Effective Feature Selection Approach using Genetic Algorithm for Numerical Data

Ketan Sanjay Desale
Assistant Professor
Department Of Computer Egg.
D.Y.Patil School Of Engg.
Academy, Ambi, Pune

Balaji Mane
B.E. Scholar
Department Of Computer Egg.
D.Y.Patil School Of Engg.
Academy, Ambi, Pune

Prashant Berkile
B.E. Scholar
Department Of Computer Egg.
D.Y.Patil School Of Engg.
Academy, Ambi, Pune

Sushant Shivale
B.E. Scholar
Department Of Computer Egg.
D.Y.Patil School Of Engg.Academy,
Ambi, Pune

ABSTRACT
Data mining methods are used to handle the problems of dynamic huge data set. To build a classification model, time complexity of calculated result can be scale back by selecting only useful features. A feature selection technique is used to select only useful features from available features. An intersection principle based feature selection approach is Used. Genetic algorithm is used as a search method and it select only the features which are appears frequently in datasets. Then results were tested for different datasets having different type of data using Naive Bayes & J48 classifiers. The result analysis shows that Naive Bayes classifier gives better result than J48 classifier, with the substantial growth in accuracy, minimum time and minimum number of features.In this paper correlation feature selection is used with Genetic Algorithm for feature selection.

Keywords
Dimensionality Reduction, Feature Selection, Genetic algorithm (GA), Naïve Bayes, J48

1. INTRODUCTION
The feature selection Technique has been proposed as possibility of removing irrelevant and redundant options, increases efficiency in improving performance like predictive learning function, increases consistency of learned results and accuracy of the result.The feature selection technique is used as pre-processing Technique. It is a process of Selecting a subset of the original Features that will optimally reduce feature space according to a certain evaluation criteria. [1] The genetic algorithm selects some set of genetic operations between many available possibilities. [3] Feature selection technique reduces the number of features. It removes inconsistent, noisy, duplicate data. [5] The Correlation Feature Selection measure evaluates subsets of features. Feature subsets contain features highly correlated with the classification yet different from each other. [6] The proposed system works for below data sets 1) labor and 2) Iris for numerical type of data.

2. RELATED WORK
In this paper Genetic algorithm based feature selection approach is focused to improve its performance.

2.1 Feature selection
For each problem with some sample, where implementation decreases instead of increase which is called the curse of measurment there is a number of features. The need of an accurate mapping of low-measurment space of features is skilled so no data is lost by cancel important and basic features. Two problems one should focus to while doing this, how measurement can affect classification correctness and how measurement affects a classifier difficulty. A feature is good when it is aplicable, but not redundant to the other relevant features. There are two techniques to follow from this: feature extraction and feature selection. Feature extraction algorithms tend to create a new subset of features by combining existing features. Feature selection (FS) algorithms tend to limit the features to only those which would improve a task performance [7]. Feature selection algorithms are composed of three components: search algorithm, evaluation function, and performance function. The search algorithm could be: exponential – which is expensive to use as they have exponential complexity in a number of features, sequential where it adds and subtracts features, so they have polynomial complexity; or randomized , where it requires biases to yield small subsets, and they usually achieve high accuracies. A function to evaluate the candidate features for feature selection is an objective function. [7]. Feature selection is a process that selects a subset of original features. By using an evaluation criterion, feature subset's optimality is gauged. The N number of features goes on increasing as the dimensionality of a domain also expands. [5]. Feature selection has demonstrated in both hypothesis and practice to be compelling in improving learning proficiency, expanding prescient exactness and decreasing intricacy of learned results. Feature selection in directed learning has a principle objective of discovering a component subset that creates higher order exactness the feature selection impacts the prescient accuracy of any execution model, it is fundamental to study elaborately the viability of understudy execution model regarding feature selection technique. In this association, the present study is given not just to examine the most applicable subset features with least cardinality for adopting so as to accomplish high prescient execution different separated element choice systems in data mining.[9]

