Hybrid Genetic and Dempster Shafer Theory based Classifiers for Multi-Class Classification Tasks

Jitendra Kumar, Patel Himanshu Yadav, Anurag Jain

University School opf I.T.
Guru Gobind SIngh Indraprastha University, Kashmere Gate , Delhi, India

Abstract
Classification is a data exploration and learning mechanism, which has been widely studied and a wide range of applications subject. Supervised Classification is based on association rules and if we increase number of association rule degree of accuracy of classification is also being increase but larger number of rule take longer time to classify. Recently researcher is focus to develop an model that increase the accuracy in minimum time. In this paper Genetic based multi class classification model is proposed. Proposed model also use Dumpster shafer theorem for confining resultant rule set generated by GA algorithm. This paper used wine data set available at UCI machine learning website for classification and applies 3 cross fold mechanism for cross validation.
Keywords-
Data mining, classification , Genetic algorithm, Multi Class classification, Dempster shafer theorem

1. Introduction

Data Mining is a technique you can find valuable information or knowledge of many primitive data. Unlike static database mining, mining of data streams many new challenges. [21] First, each data element must be examined more than once. It is unrealistic to keep the whole sequence in the main memory. Second, each data item in the data stream must be processed as quickly as possible. Third, the use of the memory for mining data stream should be limited, even if new data items are continuously generated. Finally, the results generated by the algorithms must be available online instantly when the user requests. [6]

Traditional mining algorithms make data processing, considering that the data is organized in a single structure, in general, a file or table. This limitation makes it difficult to use such algorithms, for example, a relational database consists of multiple tables semantically related [1]. Mining algorithms Multi-relational data come as a viable proposition to the limitations of traditional algorithms, making it possible to extract multiple records patterns of direct and effective manner without the need to transfer data to a single table [2] [3] . However, the available memory space used equipment may not be sufficient to process the removal of large volumes of data. Because of this, the performance and memory space usage of such algorithms become an inherent concern for the exploration of large repositories.

Classification [13] is an important data mining and machine learning technique, which has been studied extensively and has a wide range of applications subject. Classification based on association rules, also called associative classification is a technique that uses association rules to build the classifier.

This paper presents an original contribution as a multiple relational algorithm for mining association rules focused on use with large relational databases. This proposal takes into account the limited memory available, since the goal of this work is the removal of large volumes of data. To overcome this limitation, the concept of partition database is divided into units of a size that can be allocated memory is used to provide treatment. In addition, the multi-relational extraction algorithm had satisfactory performance which allowed a detailed analysis of databases. Algorithms multi-relational association rule mining using different approaches to represent and extract patterns. A first approach involves algorithms based on logic, also known as inductive logic programming (ILP) algorithms. The main feature of this approach is that the data and patterns are represented as data records, which are written in first order logic [3]. The most widespread ILP algorithm mining association rules is WARMR [8], which operates on the basis of Apriori [9] algorithm. Another ILP algorithm found in the literature is the farmer [10], which is similar to WARMR, but is more efficient in the use of trees ordered for organizing common elements. There are multi-relational algorithms based on the function and structure of data using traditional techniques, mainly the A priori and FP-growth. The intention of this approach is to extend traditional mining algorithms that work properly with the extraction of patterns from a single table, adapting to a multirelational context to allow processing directly pattern in relational databases. Among the algorithms that are extending the Apriori Apriori-Group [11] and AprioriMR [12].

2. Related Work

Dewan Md. Farid [1] proposed two independent hybrid algorithms for DT and NB classifiers. The hybrid

Manuscript received February 5, 2026
Manuscript revised February 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.2.21

decision tree is able to remove noisy data to avoid over fitting. The hybrid Bayes classifier identifies a subset of attributes for classification. Hybrid DT algorithm used a NB classifier to remove the noisy troublesome instances from the training set before the DT induction, while the second proposed hybrid NB classifier used a DT induction to select a subset of attributes for the production of naive assumption of class conditional independence. But Hybrid model not deal with dynamic feature set.

