Hybrid Intelligent Modeling Technique for Data Classification

[1]Tanu Rani, [2]Narender Kumar
[1] PG student, [2] Assistant Professor
[1, 2]Department of Computer Science and Engineering
[1, 2]Guru Jambheshwar University of Science & Technology, Hisar, India
[1]tanusinha786@gmail.com, [2]narenderster@gmail.com

Abstract:
Classification is technique of data mining to Predicts the categorical or class of unseen data. It is supervised learning method. In supervised learning, class of each samples are given. It can be separated into binary classification and multiclass classification. In binary classification, two classes are used and in multiclass classification more than two classes are used. The classification of multi-class datasets is more difficult as compared to the binary data classification. In this paper, we present a hybrid technique of GA (Genetic algorithm) and ANN (Artificial Neural Network) for multiclass problem. Genetic algorithm is used to improve the performance of neural network for multiclass data classification. GA optimizes the feature and provides the weight to ANN classifier. Proposed technique classifies IRIS, LYMPHTICS, ZOO, ECOLI and WINE multiclass datasets. To demonstrate the results, all dataset taken from UCI machine learning repository and compared the accuracy, specificity, sensitivity and f-score of proposed algorithm with respect to the standard ANN algorithm.

1. Introduction

The amount of data stored in the database is growing exponentially. This stored data contains the beneficial and valuable information, which usually upgrade the decision-making procedure of an organization. It is necessary to analyze this huge amount of data to extract useful information from it. Thus, there is need of automatic methods to mine knowledge from huge data. The method that is used for mining the knowledge is called data mining and knowledge discovery. Knowledge discovery in database (KDD) analyze and modeling the large dataset. It includes various pre-processing steps that make dataset more suitable for mining the knowledge. Data-mining is the core step of a knowledge discovery. Data mining is the process of mining useful knowledge from large amounts of data. Data mining provides many techniques for extracting useful information likes clustering, classification, rule-mining etc. Classification is building models that analyze and classify the data. Machine Learning breaks classification into binary, multi-class, multi-labeled, and hierarchical tasks [1]. To classify multi-class datasets is more difficult than the binary data classification. In multiclass makes a presumption that each instance is assigned only one label. Our paper focuses on the multi-class problem, and presents a new hybrid method based on GA and ANN. Hybridization improves the accuracy and speed of classifier. GA is a subset of EA (Evolutionary algorithm) that is widely used to build automatic model for training ANN parameter. Goal of this propose work is to build a novel model for multi-class problem. Multiclass data IRIS, ZOO, LYMPHATIC, ECOLI and WINE obtained from UCI machine learning repository. Results presents in term of the accuracy, specificity, sensitivity and f-score. Proposed technique give better performance for multiclass dataset compared with standard ANN.

2. Related Work

Tsun-Chen et al. [2] proposed GA/ANN hybrid technique for multiclass cancer dataset. The authors use a Genetic algorithm for gene subset selection. Artificial neural network (ANN) applies on selected subset of genes of cancer. Ashraf Osman Ibrahim et al. [3] presented a hybrid technique for classification. They integrate non-dominated sorting genetic algorithm-II (NSGA) with Three-Term Back propagation algorithm (TBP).This hybrid technique compared with multi-objective genetic algorithm based TBP network (MOGATBP). To optimize the accuracy and complexity, NSGA-II

hybrid with Local Optima search algorithm. Multi-class problem solves by BTP. Sungmoon Cheong et al. [4] integrate Support Vector Machine (SVM) technique with Binary Tree Architecture for the multi-class problem. SVM was primarily designed for the binary-classification problem and it was extended for the multi-class problem. SVM decompose the multi-class problem into many binary-class problems and then incorporate many binary -problem. Modified SMO convert multi-class problem into binary tree and SVM made binary tree. Chih-Wei Hsu et al. [5] compare method for multi-class Support Vector Machine (SVM). The author worked on two methods of multiclass SVM. One is combining several binary classifiers and other is taken all data together for optimization. The performance of these multiclass methods is compared with the three other methods that are based on the binary classification. These methods are one-against-all, one-against-one and directed acyclic graph (DAG).

