Prescient Precision Utilizing GABASS Approach over Bank Data

Kanika Choudhary
Research Scholar
Noida institute of Engineering and Technology, Greater Noida, India

Jaykant Pratap Singh Yadav
Assistant Professor Dept. of Computer Science
Noida institute of Engineering and Technology, Greater Noida, India

Pradeep Kumar
Assistant Professor Dept. of Computer Science
Noida institute of Engineering and Technology, Greater Noida, India

ABSTRACT
For improving accuracy in present work experiment is proposed over bank data to classify, according to the 11 existing feature. Classification problems frequently have a large number of features, but not all of them are utile for classification. Redundant and irrelevant features may be reduced the classification accuracy. Feature selection is a procedure of choosing a subset of significant components, which can diminish the dimensionality, abbreviate the running time. Genetic algorithm as an optimization tool and Naïve Bayes classifier will be used to compute the accuracy.

General Terms
Data mining

Keywords
Data mining, Feature selection subset, Data set, GABASS, Naïve Bayes classifier.

1. INTRODUCTION
Data mining (DM) is a general term that joins different computer based techniques to break down conceivably complex datasets. It is an interdisciplinary research territory where specialists from different zones, for example, measurements, software engineering, arithmetic and so on regularly cooperate with specialists from various application zones. DM strategies are generally conveyed to discover novel and valuable examples in information that may somehow or another stay obscure [1, 7]. DM is a critical thinking philosophy that finds an intelligent or scientific depiction, in the end of an intricate sort, of examples and regularities in an arrangement of information. It is Opposite to traditional research methods that are utilized to confirm or disprove a pre-characterized speculation. DM offers the likelihood to consequently produce new theories.

The feature selection problem has been important approach in DM from different perspectives. Prior to exploring the different approaches, presenting a general model from which the approaches originate is helpful. As the dimensionality of the attribute space increments, many sorts of information, investigation and arrangement likewise turn out to be fundamentally harder, and moreover the information turns out to be progressively inadequate in the space it involves which can prompt huge challenges for both administered and unsupervised learning [8]. Countless increment the clamor of the information and in this way the mistake of a learning calculation, particularly if there are just couple of perceptions (i.e. data sample) contrasted with the quantity of properties. In the most recent years, a few examinations have concentrated on enhancing highlight determination and dimensionality lessening strategies and significant advance has been acquired in choosing, separating and building helpful capabilities.

Estimation, Computation, Critical clustering, Data becomes increasingly sparse are the various problems occurs when searching in high-dimensional spaces is proposed to be applied where Naïve Bays Classifier will be used to compute the Classification accuracy that will be taken as the fitness value of the individual subset. In the area of feature subset selection for dimensionality reduction could not claimed that the solution provided by them in the most optimal solution so the scope of work remains open further and algorithm likes ACO[3], GA, co-relation based GA[4], Meta heuristic Search and PSO in past already applied for subset selection. In the present work, Genetic Algorithm based Attributes Subset Selection using Naive Bays Classifier is proposed. Aim is to enhance the execution after effects of classifiers yet utilizing a fundamentally diminished arrangement of features.

2. DATA SET
The dataset used in this model is bank information which is more precise and accurate in order to improve the predictive accuracy of data mining tasks. The data set may have missing (or) irrelevant attributes and these are to be handled efficiently by the data mining process [5]. The Bank dataset which consists of 601 instances and 11 attributes with 3 multi-valued attributes and 8 nominal attributes shown in table 1.

Table 1: Data set descriptions of bank information.