Seokho Kang [2] present a multi-class classification method based on heterogeneous ensemble. Heterogeneous classifiers consist of two phases: training heterogeneous one-class classifiers for each class using various one-class classification algorithms, and constructing an ensemble by combining the base classifiers using multi-response linear regression-based stacking. The use of various classification algorithms contributes towards increasing the diversity of the ensemble, while stacking resolves the normalization issues on different scales of outputs obtained from the base classifiers.

Joshi [3] analyze the task multi-class document classification and knows that it can achieve high classification accuracy in the context of text documents. Naive Bayes approach to dealing with the problem of classification of documents through a deceptively simplistic model is used: assume all the features are independent of each other, and class calculated based on the maximum likelihood document. Naive Bayes approach is applied in Flat (linear) and hierarchical manner to improve the efficiency of the classification. It has been found that the hierarchical classification technique is more effective then the flat classification.

In [4] Directed acyclic graph discussed , Support Vector Machine (DAGSUM) based on Analytic Hierarchy Process (AHP) to classify e-Complaints documents in four classes based on the importance and urgecy. In [5] Ahmed presents a new algorithm H-RISC (Hierarchical SISC). H-SISC captures the underlying correlation between each pair of class labels in a multi-tab environment. The development of a robust multi-label classifier allow us to apply a model of this type in the classification of streaming text data more effectively.

3. Genetic Algorithm

The genetic algorithm (GA) is a research and technical optimization [24] based on the principles of genetics and natural selection. GA provides a population composed of many people (especially candidates) to perform as specified state selection rules that maximizes the fitness.A genetic algorithm is mainly composed of three operators: selection, crossover and mutation. By selecting a good channel is selected (based on fitness) to raise a new generation; Crossing combines good strings to generate improved seeds; mutation alters a local chain to maintain genetic diversity from one generation of a population of chromosomes to the next. In every generation, the population is assessed and approved by the end of the algorithm. If the failure criterion is not met, the population is operated by the three operators GA then reassessed. The cycle continues until the GA termination criterion is reached. In feature selection, genetic algorithm (GA) is used as a random selection algorithm, capable of efficiently explore large search space, which is usually required if selected attributes. For example; If the original features set contains a number of features, the total number of subsets candidates vying to be generated is 2N, which is a huge number even for medium N. Also, unlike many search algorithms that perform local greedy search, Gas conducts a global search. This offers a correlation process on the subset of the basis of selection using GA attribute. The correlation between the attributes to determine the suitability of an individual to engage in mating. Fitness function for GA is a simple function that affects a range of individual attribute based on correlation coefficients. Since highly correlated attributes cannot be part of DW together, only the attributes will be eligible to participate in crossover operations have lower correlation coefficients. In other words, we can say the closer the correlation is above the fitness value will be [24].

A genetic algorithm [25] is a type of the search algorithm. Space solutions for an optimal solution to a problem are sought. The algorithm creates a "population" of possible solutions to the problem and allows them to "evolve" over several generations to find a solution better. Algorithm starts with a set of solutions (represented by chromosomes) called the population. Solutions of a population are taken and used to form a new population. Cycle algorithm: The algorithm works through one cycle.

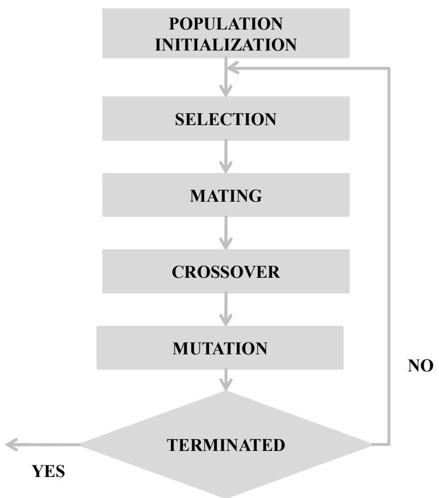

Figure 1 Flow Chart of Genetic Algorithm

4. Proposed Methodology

Recent research having lower predictive accuracy which lead to tends, combine existing [1] log-linear model with probabilistic techniques. While a search for informative aggregate features is computationally expensive, when it succeeds, the new aggregate features can increase the predictive accuracy. There are several possibilities for a combined hybrid approach.

(i) Once good aggregate features are found, they can be treated like other features and used in a decision tree.

(ii) A simple decision forest [2] is fast to learn and can establish a strong baseline for evaluating the information gain due to a candidate aggregate feature.