3. Artificial Neural Network

Neural network is a composed set of interconnected input/output element called neurons. A weight is changes. ANN is example of supervised learning method. In supervised learning output value of each training tuples are given. An activation function applied with weighted input to achieve output signal. There are many types of activation function like associated with each interconnected link. ANN architecture consists of three layers: input layer, hidden layer and output layer. ANN during training phase updates the weights in response of input/output identity function, binary step function, Threshold, sigmoid function, Gaussian, ramp function [10].

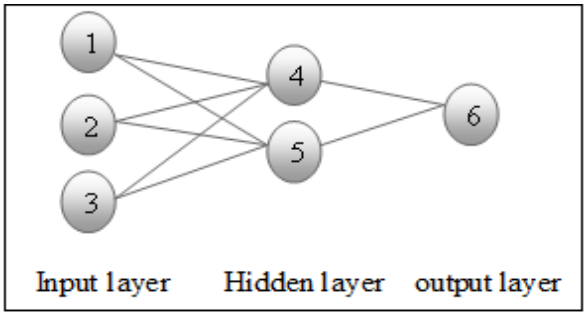

Fig: Artificial neural network

Feed forward neural network, feedback neural network type of Neural network Architecture. In feed forward neural networks signal flow in one direction. In feedback networks signal flow in both direction by using loop

Advantages of ANN

(a) Adaptive learning
(b) Fault tolerance

Limitation of ANN

(a) ANN training time is long.
(b) Computational cost of ANN is very high.
(c) Weight adjustment is difficult.

4. Genetic Algorithm

Genetic algorithm (GA) is a heuristic search and optimization algorithm. It is based on the mechanism of natural selection. It is an evolutionary algorithm inspired by Darwin theory of genetics Genetic algorithm takes two type processes, in first process generate new generation and in second process use the crossover and mutation operators [6]. It is initialized with a set of solution called chromosomes. A

chromosome is set of gene value that represents a candidate solution. Fitness function is used to measure the quality of a chromosome. Selection operators are use to determine which candidate solution are selected for reproduction. Crossover operator generates the new solution by swapping the some number genes of parents. Mutation operator also generates new solution by exchanging each gene of parents. Algorithm terminate when population converges towards the optimal solution [6].

4.1 Method of GA [7,8]

1. START - Generate random population.
2. FITNESS - Evaluate the fitness f(x) of each chromosome x in the population.
3. NEW POPULATION – Create a new population by repeating following steps until the new population is complete
REPRODUCTION OR SELECTION - Parents chromosomes are selected from population according to their fitness to crossover and produce new offspring.
CROSSOVER –crossover operator produce new two offspring from selected two parents based crossover probability.
MUTATION – Mutation operator produce new offspring by mutate single bit position in chromosome. Mutation used o maintain genetic diversity.
4. ACCEPTING - place new offspring in the new population.
5. REPLACE - Use new generated population for a further run of the algorithm
6. TEST – If the end condition is satisfied, stop, and return the best solution in current Population.
7. LOOP - Go to stop.

