A New Approach for Optimization Classification Rule Generation Technique

Leena Joshi
M.E. (IT) Student(final year)
MedicapsInstitude Of
Tech.&Mgt.,Pigdambar,Rau,Indore(MP),India-452001

C.S.Satsangi, Ph.D
Professor IT
MedicapsInstitude Of
Tech.&Mgt.,Pigdambar,Rau,Indore(MP),India-452001

ABSTRACT
Data mining techniques enable an application to analyse a rich amount of data and recover the essential information from it. This information can use for decision making, pattern recognition and other applications. This data model can be a transparent data structure or a set of rules. In this presented work the transparent data models are investigated for optimizing their performance and resource consumption. During experiments that are observed these data models are accurately identify the patterns as defined the classification rules but the number of comparisons for classification increases as the number of rules generated are increases. The proposed technique first analyse the entire data samples and then the most optimum attributes are targeted for rule development. The proposed classification rule generation technique is efficiently generating less number of rules as compared to the traditionally available techniques. The implementation of the proposed concept is provided using MATLAB simulation tool and the performance in terms of memory consumption, time consumption and numbers of rules are evaluated. According to the obtained results the performance of the proposed rule generation technique is much efficient as compared to the traditionally available techniques.

Keywords
Data mining, rule mining, classification, implementation, rule optimization.

1. INTRODUCTION
The data mining algorithms [4] are used to analyse data and recover the hidden information from the input data. Using this hidden information the data mining algorithms evaluate the upcoming data patterns and identify the patterns. According to the input training data the data mining algorithms can be categorized in two main parts first the supervised data analysis and secondly the unsupervised data analysis. The supervised techniques support the classification and the unsupervised learning process is used for cluster analysis of data [1]. The data mining techniques according to their working can divided into two main categories. First can be termed as opaque data a model, these data models are process data internally and estimate the relation and weights that represent the relationship between data. Second the transparent data models the data is arranged using a data structure or rules representation. These rules and data structures are helpful for representing the whole datasets classification. The presented study is based on the transparent rules design and development.In this work the transparent data models are investigated thus the fuzzy technique of classification rule generation and ACO (ant colony optimization)algorithms[2] are initially implemented and their rule generation is understood then after for optimizing the performance of classification and computational complexity a new technique is developed for data analysis and classification rule generation.Basically the ant colony optimization is an optimization technique which is used to find optimum solution over a number of existing solutions. Therefore sometimes this algorithm is works as a kind of search technique. Basically using ant colony optimization algorithm generates one rule for each instance that represents the whole dataset as rule. According to observation that is found that the ACO design similar number of rules as the number of data instances found in dataset. On the other hand the Fuzzy based rule generation techniques finds the probability distribution of the data instances over the available class labels. Using the estimated probability distributions the membership of attributes with respect to the other attributes is computed to convert into the rules sets. The given techniques are significantly efficient for rules generation but due to observation that is found that the amount of rules can be more optimized. This chapter provides the basic understanding of the proposed study domain and the further chapter provides the design and implementation of the proposed system.

2. PROPOSED WORK
The classification rule development is a classical transparent classification technique, in this technique the data set analysed and for providing the classification pattern a sequence of rules are generated. These rules are helpful to understand the attributes distributions over the class labels[3]. The classical rule development technique suffers from the following key issues that are required to resolve in this study.

1. Most of the classification rule generation techniques offers a single rule at a time for the input training samples
2. The number of rule generated is equal to the number of examples available in dataset thus the number of comparison cycles are required for classy the entire set of data.
3. Available rule optimization technique only written on the basis of rule quality in other words the optimization techniques works after generating the rules for classification thus the time consumption is higher
4. More number of rules consumes more time for comparison and significant amount of main memory, therefore required to reduce the time and space complexity for rule based classification.

Proposed Solution

In order to find an solution for optimizing the traditional rule generation technique the following methodology is suggested.