(iii) The regression weights can be used to quickly prune uninformative join tables with or small weights, which allows the search for aggregate features to focus on the most relevant link paths.

(iv) Whereas in [9] a hybrid mining algorithms to improve the classification accuracy rates of decision tree (DT) and naïve Bayes (NB) classifiers for the classification of multi-class problems but it's don't have any genetic algorithms, rough set approaches and fuzzy logic, be used to deal with real-time multi-class classification tasks under dynamic feature sets.

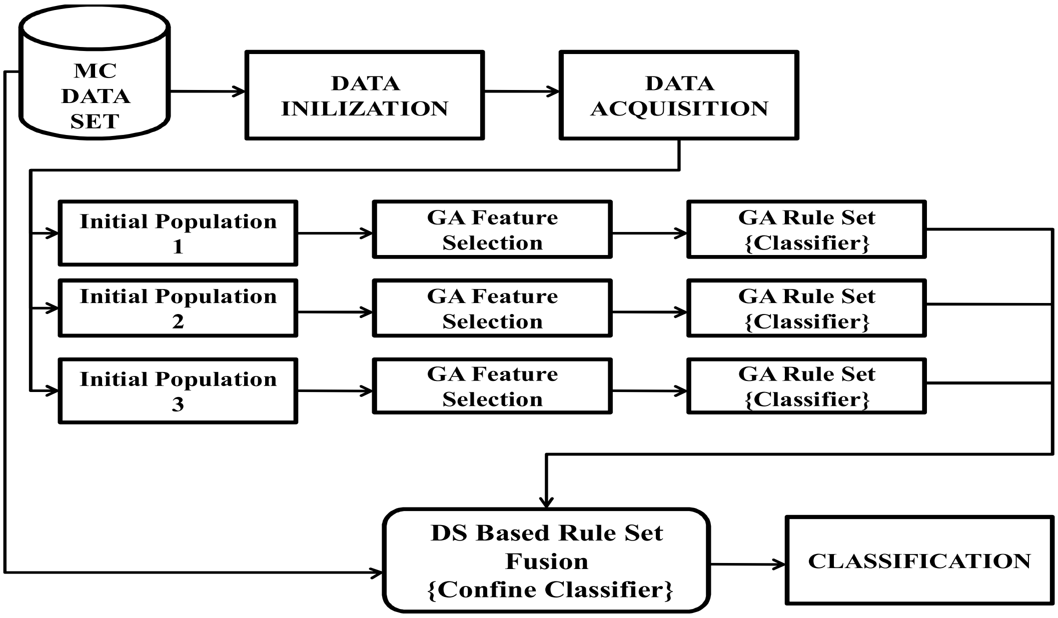

Figure 2 Proposed Model For GA-MCC

There are some limitation and problem of classification algorithm. now we choose association classification for classification algorithm and we used genetic algorithm along with probabilistic approach for the optimization of classification rate of association classification. In our case the results are improved because the genetic algorithm is a heuristic function. The heuristic function gives an optimal result. Whereas D-S theory approaches apply over historical data and give better result. Now we adopted optimization of classification of association rule with the help of genetic algorithm and D-S theory approach. In proposed model as shown in figure initially data initialized then acquiesced into different population on the basis of GA algorithm. GA algorithm also responsible for crossover and mutation based feature section and at last on the basis of that feature generate rule set for each population separately.

Whereas DS theorem apply for optimization and fusion of these rule set and generate confine classifier rule set. This resultant confines classifier responsible to classify mutli-class data set.

5. Result Analysis

Experimental results show that GA based MCC(Multi class classifier) gets higher accuracy comparing with the existing HMCC (Hybrid decision tree and naive Bayes classifiers). Rules discovered by HMCC have a more comprehensive characterization of databases. There is large possibility to extend HMCC rule set. Currently, MCC uses a different initialization of data set in proposed framework to discover feature set and generate classification rules. It may discover more relevant features of each class label by using related measures extending current framework. Also the current algorithm could be improved in terms of efficiency by using the optimization technique. Multiple relational classification algorithm modified by GA so improved rate of classification in comparison of HMCC. Our proposed algorithm test wine data set. In this data set the rate of classification is 92%.We also use another data set (abalone data set) and estimate some little bit difference of rate of classification is 91%.