ATTRIBUTES	VALUE
Age	Numerical value
Sex	male(1), female(2)
Region	character value
Income	numerical value
Married	no(1), yes(2)
Children	numerical value
Car	no(1), yes(2)
Save_ act	no(1), yes(2)
Current_ act	no(1),yes(2)
Mortgage	no(1), yes(2)
Pep	no(1), yes(2)

3. GABASS

It was firstly proposed by Patel and Rao under the motivation of instructing learning phenomena of a classroom to understand multi-dimensional, direct and nonlinear issue. The effectiveness of the GABASS has been compared with the other population intelligence optimization algorithms based on the best solution, average solution, convergence rate and computational complexity. The result of GABASS is better performance, specific algorithm with less parameter, effective and efficient to solve optimization problem [9]. This optimization method is based on the effect of the influence of the Naïve Bays Classification. It is a population based method and like other population based methods it uses a population of solutions to proceed to the global solution. GABASS is an evolutionary based optimization algorithm**.**

3.1 Algorithm

1. Generation of subsets of attributes from the BANK dataset.
2. Initial population is created by taking the chromosome (selected subset) of attributes in step 1.
3. Naïve Bayes Classifier is applied on individual subset to compute the fitness (accuracy) value.
4. Applying the Roulette Wheel Selection chromosome are taken for the crossover operation with probability Pc= .8.
5. Some selected offspring generated in step 4 undergoes the Mutation operation with probability Pm= .001.
6. Offspring generated in step 4 & 5 are used to create a new population.
7. Until the stopping criterion is satisfied step 3 & 6 will repeated.

3.2 Cross Validation

The data is divided into training and test set and the validation are performed multiple times so that each example in the dataset is used as the training data at least once in cross-validation. The most common approaches are k-fold cross-validation, 2-fold cross-validation and leave-one-out cross-validation. In a k-fold cross validation, the data is divided into k mutually exclusive parts. One of the k part is taken as a test set and rest are combined together to induce the classifier. In the proposed methodology classification algorithm is 10times trained and tested. The cross validation divides the data into 10 subgroups and each subgroup is tested through classification rule constructed from the remaining 9 groups. Ten different test results are obtained for each train–test configuration and the average result provides the test accuracy of the algorithm.

4. NAÏVE BAYES CLASSIFIER

Naïve bayes is also known by Credulous Bayes. Credulous means contingent freedom among attributers of components. The "innocent" suspicion significantly lessens calculation unpredictability to a basic augmentation of probabilities. The Naive bayes handles numeric qualities utilizing managed discretization and employments part thickness estimators that will enhance the execution. It needs just little arrangement of preparing information to create exact parameter estimations since it requires as it were the computation of the frequencies of qualities and characteristic result combines in the preparation informational collection. The objectives of the research activity can be defined as follows:

• A reduction in the cost of acquisition of the data.

• A faster induction of the final classification model.

• An improvement in classification accuracy [10].

5. PROPOSED METHADOLOGY

In this proposed method a source bank dataset will be taken as input in arff file; arff is Attribute relation file format. After that all the attributes of datasets are encoded. A number of attributes are select randomly. The classification accuracy is compute with selected attributes. GABASS is applied to improve the Classification accuracy and update the attribute set accordingly. The algorithm is repeated number of times until termination criteria is satisfied. After termination we will get the subset of best feature and classification accuracy. Generalized form of algorithm is shown in Fig.1 (next page).

6. RESULT ANALYSIS

In this work, five different methods are used for feature selection. Forward Selection Multicross Validation, Bootstrap backward elimination, Relief, MIFS and proposed GABASS method are implemented and five different feature subsets were obtained. Forward Selection Multicross Validation and Bootstrap backward elimination are wrapper based method; Relief and MIFS are filter based method. To calculate the classification accuracy for above described methods; SIPINA tool of TANAGRA software is used. The selected feature subsets by these five methods are detailed in following table. The k-fold cross validation method was used to measure the performances. The performances thus measured are shown in table 2.