5. **Obtain the class distribution of the input dataset:** The training sample for classification rule generation contains the entire knowledge for representing the data and the patterns are identified as the instance class labels. Thus the distributions of the data over each class are required to estimate first.
6. **Obtain the class distribution over attributes:** In this step the class distribution over individual data set is measured, for finding the most informative attribute selection.
7. **Find the attributes correlation and frequency for class distribution:** Before finalizing the attribute selection by which the rules are covered the total frequency of the attribute is measured and unique numbers of attributes are selected and their upper and lower domain of distribution is measured.
8. **Lowest frequency attribute is selected for rule development:** Finally optimum attribute according to the above constrains rules are developed.

2.1 Proposed rule generation technique

This section describes the proposed technique of the rule generation using attribute evaluation based algorithm. This technique first optimizes the data and then numbers of rules are generated. The algorithm steps are included in the below given table 2.1.

Table 2.1 The Proposed Rule Generation Technique

Input: Training examples
Output: Optimum rules list

Process:
1. Initialize the training samples
2. Find total number of classes in training example
3. Find number of attributes
4. Find unique attributes in attribute list
5. for each instance in population
6. if isunique(instance) then
7. listFinalattributes.add(instance)
8. else
9. remove(instance)
10. end if
11. end for
12. initializeatt[]; // array for holding the least number of attributes
13. for i=0 to number of attributes
14. if attribute[i].length > attribute[i+1].length
15. initializeatt[] = attribute[i]
16. End if
17. End for
18. For j=0 to initializeatt.length()
19. Rule ←Createrule (initializeatt[j])
20. If isruleexist(rule)
21. Remove(rule)
22. Else
23. Addtorulelist(rule)
24. End if
25. End for

Below given using figure 2.1 provides the overview of the proposed technique of rule generation. In this technique first data set is filtered for obtaining the unique list of data instances. Then after the total number of classes, number of attributes is extracted. Here the numbers of attributes are used to evaluate the attributes set available in training samples. After that the list of unique symbols are created attributes wise and less number of attributes is considered as optimum distribution of the class labels. Finally the class distribution matrix is applied for generating rules. After rule generation there are two cross checks are employed for finding the generated rules list. In first check the similar rules are removed from the final rules list. And in second rule the quality of rules are checked and assured that the generated rule having unique definition or not.