Proposed model have a data set means wine data set and create a transaction table then select operation is performed of that data set. Now we write the working steps of this model.

a. Initialized data set.
b. Acquiesced data set
c. Initialized different population.
d. Set single point selection on each population.
e. Cross over the population on generate bit encoded.
f. Generate feature set for each population
g. By mutation generate classification rule
h. Finally apply DA for rule fusion.

Proposed methodologies build a hybrid model through classification association and genetic algorithm to increase data classification rate in order to hybrid Multi Relational association rule classification as shown in table 1. For implementing our proposed model for multiple class classification algorithms using GA simulated in the Matlab 10 and used wine data set and 3 fold cross validation for classification results validation.

Table 1: Resultant table for GA-MCC and HMCC

			Accuracy		Processing Time	
S. No	Cross Validation	Initialized Data Set	Proposed	Existing	Proposed	Existing
1	0.3	0.1	90.4415	82.3691	3.15122	2.37122
2	0.3	0.2	90.0846	80.4098	1.98121	2.07481
3	0.3	0.3	92.1243	82.4506	1.84081	1.98121
4	0.3	0.4	90.1639	82.4914	1.82521	1.80961
5	0.3	0.5	91.36	84.5322	1.77841	1.76281
6	0.3	0.6	90.2432	82.5729	1.82521	1.66921
7	0.3	0.7	91.2829	83.6137	1.82521	1.68481
8	0.3	0.8	90.3226	82.6545	1.74721	1.76281
9	0.3	0.9	90.3622	82.6953	1.85641	1.66921

In HMCC, DCT algorithm is used which classified only one type of data means high order data not low and average by using this type of algorithm low order data is unclassified and high order data is classified .So this lead to negative rule generation and classification rate would not be above 90% as show in figure 3.

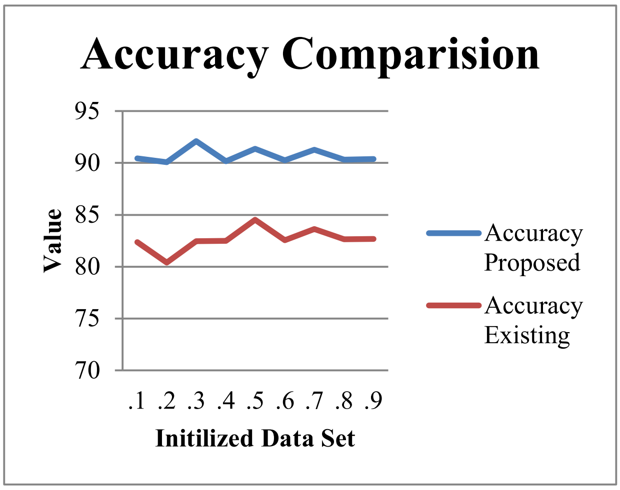

Figure 3: Comparative accuracy of HMCC and GA-MCC

Whereas In proposed GA-MCC low, average and high order data are classified easily because proposed methodology used GA which is computerized and optimization algorithm based on the mechanism of natural genetic and natural selection ,used genetic operation such as selection ,crossing and mutation and fitness function on the basis of data is optimized.