4.2 Flow chart of GA

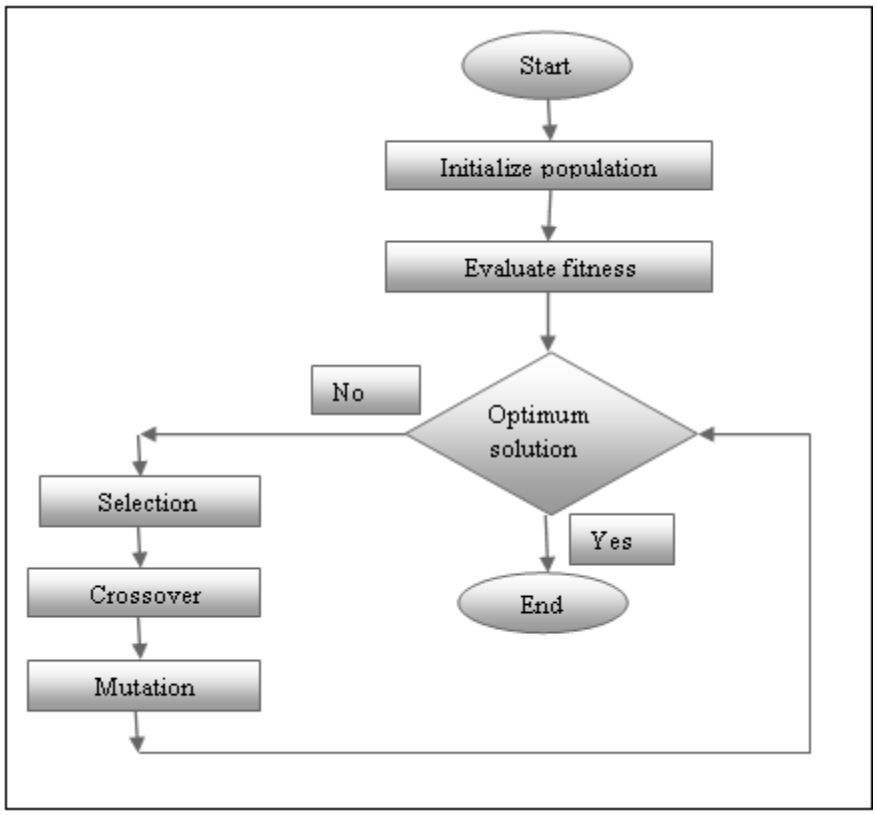

Fig: Flow chart of GA [9]

Advantages of GA

(a) GA has more efficient and faster compared to the traditional methods.

(b) It gives the facility of parallelism.

(c) It optimizes both continuous and discrete functions and also multi-objective problem.

5. PROPOSED METHOD (GA with ANN)

In the proposed research work genetic algorithm is embedded with neural network classification. In this algorithm, genetic algorithm is used to improve the performance of neural network classifier for multi class data classification. Although neural networks have proven to be a better approach for classification yet its training efficiency is a problem. Training performance of neural network depends upon various parameter like hidden layer size, weight and bias. Various algorithms like genetic algorithm, particle swarm optimization and gravitational search algorithm have been used in literature in order to get optimum value of these parameters. Genetic algorithm has proven to be a better approach when applied in field of optimization and parameter setting. Optimized Neural network provide better performance with simple architecture for binary class problems. Its architecture has to modify for multi class problem. One can design neural network with more than one output neurons, one for each class. Neuron with highest prediction is used to predict the class label. This approach is able to provide information regarding correlations between classes. Another approach for this is to design a separate network for each class.

This paper describes how genetic algorithm can be used to improve the performance of neural network. It also describes how neural network can be designed to solve multi class classification problem.

This Proposed algorithm is designed into 3 phases:

1. Genetic algorithm is employed to find optimum value for feature weighting
2. Neural Network is designed for multi class classification.
3. Optimum value obtained through GA is used to improve the performance of classifier.

5.1 Genetic algorithm

Genetic algorithm is a heuristic search based algorithm. It is an evolutionary algorithm inspired by Darwin theory of genetics. Genetic algorithm is able to provide optimum solution in large search space. Genetic algorithm is employed here to find optimum value of parameter for neural network. The implementation phase of genetic algorithm is described as:

Initialization

Population is initialized based on the acceptable minimum and maximum value. The initialization function is described as

$$pop_i = min_{value} + random_{value}\,(max_{value} - min_{value})$$

Here pop_i represents i^{th} individual of the population.

Fitness Evaluation

Fitness defines the quality of the individual. Fitness of the individual is computed using following fitness function.

$$fitness = \sum_{i=1}^{n} x_i \sin x_i$$

Here x_i represents the i^{th} part of the individual.