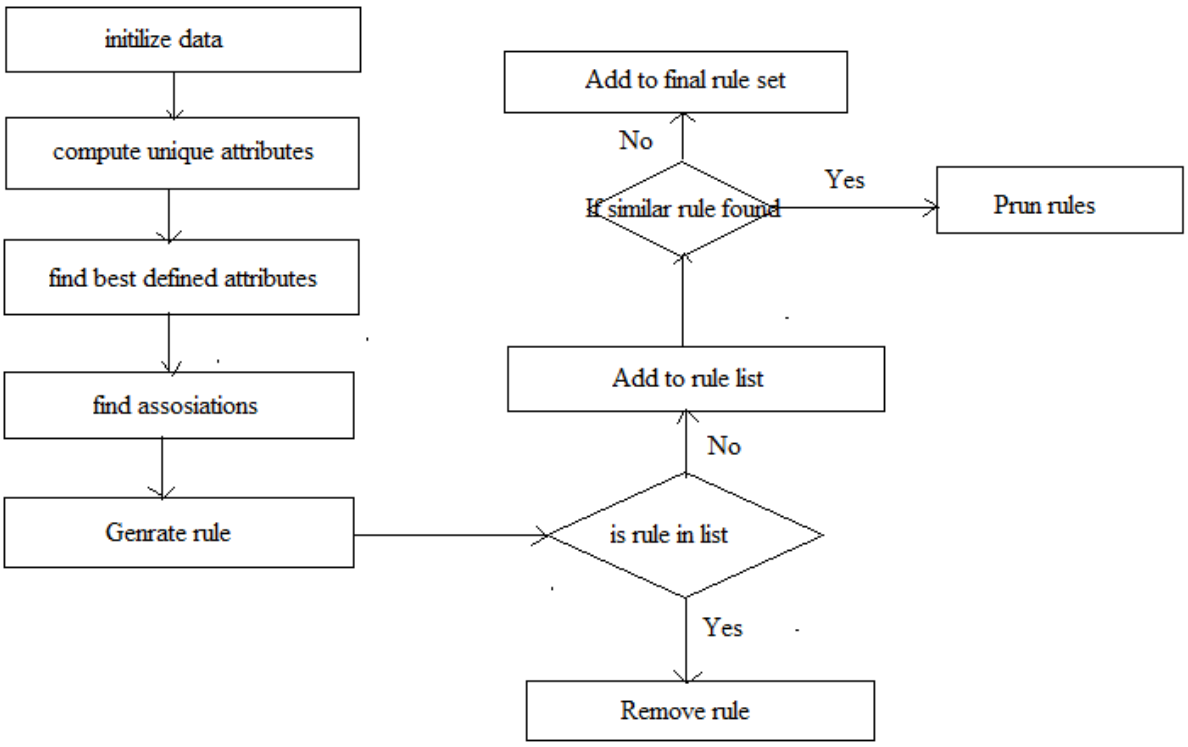

Fig 2.1 Proposed Rule Generation Techniques

2.2 Simulation Architecture

The simulation strategy of the proposed algorithm comparison and performance evaluation is given using figure 2.2. In this given system three different techniques for optimum classification rule generation process is given according to the given model first the system required to find the training samples for rule development than after there are a provision is made to select the appropriate algorithm. The selected algorithm processes the data and generates the rules for classification. After rule generation system computes the performance of algorithms and compared to other algorithm implemented with the system.

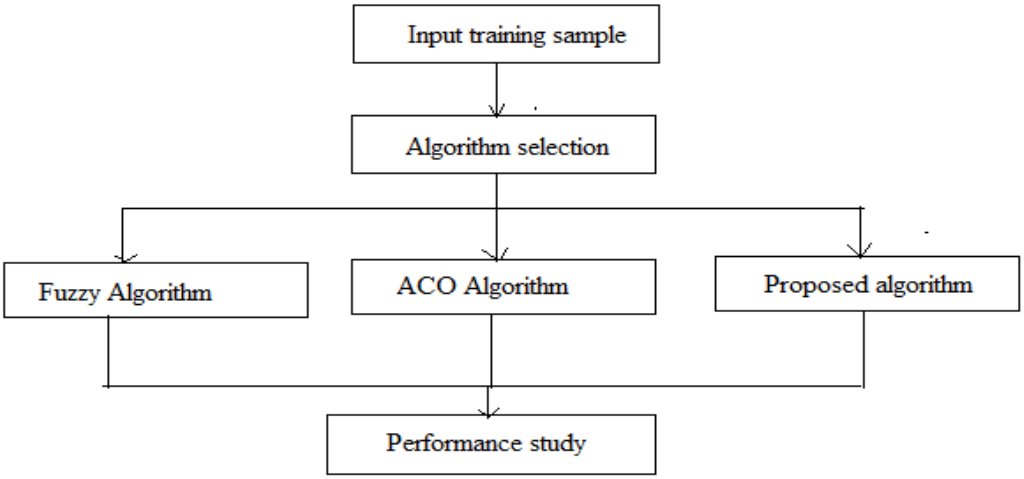

Fig 2.2 Simulation Architecture

3. RESULT ANALYSIS

After implementation of the proposed technique the performance of the different implemented algorithms are computed in terms of performance parameters[7]. The estimated performance using different datasets are given this section.

3.1 Time consumption

The amount of time required to generate the classification rules using different algorithms are known as the time consumption of the system.

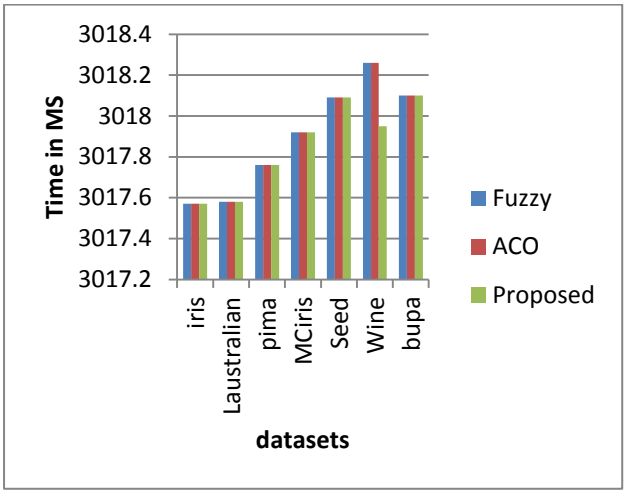

Fig 3.1 Time consumption

Table 3.1 Time Consumption

	Fuzzy	ACO	Proposed
iris	3017.57	3017.57	3017.57
Laustralian	3017.58	3017.58	3017.58
pima	3017.76	3017.76	3017.76
MCiris	3017.92	3017.92	3017.92
Seed	3018.09	3018.09	3018.09
Wine	3018.26	3018.26	3017.95
bupa	3018.1	3018.1	3018.1

