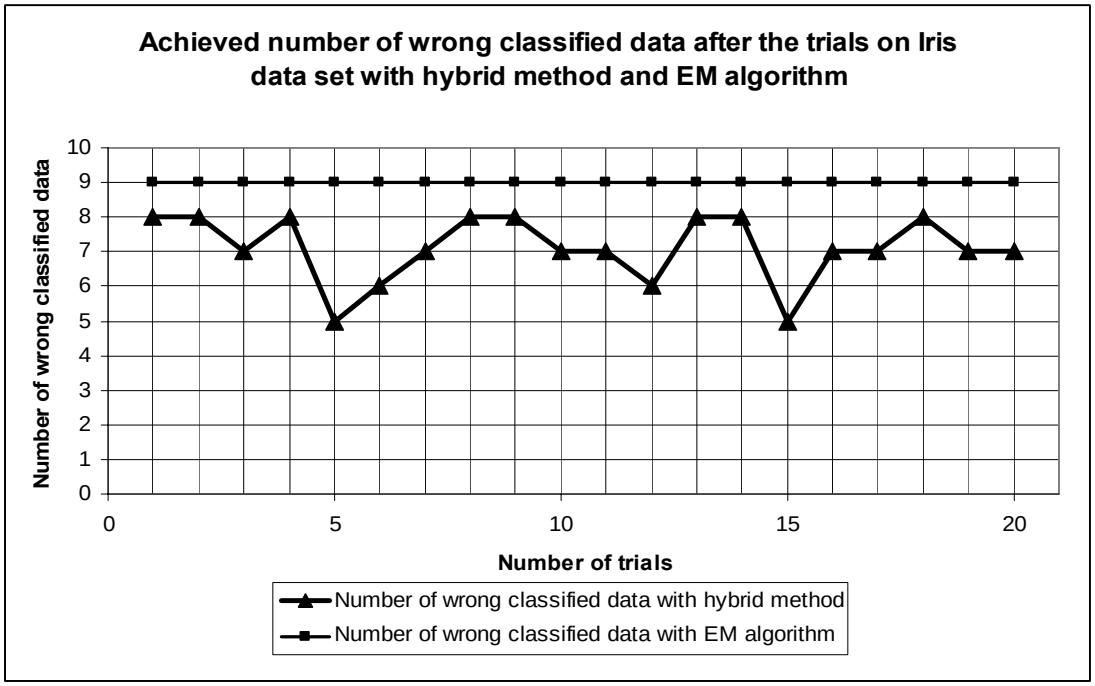

Figure 5.1. Achieved number of wrong classified data after the trials on Iris data set with hybrid method and EM algorithm

Table 5.4 shows the percentage values of the improvement on number of wrong classified data on different data sets after applying the hybrid method. 20 trials are done over Iris, Breast Cancer, Glass, Yeast and Wine data sets. The results are listed as percentage values to show the development about reducing the number of wrong classified data. Comparison and calculation of the improvement are made according to the results of EM algorithm. The best improvement is achieved on Breast Cancer data set and at fourth trial the improvement is resulted as 75.52%. Glass data set is also generally gives good results and its best improvement is achieved at second trial as 28.50%. Yeast data set gives stable results. It has only 10.15% improvement and this value is constant for all trials. Wine data set is the only data set that does not show any improvement. It also gives stable results like Yeast data set and has retrogression with -1.67% values. The whole results of the work are listed in Table 5.4 and Figure 5.2.