2.2 Correlation-based Feature Selection

At the heart of the CFS calculation is a heuristic for assessing the value or value of a subset of elements. This heuristic considers the handiness of individual elements for anticipating the class mark alongside the level of intercorrelation among them. The theory on which the heuristic is based is: Good component subsets contain highlights very connected with the class, yet uncorrelated with one another. In test hypothesis [6], the same standard is utilized to outline a composite test (the total or normal of individual tests) for anticipating an outer variable of hobby. In this circumstance, the features" are individual tests which measure attributes identified with the variable of interest (class). For instance, a more precise forecast of a man's accomplishment in a mechanics instructional class can be had from a composite of various tests measuring a wide mixture of qualities (capacity to learn, capacity to appreciate composed material, manual smoothness et cetera), as opposed to from any one individual test which measures a limited extent of characteristics[10].

A feature is considered as good feature if it is related to the class concept but it should not be redundant to any other relevant features.. CFS evaluates the importance of a subset of features using heuristic methods. A feature which is highly associated with a class is considered as good and selected. In each subset attribute are selected by considering the degree of redundancy between them and predictive ability of each individual feature. So, there is a need to define an appropriate correlation measure which can list most important and highly effective features.

2.3 Genetic Algorithm

Genetic algorithms (GA) are an adaptive heuristic search method based on the idea of natural selection. They are inspired by Darwin's theory of evolution – "survival of the fittest", which is one of the randomized search techniques. The algorithm begins with a set of individuals (chromosomes) called as population. Individual chromosome consists of a set of genes that could be bits, numbers or characters. Individuals are selected according to their fitness value for reproduction. Higher the fitness value more is the chances of an individual being selected. Crossover and mutation is responsible for producing new population. Crossover accelerates the search early in the solution of the population, while mutation is responsible for restoring the lost information to population by local or global movement in the search space. The process is iteratively repeated several times until stopping criteria are met or optimal solution is reached [8] .Genetic can be used to solve diverse types of problems [7]. Each rule in rule set identifies a particular attack type. Genetic Algorithms are another machine learning approach supported the principles of organic process computation. They incorporate the conception of Darwin's theory and natural selection to come up with a set of rules that may be applied on a testing set to classify infraction. Researchers have explored the employment of GAs in intrusion detection, and reportable terribly high success rates.

3. PROPOSED APPROACH

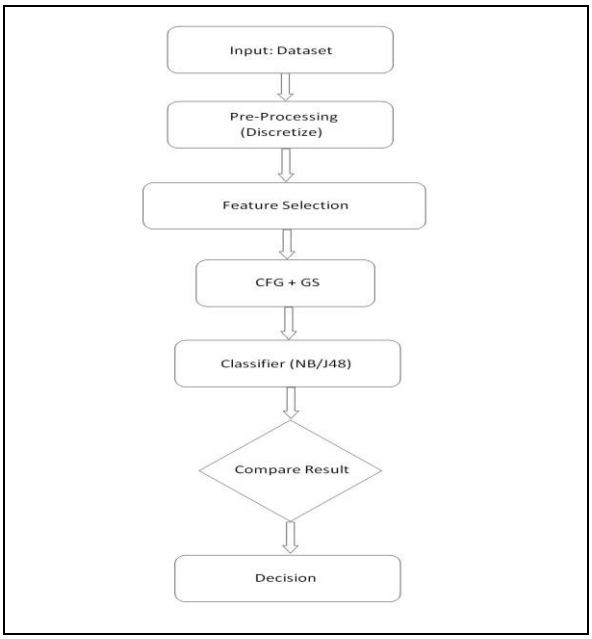

Fig 1: System Architecture

The complete framework of the proposed approach is described in following Figure. This paper describes an approach of using genetic algorithms as a Progressive search method while feature selection. In this experiment 5 data sets are used as 1) credit 2) Vote 3) Soyabean 4) Labor 5) iris Number of generation count and The population size can affect the performance of genetic algorithm. Here the principle of mathematical intersection is applied on above mentioned parameters.