Table 2: Performance comparison between GABASS and Random

Sr. No	Attributed subset selected	methods	Classification accuracy %
		WRAPPER BASED METHOD[2]	
1	(region, married, mortgage, age, children)	Forward selection Multicross validation	72.33
2	(Sex, region, income, married, children, car, save_acc, current_acc	Bootstrap backward elimination	73.33
		CHANNEL (FILTER) BASED METHOD[6]	
3	(Children, Save_acc)	Relief	68.50
4	(Children, Married, Income, Sex)	MIFS	72.66
		PROPOSED METHOD	
5	(Children, Married, mortgage, sex, income, Save_acc)	GABASS	76.07

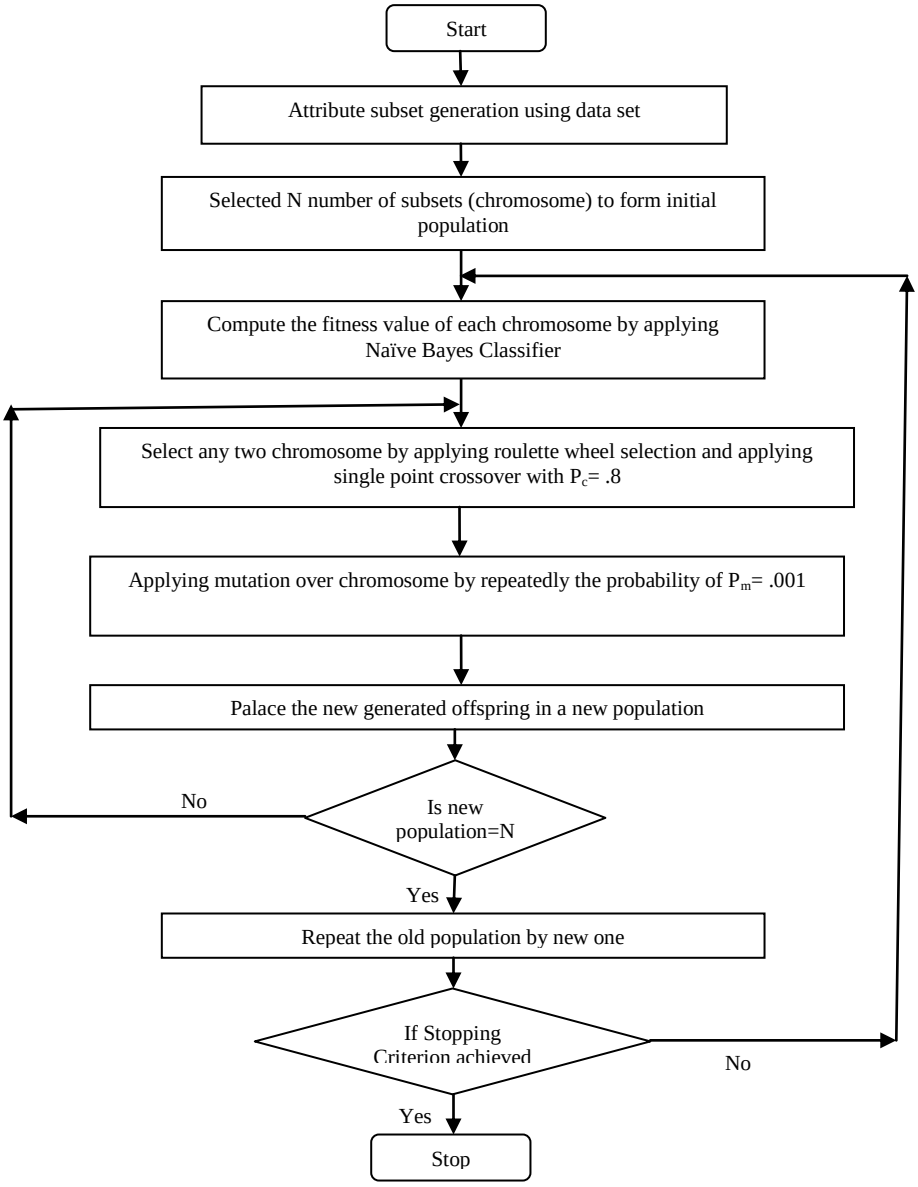

Figure 1: Generalized flow chart for GABASS

The comparison between Random and GABASS is shown with help of graph in Fig 2. This comparison is based on classification accuracy and it shown that GABASS is better than other methods.