Table 5.4. The percentage values of the improvement on number of wrong classified data on different data sets after applying the hybrid method

Number of trials	Iris (%)	Breast Cancer (%)	Glass (%)	Yeast (%)	Wine (%)
1	11.11	67.13	15.89	10.15	-1.67
2	11.11	63.64	28.50	10.15	-1.67
3	22.22	67.13	14.95	10.15	-1.67
4	11.11	75.52	14.49	10.15	-1.67
5	44.44	53.85	15.42	10.15	-1.67
6	33.33	50.35	13.08	10.15	-1.67
7	22.22	58.04	14.95	10.15	-1.67
8	11.11	53.85	18.22	10.15	-1.67
9	11.11	67.13	24.30	10.15	-1.67
10	22.22	69.23	25.70	10.15	-1.67
11	22.22	65.03	28.04	10.15	-1.67
12	33.33	68.53	17.76	10.15	-1.67
13	11.11	61.54	17.76	10.15	-1.67
14	11.11	67.83	18.22	10.15	-1.67
15	44.44	48.25	14.95	10.15	-1.67
16	22.22	35.66	16.82	10.15	-1.67
17	22.22	33.57	25.70	10.15	-1.67
18	11.11	69.93	16.82	10.15	-1.67
19	22.22	53.85	17.29	10.15	-1.67
20	22.22	59.44	23.36	10.15	-1.67

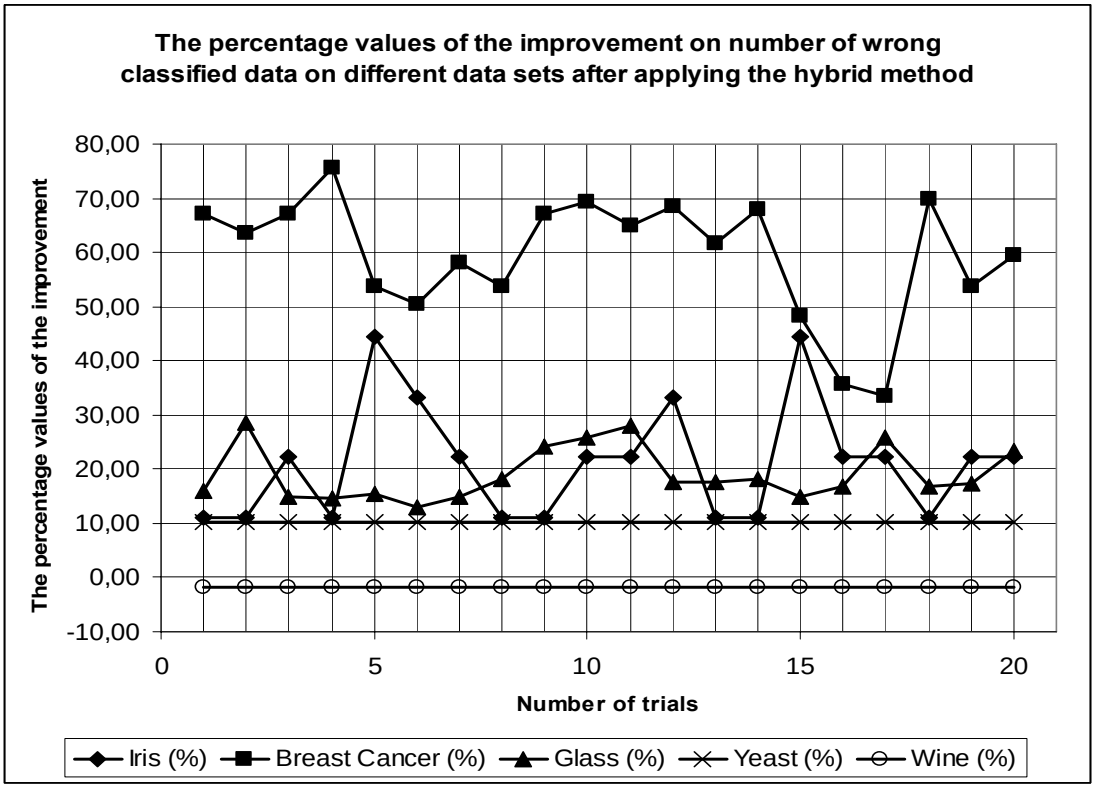

Figure 5.2. The percentage values of the improvement on number of wrong classified data on different data sets after applying the hybrid method

Table 5.5. The number of Correct (C) and Wrong (W) classified data on the test data set with Original Data Set (ODS) and New Data Set (NDS) with ANNs for 5 folds