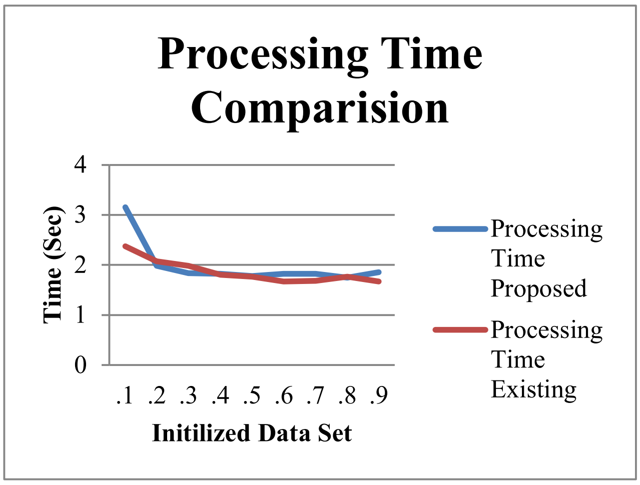

Figure 4: Comparative time complexity of HMCC and GA-MCC

Classification rate/accuracy increased above 90% as show in figure 3. Total execution time of multiple relation classification algorithm on wine data is 3.27602 Sec and classification rate accuracy is 82.6137% whereas total execution time of multiple relational classification algorithm using Genetic algorithm ie MCC Using GA on wine data set is 2.24641 Sec as shown in figure 4 and classification rate accuracy is 90.2829% as show in table 1, which is showing that classification rate accuracy increased above 90%.

6. Conclusion

In this paper a GA based multi class classification algorithm is proposed. GA-MCC use GA initialized data then acquiesced into different population. GA algorithm also responsible for crossover and mutation based feature section and at last on the basis of that feature generate rule set for each population separately. Whereas DS theorem apply for optimization and fusion of these rule set and generate confine classifier rule set. This resultant confines classifier responsible to classify mutli-class data set. GA-MCC classified Multi class data up to 92% in minimum time. In this paper whole result is verified through 3 fold cross validation.

References

[1] Dewan Md. Farid, Li Zhang, Chowdhury Mofizur Rahman, M.A. Hossain, Rebecca Strachan, "*Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks*", Expert Systems with Applications, Volume 41, Issue 4, Part 2, Pages 1937-1946, March 2014

[2] Seokho Kang, Sungzoon Cho, Pilsung Kang, Multi-class classification via heterogeneous ensemble of one-class classifiers, Engineering Applications of Artificial Intelligence, Volume 43, August 2015, Pages 35-43

[3] Joshi, S.; Nigam, B., "Categorizing the Document Using Multi Class Classification in Data Mining," in *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on* , vol., no., pp.251-255, 7-9 Oct. 2011

[4] Cholissodin, I.; Kurniawati, M.; Indriati; Arwani, I., "Classification of campus e-complaint documents using Directed Acyclic Graph Multi-class SVM based on analytic hierarchy process," in *Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on* , vol., no., pp.247-253, 18-19 Oct. 2014

[5] Ahmed, M.S.; Khan, L.; Rajeswari, M., "Using Correlation Based Subspace Clustering for Multi-label Text Data Classification," in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on* , vol.2, no., pp.296-303, 27-29 Oct. 2010

[6] Yingqin Gu1,2, Hongyan Liu3, Jun He1,2, Bo Hu1,2 and Xiaoyong Du1,2 "A Multi-relational Classification Algorithm based on Association Rules" pp.4-9 2009 IEEE.

[7] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient Classification Based on Multiple Class-Association Rules", Proceedings of the ICDM, IEEE Computer Society, San Jose California, 2001, pp. 369-376.

[8] X. Yin, and J. Han, "CPAR: Classification based on Predictive Association Rules", Proceedings of the SDM, SIAM, Francisco California, 2003.

[9] Xiao-Lin Li , Xiang-Dong He "A hybrid particle swarm optimization method for structure learning of probabilistic relational models" in transaction of Elsevier Information Sciences 283 (2014) 258–266

[10] Bahareh Bina, Oliver Schulte , Branden Crawford, Zhensong Qian, Yi Xiong "Simple decision forests for multi-relational classification " in transaction of Elsevier Decision Support Systems 54 (2013) 1269–1279

[11] Geetha Manjunath , M. Narasimha Murty , Dinkar Sitaram "Combining heterogeneous classifiers for relational databases" in transaction of Elsevier Pattern Recognition 46 (2013) 317–324

[12] Tahar Mehenni , Abdelouahab Moussaoui "Data mining from multiple heterogeneous relational databases using decision tree classification" in transaction of Elsevier Pattern Recognition Letters 33 (2012) 1768–1775

[13] Marko Debeljaka , Aneta Trajanova, Daniela Stojanovaa, Florence Leprincec, Sa D zeroski "Using relational decision trees to model out-crossing rates in a multi-field setting" in Ecological Modelling 245 (2012) 75– 83

[14] Dewan Md. Farid , Li Zhang "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks" in transaction of Elsevier of Expert Systems with Applications 41 (2014) 1937–1946