The time consumption of the implemented algorithms are given using figure 3.1, in this diagram the X axis shows the experimental dataset used for rule extraction and the Y axis represents the amount of time consumed for rule extraction in terms of milliseconds. According to the obtained results most of the time the time consumption of all three algorithms are similar to the other methods implementation for comparative study.

3.2 Number of rules

The number of rules extracted from different algorithms is given using figure 3.2, the number of rules is demonstrating the number of comparisons required to classify the dataset. In this diagram the X axis shows the datasets which used for experiments and the Y axis shows the number of generated rules. According to the obtained results the proposed algorithm generates less number of rules as compared to the fuzzy rule generation technique and the ant colony optimization based techniques. Thus the proposed method of classification rule generation process is efficient and able to provide more optimum rules as compared to traditional approaches.

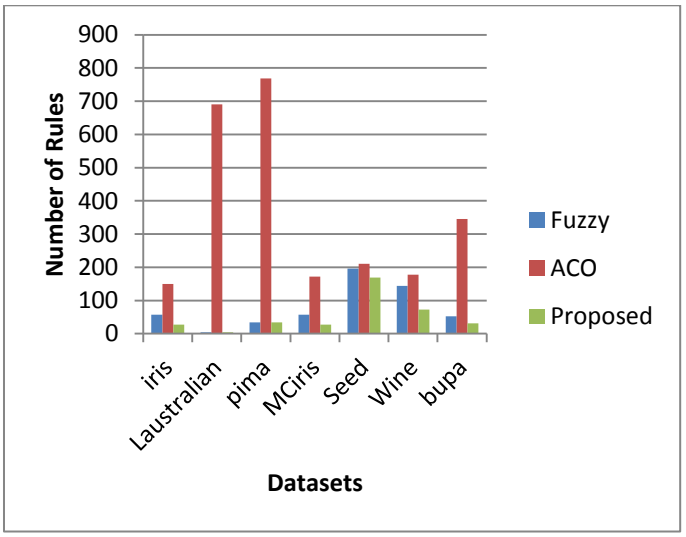

Fig 3.2 Numbers of Rules

Table 3.2 number of rules

	Fuzzy	ACO	Proposed
Iris	57	150	27
Laustralian	4	690	4
Pima	34	768	34
MCiris	57	172	27
Seed	196	210	169
Wine	144	178	73
Bupa	52	345	31

3.3 Memory consumption

The amount of main memory required to execute the algorithm for generating the classification rules is known as the memory consumption or memory utilization[9]. The amount of memory consumption with different dataset processing is given using figure 3.3 in this diagram the X axis shows the experimental dataset and the Y axis shows the Memory consumption in terms of MB (megabytes). According to the obtained results the proposed technique consumes less memory as compared to other traditional approaches of classification rules generation techniques.

Table 3.3 Memory Consumption

	Fuzzy	ACO	Proposed
Iris	0.171	0.172	0.078
Laustralian	0.125	1.344	0.109
Pima	0.173	0.938	0.125
MCiris	0.203	0.172	0.094
Seed	0.437	0.375	0.377
Wine	0.501	0.484	0.309
Bupa	0.14	0.328	0.094