Genetic Operators
Genetic operators are used in genetic algorithms to maintain diversity among the individuals. It consists of selection operator that selects the individuals that can be recombined with other individuals using recombination operator.

Selection Operator
The selection operator deals with selecting the individuals to be preserved for next generation. Many selection methods are there like roulette wheel, rank-based, Proportionate selection, Steady state selection and tournament selection. In proposed work roulette wheel selection method is used.

Crossover operator:
It is also called recombination operator. It deals with generating new individuals by swapping the genes between two individuals. Many types of crossover are defined like one point cross over; two points cross over, uniform crossover and arithmetic crossover. In the proposed work arithmetic crossover is used. Arithmetic crossover produces new offspring by combining two individuals by using following equations

$$o_1 = t * parent1 + (1 - t)parent2$$

$$o_2 = (1 - t)parent1 + t * parent2$$

Here t is a random weighting number.

Mutation
Mutation operator aims to maintain diversity among the individuals by introducing new genetic material. This operator is applied on only one chromosome at a time and changes one or more gene value of the chromosome with a probability called mutation probability. Various type of mutation operators are there like flip bit, Boundary, non-uniform, uniform and Gaussian. In the proposed work Gaussian mutation operator is applied. Gaussian operator changes the gene value by adding a Gaussian distributed random value to the gene.

5.2 Neural Network

Neural network is a composed set of interconnected input/output units called neurons. The interconnection link is associated with weights. Neural network is able to classify any type of data without prior training of it. It is also able to deal with noisy data and data with complex decision boundary [1]. In the proposed work neural network is applied for multi-class classification.

Neural Network Design for Multi Class Classification
For multi class classification problem, class attribute is defined by more than two classes It can be represented with a variable of length *K* where *K* is the number of classes [11]. To map the multi class problem in neural network, architecture of neural network has to redesign. Neural network for multi class classification can be designed using: OAA, OAO and P AQs modeling technique In OAA technique, each class is trained against all other classes [12]. It can be implemented either using a

single neural network with m output nodes where m is the number of output classes or with a m binary neural network. In OAO technique, each class is trained against every other class separate neural network with a system of k (k-1)/ 2 binary neural networks [13].
In the proposed work, feed forward back propagation neural network is used.OAA technique is used to design neural network for multi class classification. In this method all classes is represented by a single neural network. Neural network is represented by m×n architecture where m is the number of input units and n is the number of output units. Number of input units depends on the number of features and number of output units depends in the number of classes in the data set. Size of hidden layer also depends on number of input units.

Input and Output Data
Input and output for neural network is prepared by dividing the data set into two matrix one for input and other for output where input matrix is represented by $m \times x$ .In which m is the number of attributes and x is the number of rows in training data set. Output matrix is also represented by $K \times x$ .In this K is the number of classes and x is the number of columns. Here x represent the training instance. For a given training instance x its class is represented by setting the Kth row set to 1.

Activation function and Training
In the proposed work, sigmoidal function is used as activation function for hidden layer and linear function is used as the activation function for output layer. Lavender-Marquardt back propagation is used for training the neural network.
Proposed algorithm is represented as

1. Divide the data set into training and test data.
2. Find the optimized weight using Genetic Algorithm.
3. Design neural network for multi –class classification using OAA technique.
4. Train the neural network using weight optimized using genetic algorithm.
5. Classify the test tuple.
6. Evaluate the performance.

6. EXPERIMENTAL DATASET
A large number of datasets are available on UCI machine learning repository that is used for experiment purpose. In this dissertation, we use the multi-class dataset. So many multi-class datasets are available on UCI machine learning repository. Five datasets are used in this dissertation. Name of the dataset is IRIS, LYMPHATICS, ZOO, ECOLI and WINE. The classification is the most widely used technique in data mining. Classification used many algorithms to classify the data in different ways. The performance of the classifier is measured in terms of accuracy. Datasets that are used in this dissertation are multiclass dataset. Details of these datasets are given in below table:

Table: Details of datasets

NAME of DATASET	Instance	Attribute	Class
IRIS	150	5	3
LMYPHATICS	148	18	5
Zoo	101	17	7
ECOLI	336	8	8

WINE	178	13	3

7. EXPERIMENTAL RESULTS

Proposed algorithm (ANN with GA) classify multiclass dataset and its performance shown in terms of accuracy, specificity, sensitivity, f-score and time taken by algorithm to complete scan the dataset. The performance of this technique is compared with the Standard ANN algorithm. **T**he performance of proposed technique and standard algorithm is analyse on different five multiclass dataset. All dataset are taken from UCI Machine learnig repository. All datasets are different in size, number of class and in number of instance. The performance is described with the help of table and graph. Table 1 show the result of proposed technique and result of standard ANN algorithm. Table 2 shows the training time of both algorithm. The performance of both algorithm can be presented more clearly with the help of graph.

TABLE 1 Experiment result for proposed Algorithm and Standard algorithm

Data set	Accuracy	Accuracy	Specificity	Specificity	Sensitivity	Sensitivity	F-score	F-score
IRIS	98.40	55.55	1	77.77	33.33	44.44	98.07	16.66
LYMPHATICS	91.25	86.95	1	1	80.49	81.59	82.58	73.66
Zoo	85.34	83.24	1	1	76.60	77.02	62.26	56.75
ECOLI	95.98	95.95	1	1	93.58	94.05	86.97	86.90
WINE	98.98	96.62	1	1	33.33	95.06	98.53	94.90

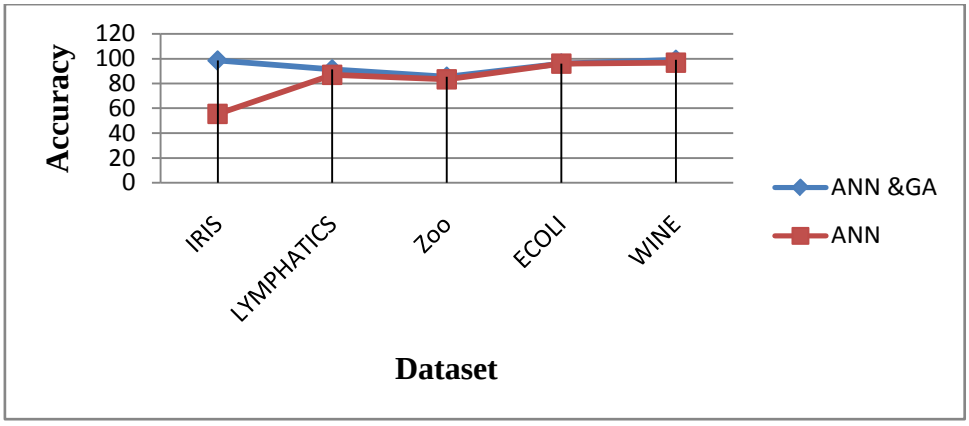

Fig: Accuracy of Standrad ANN and ANN with GA

The graph shows that accuracy of ANN with GA is better as compared to standrad ANN. For large dataset ANN with GA and standrad ANN gives comparable results.

Training time for Proposed algorithm and standard algorithm

TABLE 2 Training times of both Algorithm

Data set	ANN with GA	Standard ANN
IRIS	3	4
LYMPHATICS	3	6
Zoo	4	6
ECOLI	4	6
WINE	4	5