Dataset	Type	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5		Average	
		C	W	C	W	C	W	C	W	C	W	C	W
Iris	**ODS**	15	0	14	1	13	2	14	1	15	0	**14,2**	**0,8**
	NDS	15	0	14	1	14	1	14	1	15	0	**14,4**	**0,6**
Breast Cancer	**ODS**	135	5	133	7	134	6	136	4	136	4	**134,8**	**5,2**
	NDS	137	3	135	5	135	5	136	4	137	3	**136**	**4**
Glass	**ODS**	15	27	15	27	15	28	14	30	16	27	**15**	**27,8**
	NDS	17	25	18	24	16	27	16	28	17	26	**16,8**	**26**
Yeast	**ODS**	118	181	120	178	130	168	131	165	122	171	**124,2**	**172,6**
	NDS	120	179	121	177	130	168	130	164	123	170	**124,8**	**171,6**
Wine	**ODS**	27	9	32	4	29	7	33	2	32	3	**30,6**	**5**
	NDS	27	9	30	6	28	6	31	4	31	4	**29,4**	**5,8**

Table 5.5 shows the number of correct and wrong classified data on the test data set with original data set and new data set with ANNs. The results are obtained over 5 different folds. Generally results of ANNs are parallel to the hybrid method's results. However; the improvement is not high as much as hybrid method's results.

5.2. Test and Performance Results of Second Study: Utilization of K Nearest Neighbor Method for Expectation Maximization Based Classification Method

According to the results on Table 5.6, 5.7 and 5.8; for all k and λ values hybrid method results better than Bayesian algorithm. With the hybrid method for small k values better results are found with respect to KNN algorithm. According to the λ values distance measurements do not determine stable results. For Breast Cancer dataset Manhattan distance determines better results. Generally Euclidean and Minkowski distances give similar results. Minkowski distance obtains best performance on Wine dataset. According to the general results Euclidean distance is more applicable for the measurements. It obtains acceptable results for all datasets.

For Iris data set maximum clustering errors are 7 at hybrid method for all distances. With KNN algorithm maximum clustering errors are 7 for Manhattan and Euclidean distances and 9 with Bayesian algorithm. For Wine data set there are maximum 81 wrong classified data with Bayesian algorithm, 49 with KNN algorithm when λ is equal to one and 44 with hybrid method when λ is equal to one and two. Although Bayesian algorithm did not perform a good classifying success at Breast Cancer data set, hybrid method classified only 20 data as wrong with Manhattan distance. Both hybrid method and KNN algorithm obtained good results at Glass and Yeast data sets. Hybrid method classified 45 and 406 data as wrong respectively with Euclidean distance. KNN algorithm's best results were 55 and 456 at these data sets. Maximum error numbers after classifying with different k values and best distance measurement are listed in Table 5.9.

Table 5.6. Distribution of wrong classified data after classifying with hybrid model (H) and KNN method (K) at Iris, Wine and Breast Cancer data sets