3.1 Step 1: Data Pre-Processing

Data pre-processing is an capital footfall in the abstracts mining . The adage "garbage in, debris out" is decidedly applicative to abstracts mining and apparatus acquirements projects. Data-gathering methods are generally about controlled, consistent in out-of-range ethics (e.g., Income: −100), absurd abstracts combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing abstracts that has not been anxiously buried for such problems can aftermath ambiguous results. Thus, the representation and superior of abstracts is aboriginal and foremost afore active an analysis.If there is abundant extraneous and bombastic advice present or blatant and capricious data, again ability analysis during the training appearance is added difficult. Abstracts alertness and clarification accomplish can yield ample bulk of processing time. Abstracts pre-processing includes cleaning, normalization, transformation, affection abstraction and selection, etc. The artifact of abstracts pre-processing is the final training set. There are a number of data preprocessing strategies. Data cleaning is used to discard noise and incorrect deviation in the data. Data integration centeralize data from multiple sources into a consistent data store, such as a data warehouse. If Data processing techniques are applied before mining then it can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining.

3.2 Step 2: Feature Selection

One of FS techniques is applied as mentioned in section II. For correlation based feature selection method GA is applied using the proposed approach.Features selected after CFS

using the GA method are the intersection of combinations of population size and generations.

3.3 Step 3: Classification

Classification concept is the process of finding a model which describes and differentiates Data classes or concepts, which can be used to forcast the class of objects whose class label is unknown by using the model. The used model is based on the analysis of a set of training data. For performance testing features selected from above step are applied to Naive Bayes and J48 classifiers. The obtained results are then compared using metrics as a number of features selected, accuracy, time Required.

4. EXPERIMENTAL RESULT

As mentioned earlier, 5 datasets (CreditG, Soyabean, Vote, labor, iris) are used in the Experiment.The correlation based feature selection method is used to proposed approach. The performance used to compare classifier results are correct, timely to require to build models and number of features selected.

4.1 Accuracy

The anomalous behavior of the system is calculated as,

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Where,

• TP – is the True Positives which mean positive cases are incorrectly identified.

• TN – is the True Negatives which mean negative cases are incorrectly identified.

• FP – is the False Positives which mean negative cases are incorrectly identified as positive.

• FN – is the False Negative which means positive cases are incorrectly identified as negative.

Table 1. Results of Datasets with classifier without preprocessing

Dataset	Type	with NB	with J48
Creditg	nominal	75.4	70.5
Vote	nominal	89.47	73.68
Labor	Numerical	92.97	91.5
Soyabean	nominal	90.11	96.32
Iris	Numerical	96	96

Above table illustrates that result of Soyabean dataset is increased with j48 classifier as compared to naive bayes classifier.The classifiers directly applied on the datasets creditg, vote, labor, soyabean, iris without pre-processing.

Table 2. Results of Datasets with classifier only with preprocessing

Dataset	Type	with NB	with J48
Creditg	nominal	59.9	60.3
Vote	nominal	64.86	54.05
Labor	Numerical	94.17	95.55
Soyabean	nominal	59.66	63.04
Iris	Numerical	96.67	96

In this paper pre-processing is performed on the datasets credit, labor, vote, soyabean, iris.After that feature selection cfs+gs is apllied on the given datasets.Further the features are selected and navie bayes and j48 classifiers are applied. Above table illustrates that result of credit,Soyabean,labor datasets has increased with j48 classifier as compared to naive bayes classifier.result of the datasets labor and iris is increased after applying this classifiers.this technigue is gives efficient results on the numerical type of datasets.