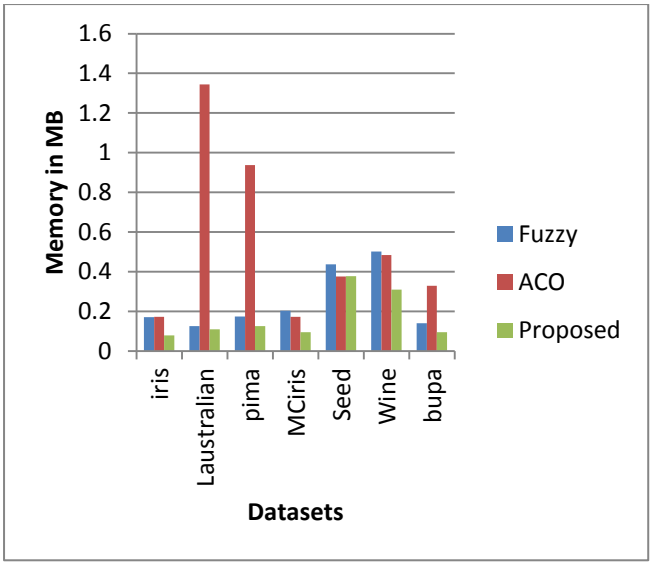

Fig 3.3 Memory consumption

4. CONCLUSION

Data mining is a process of data analysis and extraction the essential knowledge from the data. Thus different algorithms are applied to data for finding the essential hidden knowledge. After evaluation of data using the algorithms the algorithm generates the data model which is used for classification of new arrived samples. In this proposed work the key aim is to work with the transparent data models for rule extraction and classification and also intended to find an efficient and accurate technique for classification rule generation. Therefore a number of research articles are evaluated and found that the classification rules are help to convert the dataset knowledge in terms of transparent classifier. These rules are help to classify the data using less number of attribute comparisons thus the time consumption and memory consumption is optimized. Using this concept the three different classification rule generation techniques are selected for implementation namely ant colony optimization technique, fuzzy membership function based technique and the proposed classification rule generation technique.

Future work

The proposed work for classification rule generation technique is implemented successfully and able to extract the classification rules according to the given data samples additionally efficiently able to generate classification rules. The proposed technique is consumes similar amount of time as traditional approaches are consumed thus in near future the study in made to optimize the time consumption for data analysis.

5. REFERENCES

[1] KrasimiraKapitanova, Sang H. Son, Kyoung-Don Kang, "Using fuzzy logic for robust event detection in wireless sensor networks",2011 Elsevier B.V. All rights reserved.

[2] Fernando E. B. Otero, Alex A. Freitas and Colin G. Johnson, "A New Sequential Covering Strategy for InducingClassification Rules with Ant Colony Algorithms", IEEE Transactions On Evolutionary Computation, Volume 17, Issue 1, Pages 64–76, February 2013 (DOI: 10.1109/TEVC.2012.2185846)

[3] DataMining - Cluster Analysis,http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

[4] "Data Mining - Classification & Prediction Introduction", http://www.idc-online.com/technical_references/pdfs/data_communications/Data_Mining_Classification_Prediction.pdf

[5] Brain Decoding Of FMRI Connectivity Graphs Using Decision Tree Ensembles, 978-1-4244-4126-6/10/$25.00 ©2010 IEEE

[6] S Keskar, R Banerjee, "Time-Recurrent HMM Decision Tree to Generate Alerts for Heart-Guard Wearable Computer", Computing in Cardiology, 2011, 2011 - ieeexplore.ieee.org

[7] Khalid M. Salama, Ashraf M. Abdelbar, Fernando E.B. Otero, Alex A. Freitas, "Utilizing multiple pheromones in an ant-based algorithm forcontinuous-attribute classification rule discovery",© 2012 Elsevier B.V All rights reserved

[8] Burkay Orten, PrakashIshwar, W. Clem Karl,VenkateshSaligrama, Homer Pien, "Sensing-Aware Classification With High-Dimensional Data", 978-1-4577-0539-7/11/$26.00 ©2011 IEEE

[9] José Antonio Sanz, MikelGalar, AranzazuJurio, Antonio Brugos, Miguel Pagola, HumbertoBustincea, "Medical diagnosis of cardiovascular diseases using an interval-valuedfuzzy rule-based classification system", © 2013 Elsevier B.V. All rights reserved.

[10] Yoichi Hayashi, Tomohiro Takagi, Hiroyuki Mori, Hiroaki Kikuchi, Takamichi Saito, Hideaki Iiduka and SayakaAkioka, "Survey on the Family of the Recursive-Rule Extraction Algorithm",Journal of Computer Science Technology Updates, 2014, 1, 26-34