	Iris						Wine						Breast Cancer					
	H			K			H			K			H			K		
λ / k	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	3	3	3	13	12	11	15	19	19	3	3	4	10	14	17
3	3	3	3	6	6	5	24	25	24	18	25	26	8	5	5	17	16	15
4	3	5	5	5	5	4	25	23	25	25	30	33	12	11	9	18	15	15
5	2	5	5	5	5	5	36	36	37	25	35	35	16	12	12	16	19	21
6	2	4	3	4	4	5	35	33	32	27	35	36	14	15	15	18	17	20
7	4	3	3	4	4	3	29	32	32	26	42	41	16	15	16	19	17	16
8	3	3	3	5	5	3	34	34	34	29	39	39	17	16	18	21	15	18
9	3	2	2	5	3	3	36	34	33	27	38	42	17	17	18	24	20	20
10	3	3	2	4	4	3	38	34	34	34	39	41	16	17	16	22	19	21
11	2	3	2	4	4	2	38	38	38	34	39	40	18	18	15	24	21	21
12	3	3	3	5	3	2	42	37	37	33	42	42	18	18	15	23	19	20
13	2	3	3	5	3	2	41	39	39	39	46	47	18	18	16	25	21	22
14	2	3	3	4	4	3	42	36	37	35	50	50	17	18	16	23	19	20
15	3	3	3	4	2	2	40	33	32	41	48	47	17	19	19	25	22	22
16	4	3	3	4	4	3	41	36	34	42	48	48	18	22	20	25	20	22
17	4	3	4	5	3	3	38	36	35	44	48	47	19	21	21	25	23	22
18	4	4	3	5	3	3	41	39	41	43	48	48	19	22	21	25	22	22
19	4	4	3	4	3	3	41	37	38	45	49	48	19	21	22	25	22	23
20	3	3	3	4	3	3	40	37	36	43	48	48	20	21	22	25	22	21
21	3	3	3	4	3	3	42	42	38	45	48	49	19	22	23	26	25	21
22	3	3	3	3	3	3	43	39	39	45	46	48	20	22	23	26	23	21
23	4	3	4	4	3	3	44	43	43	46	49	49	20	22	23	29	25	22
24	4	3	4	3	4	3	42	42	43	44	47	48	20	22	24	29	23	21
25	4	4	4	4	3	3	42	41	42	46	48	49	20	20	24	30	24	23
26	4	4	4	3	4	3	43	41	42	44	48	49	20	20	24	29	24	22
27	5	5	4	4	4	5	42	42	40	49	49	49	20	20	24	28	24	22
28	6	6	4	6	3	3	43	42	42	47	49	49	18	20	23	27	24	23
29	6	6	6	5	4	5	45	44	43	47	49	49	18	20	24	28	25	24
30	6	6	6	5	4	4	43	44	43	47	48	48	18	19	24	27	25	24
31	6	6	6	6	6	6	42	42	41	47	50	49	19	18	24	28	25	25
32	7	7	7	5	7	6	42	42	41	45	49	49	19	18	23	28	25	25
33	7	7	7	6	6	6	41	43	42	48	51	51	19	19	23	29	25	26
34	7	7	7	5	7	7	43	43	42	45	48	48	18	19	23	28	25	26
35	7	7	7	6	6	7	43	41	41	46	49	49	18	20	22	29	25	26
36	7	7	7	5	6	6	43	40	40	46	49	49	18	20	22	29	25	26
37	7	7	7	4	6	7	44	41	41	45	45	48	18	20	22	30	25	26
38	7	7	7	5	5	6	46	40	40	43	45	48	17	20	22	30	25	26
39	7	7	7	7	7	8	45	40	40	45	45	48	17	20	22	31	26	26
40	7	7	7	6	6	7	45	44	44	45	49	49	16	20	22	31	26	26

Table 5.7. Distribution of wrong classified data after classifying with hybrid model (H) and KNN method (K) at Glass and Yeast data sets

	Glass						Yeast					
	H			K			H			K		
k \ λ	1	2	3	1	2	3	1	2	3	1	2	3
1	0	0	0	0	0	0	0	0	0	0	0	0
2	15	13	13	25	24	24	197	179	177	303	316	318
3	15	20	21	32	35	37	316	305	282	379	383	383
4	30	28	25	37	46	42	387	357	385	415	423	404
5	33	34	34	40	47	51	419	406	412	457	457	456
6	39	37	40	43	51	54						
7	41	45	44	50	55	55						
8	42	44	44	55	58	60						
9	46	43	46	54	58	62						

Table 5.8. Number of wrong classified data after classifying with Bayesian method

Iris	Wine	Breast Cancer	Glass	Yeast
9	81	458	135	1235

Table 5.9. Maximum error numbers after classifying with different k values and best distance measurement (H: Hybrid model, K: KNN method, B: Bayesian method)

Data Set	H	K	B
Iris	7	7	9
Wine	44	49	81
Breast Cancer	20	26	458
Glass	45	55	135
Yeast	406	456	1235

6. CONCLUSIONS

In first study a hybrid classification method based on combination of KNN, Bayesian methods and GA is introduced. Approach of the method contains generation of a new data set according to the given data. Utilizing this new and expanded data set, better classification of the original data set is aimed. The developed method is tested with the Iris, Breast Cancer, Glass, Yeast and Wine data sets. Generally better classification performance is obtained with respect to classic methods such as EM algorithm. On Breast Cancer data set 75.52 % improvement is achieved. Also retrogression is found on Wine data set. Generally results of ANNs are parallel to the hybrid method's results. However; the improvement is not high as much as hybrid method's results.