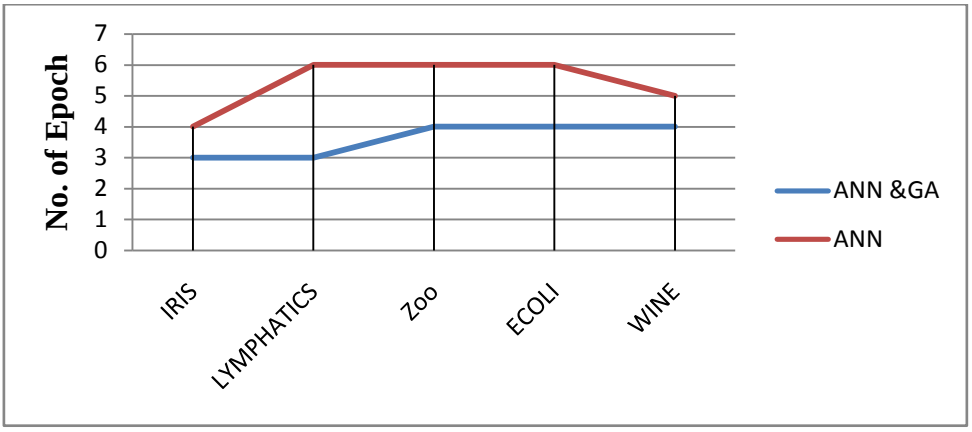

Fig: Training Time of Both Algorithms

The Graph shows the training time of Standard ANN and ANN with GA. It is analyzed from the graph that training time of ANN with GA is less as compared to standard ANN. Hence training time of ANN is improved.

From the above graphs and table it is concluded that performance of ANN is improved when it combined with GA. It gives better accuracy and takes less training time as compared to standard ANN. Hence proposed algorithm optimized the performance of ANN algorithm.

8. CONCLUSION

In this paper, standard ANN algorithm compare with the modify ANN algorithm. GA is used for feature optimization and ANN for classification. Modify ANN gives better result by taking less time in comparison to standard algorithm. So, it concludes that proposed algorithm perform better on multiclass dataset. In future, we combine the ANN with other technique for better classification.

REFERENCES

[1] Marina Sokolova , Guy Lapalme, " A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, pp. 427–437, 2009

[2] Tsun-Chen Lin , Ru-Sheng Liu , Ya-Ting Chao and Shu-Yuan Chen, "Multiclass Microarray Data Classification Using GA/ANN Method," *Springer*, pp. 1037-1041, 2006.

[3] Ashraf Osman Ibrahim, Siti Mariyam Shamsuddin,and Mohd Najib Mohd Salleh, "Hybrid NSGA-II of Three-Term Backpropagation Network for Multiclass Classification Problems," *IEEE, computer and information Science (ICCOINS) international conference*,2014.

[4] Sungmoon Cheong, Sang Hoon Oh, Soo-Young Lee, "Support Vector Machines with Binary Tree Architecture for Multi-Class Classification," *Neural Information Processing – Letters and Reviews*, Vol. 2, No. 3, pp. 47-51, March 2004.

[5] Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 415-425, MARCH 2002

[6] Ankit Maheshwari, Richa Garg, Er. Naveen Sharma, "A Review Paper on Brief Introduction of Genetic Algorithm," *International Journal of Emerging Research in Management & Technology*, Vol. 5, Issue-2, pp. 87-89, feb-2016.

[7] Jan H. Witten & Eibe Frank, "DATA MINING Practical Machine Learning Tools and Techniques," *ELSEVIER*. 2005.

[8] Tanu Rani, Mr. Narender Kumar, "A Survey: Hybrid Intelligent Modeling Technique for Data Classification," *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 6, Issue 5, May 2017.

[9] Patrick Kwaku, Elias Nii Noi Ocquaye and Wolali Ametepe , "Review of Genetic Algorithm and Application in Software Testing," *International Journal of Computer Applications*, Vol. 160, pp. 1-5, February 2017.

[10] Ms. Sonali. B. Maind, Ms. Priyanka Wankar, "Research Paper on Basic of Artificial Neural Network," *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 2, Issue. 1, pp. 96 – 100, January 2014.

[11] Thiagogm, "4th and 5th week of Coursera's Machine Learning (neural networks)," *Thiago G. Martins*, 05-Jun-2013.

[12] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," *Pattern Recognition*, vol. 40, no. 1, pp. 4–18, Jan. 2007.

[13] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," *Pattern Recognition*, vol. 40, no. 1, pp. 4–18, Jan. 2007.