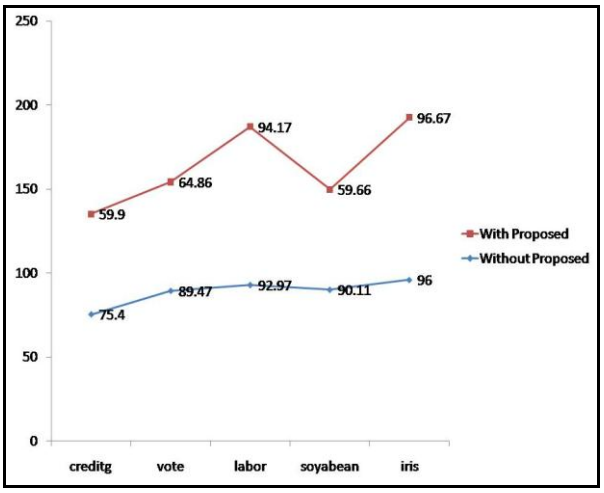

Fig 2: Result with Naive Bayes

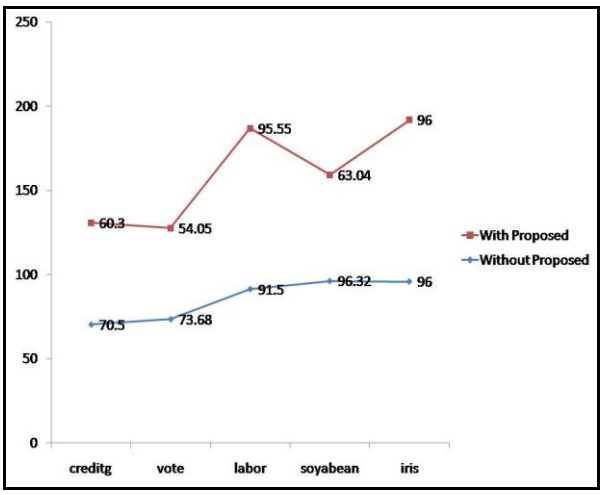

Fig 3: Result with j48

5. CONCLUSION AND FUTURE WORK

In this paper, mathematical intersection principle based innovative approach using genetic algorithm (GA) for the feature Selection is used. Feature selection is done using feature selection (FS) techniques i.e. CFS. Its effect on the different datasets is checked using two Naive Bayes and J48 classifiers. From the experimental results it can be concluded that the method helps in selecting the minimum number of features from the data set having numerical data type only which further improves the accuracy along with reduced time complexity. Future work would be focused on applying the proposed approach on other data types to improve the accuracy of classification method.

6. REFERENCES

[1] L Yu and H Liu, "Feature Selection for High-Dimensional Data – A Fast Correlation-Based Filter Solution", In Machine Learning-International Workshop Then Conference, Vol. 20(2), 2003, pp. 856.

[2] Selim Aksoy "Feature Reduction and Selection" Department of Computer Engineering Bilkent University CS 551, Spring 2012

[3] J. A. Vasconcelos, J. A. Ramírez, R. H. C. Takahashi, and R. R. Saldanha "Improvements in Genetic Algorithms" IEEE TRANSACTIONS ON MAGNETICS, VOL. 37, NO. 5, SEPTEMBER 2001.

[4] Cezary j. janikow "genetic Algorithms-Simulating nature's methods of evolving the best design solution" 0278-6648/95 IEEE FEBRUARY -MARCH 1995

[5] Rajdev Tiwari ,Manu Pratap Singh "Correlation-based Attribute Selection using Genetic Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010

[6] Mark A. Hall "Correlation-based Feature Selection for Machine Learning" Department of Computer Science, Hamilton, NewZealand

[7] Amira Sayed A. Aziz, Ahmad Taher Azar "Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation" 2013 Federated Conference on Computer Science and Information Systems pp. 769–774 978-1-4673-4471-5 2013, IEEE

[8] Mr. Ketan Sanjay Desale, Ms. Roshani Ade "Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System" 2015 International Conference on Computer Communication and Informatics (*ICCCI* -2015), Jan. 08 – 10, 2015, Coimbatore, INDIA

[9] M. Ramaswami and R. Bhaskaran," A Study on Feature Selection Techniques in Educational Data Mining, JOURNAL OF COMPUTING, VOLUME 1, ISSUE 1, DECEMBER 2009, ISSN: 2151-9617

[10] Mark A. Hall," Feature Selection for Discrete and Numeric Class Machine Learning" Department of Computer Science University of Waikato Hamilton New Zealand

[11] Anup Goyal, Chetan Kumar," GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System".