The method is proposed for low cost hardware based clustering solutions, noisy data set classifying applications and classifying data sets with few data. According to test results the proposed method shows more effective classifying performance than EM classifiers. The method is useful on the data sets that have few numbers of data. It generates unlimited data that have similar characteristics with original data and then improve them according to the suggested algorithm.

In second study a hybrid classification method based on combination of KNN and Bayesian methods is introduced. Utilizing this hybrid method, better classification of the data sets is aimed. The developed method is tested with the Iris, Wine, Breast Cancer, Glass and Yeast data sets. Better classification performance is obtained with respect to classic methods such as KNN and EM algorithm.

With different k values and distance measurements, different results are achieved. According to test results the proposed method shows more effective classifying performance than KNN and EM classifiers.

REFERENCES

AMASYALI, M. F., 2006. Introduction to Machine Learning. http://www.ce.yildiz.edu.tr/mygetfile.php?id=868.

AMINZADEH, S. 2006. Using Genetic Algorithm for High-Level Synthesis. Available at http://www.computer.org/portal/cms_docs_ieeecs/ieeecs/Communities/students/looking/2006Fall/10.pdf.

ANBEEK, P., VINCKEN, K. L., GRONENDAAL, F., KOEMAN, A., OSCH, M. J. P. V., and GROND, J. V. D., 2008. Probabilistic Brain Tissue Segmentation in Neonatal Magnetic Resonance Imaging. International Pediatric Research Foundation, Inc., 63(2): 158-163.

BAOLI, L., SHIWEN, Y., and QIN, L., 2003. An Improved k-Nearest Neighbor Algorithm for Text Categorization. 20th International Conference on Computer Processing of Oriental Languages.

BAY, S. D., 1999. Nearest Neighbor Classification from Multiple Feature Subsets. Intelligent data analysis, 3(3): 191.

BEASLEY, D., BULL, D.R., and MARTIN, R.R., 1993a. An Overview of Genetic Algorithms: Part 1, Fundamentals. University Computing, 15(2): 58-69.

BEASLEY, D., BULL, D.R., and MARTIN, R.R., 1993b. An Overview of Genetic Algorithms: Part 2, Research Topics. University Computing, 15(4): 170-181.

BIDE, Z., YANRUI, L., ZHAOHUI, L., and CHUNEN, F., 2008. A Method of Diagnosing Power Transformer's Fault Based on Fuzzy c and Immune Genetic Algorithm. International Conference on Intelligent Computation Technology and Automation 2008 (ICICTA), 2: 875-879.

BLANZIERI, E., and MELGANI, F., 2008. Nearest Neighbor Classification of Remote Sensing Images with the Maximal Margin Principle. IEEE Transactions on Geoscience and Remote Sensing, 46(6): 1804-1811.

CHEN, J., HUANG, H., TIAN, S., and QU, Y., 2008. Feature Selection for Text Classification with Naïve Bayes. Expert Systems with Applications, doi:10.1016/j.eswa.2008.06.054.

CHIPPERFIELD, A., FLEMING, P., POLHEIM, H., and FONSECA, C. Genetic Algorithm Toolbox User's Guide. Department of Automatic Control and Systems Engineering, University of Sheffield, Version 1.2.

CORSO, J. J., SHARON, E., DUBE, S., EL-SADEN, S., SINHA, U., and YUILLE, A., 2008. Efficient Multilevel Brain Tumor Segmentation with Integrated Bayesian Model Classification. IEEE Transactions on Medical Imaging, 27(5): 629-640.