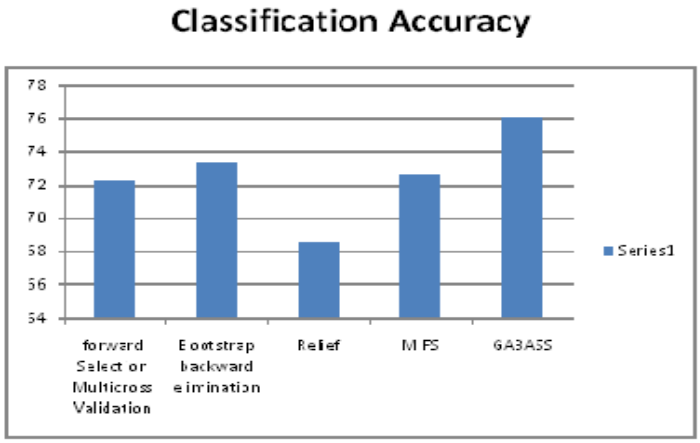

Figure 2: Graph Shows Comparison Between Random and GABASS

7. CONCLUSION AND FUTURE SCOPE

Feature selection is the primary errand of any learning way to deal with characterizes a pertinent arrangement of components. A few strategies are proposed to manage the issue of highlight choice including channel (filter), wrapper and installed (embedded) techniques. In this work, I focus on feature selection to choose an insignificantly measured subset of ideal elements .High dimensionality of data in the sense of many instances also increases the computational cost for the learning algorithm but has usually a positive influence on the classification accuracy. Feature reduction process should have a positive (or at least no negative) effect on the characteristics accuracy of the learning calculation. Feature Selection is streamlining issue; genetic algorithm based characteristic subset determination utilizing naive bays classifier is utilized for this reason. GABASS are observed to be the best system for choice reason when there is vast populace. The GABASS gives great outcomes and their energy lies in the great adjustment to the different and quick changing situations.

Future work will involve experiments on the datasets from different domains. The GABASS algorithm tested on different domains previously. The difference in performance and accuracy of different ensemble approaches will be evaluated. GABASS can give more efficient results and the optimization process can become much easier and faster. The one more important aspect of future work is of finding more factors that can compare two test suits for their goodness, so that efficiency of selection process can be improved.

8. REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Simon Fraser University, 2000

[2] Ron Kohavi, George H. John, "Wrappers for feature subset selection", Artificial Intelligence 97, pp. 273-324, 1996.

[3] Syed Imran Ali, Waseem Shahzad, "A Feature Subset Selection Method based on Symmetric Uncertainty and Ant Colony Optimization", International Journal of Computer Applications (0975 – 8887) Volume 60– No.11, 2012

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

[5] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset" Elsevier Expert Systems with Applications 38, 5947–5957, 2011

[6] Asha Gowda Karegowda, M.A.Jayaram, A.S .Manjunath "Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning" International Journal of Computer Applications (0975 – 8887) Volume 23– No.2, 2011

[7] Divya Chaudhary "Data Mining: Techniques and Algorithms" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 8, , pp. 475-479 August 2013

[8] T.R. Jerald Beno, M. Karnan, " Dimensionality Reduction: Rough Set Based Feature Reduction" International Journal of Scientific and Research Publications, Volume 2, Issue 9, September 2012

[9] R. Venkata Rao, Vivek Patel, "An elitist teaching-learning-based optimization algorithm for solving complex constrained optimization problems", International Journal of Industrial Engineering Computations 3, pp. 535–560, 2012

[10] Kashif Javed Butt, "A study of feature selection algorithms for accuracy estimation" Master in Artificial Intelligence, 2012