CUCALA, L., MARIN, J.-M., ROBERT, C.P., and TITTERINGTON, D.M., 2009. A Bayesian Reassessment of Nearest-Neighbour Classification. To appear in Journal of the American Statistical Association.

DEMUTH, H. and BEALE, M., 2002. Neural Network Toolbox User's Guide. The MathWorks, Inc.

DUDANI, S. A., 1976. The Distance-Weighted k-Nearest Neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics, 6: 325-327.

ELLIOTT, M. N., FREMONT, A., MORRISON, P. A., PANTOJA, P., and LURIE, N., 2008. A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity. HSR: Health Services Research, 43(5): Part I.

FLOM, L., and ROBINSON, C.. Using a Genetic Algorithm to Weight an Evaluation Function for Tetris. Available at www.cs.colostate.edu/~flom/TetrisPaperpdf.pdf.

FRIEDMAN, N., 1998. The Bayesian Structural EM Algorithm. 14th Conf. on Uncertainty in Artificial Intelligence (UAI '98), 129-138.

GONCALVES, J. F., MENDES, J. J. M., and RESENDE, M. G. C., 2008. A Genetic Algorithm for the Resource Constrained Multi-Project Scheduling Problem. European Journal of Operational Research, 189: 1171–1190.

GREEN, M. L., and KARP, P. D., 2004. A Bayesian Method for Identifying Missing Enzymes in Predicted Metabolic Pathway Databases. BMC Bioinformatics, doi:10.1186/1471-2105-5-76.

GUNGOR, T., 2004. Developing Dynamic and Adaptive Methods for Turkish Spam Messages Filtering. Technical Report 04A101, Bogazici University Research Fund.

HU, Y., JIANG, X., XIN, F., ZHANG, T., YUAN, J., ZHAI, L., and GUO, C., 2008. An Algorithm on Processing Medical Image Based on Rough-set and Genetic Algorithm. Proceedings of The 5th International Conference on Information Technology and Application in Biomedicine, in conjunction with The 2nd International Symposium & Summer School on Biomedical and Health Engineering, 109-111.

KELLY, J. D., and DAVIS, L., 1991. A Hybrid Genetic Algorithm for Classification. In Proceedings of the International Joint Conference on Artificial Intelligence , 645-650.

KIM, S.B., RIM, H.C., YOOK, D.S., LIM, H.S., and ANSEO-DONG, C., 2002. Effective Methods for Improving Naive Bayes Text Classifiers. PRICAI 2002: Trends in Artificial Intelligence: 7th Pacific ,LNAI 2417, 414-423.

KOTSIANTIS, S., ATHANASOPOULOU, E., and PINTELAS, P., 2006. Logitboost of Multinomial Bayesian Classifier for Text Classification. International Review on Computers and Software (I.RE.CO.S.), 1(3): 243-250.

KUBOTA, R., UCHINO, E., and SUETAKE, N., 2008. Hierarchical k-Nearest Neighbor Classification Using Feature and Observation Space Information. IEICE Electronics Express, 5(3): 114-119.

LEE, Z. J., SU, S. F., CHUANG, C. C., and LIU, K. H., 2008. Genetic Algorithm with Ant Colony Optimization (GA-ACO) for Multiple Sequence Alignment. Applied Soft Computing, 8: 55–78.

LIU, G., ZHANG, J., GAO, R., and SUN, Y., 2008. A Coarse-Grained Genetic Algorithm for the Optimal Design of the Flexible Multi-body Model Vehicle Suspensions Based on Skeletons Implementing. First International Conference on Intelligent Networks and Intelligent Systems, 139-142.

LU, J., WANG, L., LU, J., and SUN, Q., 2008. Research and Application on kNN Method Based on Cluster before Classification. Proceedings of the Seventh

International Conference on Machine Learning and Cybernetics, Kunming, 307-313.

MANTERE, T., 2006. A Min-Max Genetic Algorithm with Alternating Multiple Sorting for Solving Constrained Problems. Finnish Artificial Intelligence Society FAIS, 61-67.

MITCHELL, T. M., (1997). Machine Learning. McGraw-Hill, New York.

MOREIRA, J. J., 2008. A System for Automatic Construction of Exam Timetable Using Genetic Algorithms. Tékhne, 6(9): 319-336.

NEAL, R. M., 2004. Bayesian Methods for Machine Learning. NIPS Tutorial.

OZCANLI, O. C., DUYGULU-SAHIN, P., and YARMAN-VURAL, F. T., 2003. Açıklamalı Görüntü Veritabanları Kullanarak Nesne Tanıma Ve Erişimi, 11. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 18-20.

PATEL, N. R., 2003. Artificial Neural Networks. Massachusetts Institute of Technology.

POHLHEIM, H., 2003. Genetic and Evolutionary Algorithms: Principles, Methods and Algorithms. Genetic and Evolutionary Algorithm Toolbox, http://www.geatbx.com/docu/algindex.html.

REMUS J. J., MORTON K. D., TORRIONE P. A., TANTUM S. L. and COLLINS L. M., 2008. Comparison of a Distance-Based Likelihood Ratio Test and k-Nearest Neighbor Classification Methods. Machine Learning for Signal Processing, 362-367.

SONG, Y., HUANG, J., ZHOU, D., ZHA, H., and GILES, C. L., 2007. IKNN: Informative K-Nearest Neighbor Pattern Classification. PKDD, 248-264.

VANNUCCI, M., SHA, N., and BROWN, P. J., 2005. NIR and Mass Spectra Classification: Bayesian Methods for Wavelet-Based Feature Selection. Chemometrics and Intelligent Laboratory Systems, 77: 139-148.

WANG, B., and SHI, C., 2006. A Hybrid Genetic Algorithm for Two Types of Polygonal Approximation Problems. D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNCIS, 345: 30–41.

WANG, N., ZHUANG, J., DU, H., and WANG, S., 2008. A Self-Organization Genetic Algorithm with Cycle Mutation. 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2: 530-533.

WEI, L., 2004. Using Genetic Algorithm for Network Intrusion Detection. Proceedings of the United States Department of Energy Cyber Security Group 2004 Training Conference.

WEINBERGER, K. Q., and SAUL, L. K., 2008. Fast Solvers and Efficient Implementations for Distance Metric Learning. ACM International Conference Proceeding Series, 307: 1160-1167.

YILDIZ, T., YILDIRIM, S., and ALTILAR, D. T., 2008. Spam Filtering with Paralellized KNN Algorithm, Akademik Bilişim 2008.

YU, J., CHENG, F., XIONG, H., QU, W., and CHEN, X., 2008. A Bayesian Approach to Support Vector Machines for the Binary Classification. Neurocomputing, 72: 177-185.

ZENG, Y., YANG, Y., and ZHAO L., 2008. Nonparametric Classification Based on Local Mean and Class Statistics. Expert Systems with Applications, doi:10.1016/j.eswa.2008.10.041.

ZHANG, J., and TONG, C., 2008. Solving TSP with Novel Local Search Heuristic Genetic Algorithms. Fourth International Conference on Natural Computation, 1: 670-674.

ZHANG, Z., CHENG, H., CHEN, W., ZHANG, S., and FANG, Q., 2008. Correlation Clustering Based on Genetic Algorithm for Documents Clustering. IEEE Congress on Evolutionary Computation, 3193-3198.

ZHENG, Z., and WEBB, G. I., 2000. Lazy Learning of Bayesian Rules. Machine Learning, 41: 53–87.

BIOGRAPHY

I was born in Adana, in 1981. I have completed my elementary education at Celalettin Seyhan Primary Education School. I went to high school at Adana Kurttepe Anatolian High School and graduated at 1999.

I have graduated from department of Computer Engineering of Eastern Mediterranean University in 2004 with OSYM Full Scholarship. Since 2008, I have been working at Çukurova University Computer Centre Software Department.

My interest areas are software development with visual programming and machine learning.