THE IMPLEMENTATION OF GABASS IN TO BANK DATA

[1]Pradeep Gupta, [2]Jay Kant Pratap Singh
[1]Assistant Professor, Ajay Kumar Garg Engineering College,Ghaziabad, UP,India
[2]Assistant Professor, Ajay Kumar Garg Engineering College,Ghaziabad, UP,India
[1]guptapradeep@akgec.ac.in [2]yadavjaykant@akgec.ac.in

Abstract: **Subgroup identification assesses the acceptability of a subset of the best characteristics as a group. Subset determination is characterized in many approaches, here we are discussing the implementation tool with what suitable parts are required and how they are applied. The method proposed is GABASS and it is applied over bank data. The implementation of the GABASS algorithm is discussed in this paper.**

Keywords: **Genetic Algorithm, GUI, JAVA, TANGARA.**

I. INTRODUCTION

In this paper, we are discussing complete details of the implemented tool along with the descriptions of the result obtained in the form of snapshots. Java tool developed to automatically select a subset of GABAS-based features. A GUI is also included with the tool, which is further illustrated in-depth.

This paper aims to improve the classification accuracy of current work tests on banking data by using 11 existing features and 601 examples. The data is divided into two parts one is training and the other is testing,the verification process is repeated several times, so that at least once during the procedure, each occurrence in the dataset can be utilized as training data. K-fold cross-validation, 2-fold cross-validation, and one-off cross-validation are the most prevalent approaches in cross-validation.K data is separated into related portions in k-cycle cross-validation. The test set is chosen from among the k components, and the remainder is combined to create the classifier [2].The classification algorithm is trained and tested ten times in the proposed approach.The cross-validation data is separated into ten subgroups, with each subgroup subdivided based on the classification rule established into the remaining nine subgroups.For each train test setting, ten separate test outcomes are obtained, and the common result algorithm determines the test's correctness.

The data set is collected from the source bank ARFF file as input in this proposed approach. The ARFF attribute is an ASCII text file that describes a set of circumstances in which a collection of characteristics can be shared. After that, all of the database's features are encrypted. Some characteristics are chosen at random. The classification accuracy is calculated using the features that have been chosen. GABASs were used to increase classification accuracy and refresh the attribute set. The procedure is repeated until the completion criteria are satisfied. We'll get a subset of the top attributes and classification accuracy after that's done.The accuracy of the system classification given above was calculated using the TANGARA programme SIPINA Toll.

II. JAVA

The appropriate language for implementing the method is JAVA, which is platform neutral thanks to the JVM. In JAVA JDK 1.8, our feature selection mechanism is implemented as GABAS. It may run on any JVM-enabled operating system, including Windows, Linux, and UNIX. The (org.um. feri. ears. algorithm. GABAS) package is used to develop the GABAS algorithm. Discretization of feature value could aid in calculating each feature's fitness and information gain value. We use the discretization package [3], which is supported by Weka 3.7.0, for each feature, to discretize a group of values. The result is a feature subset that may be used to build a Nave Bayes classifier to categorize the accuracy.

A. SIPINA Tool of TANGARA

TANGARA is SIPINA's successor, and it uses supervised learning methods, association rules, feature selection, and the creation of custom algorithms. TANGARA is open-source software because it is written in Java. TANGARA's major goal is to make data mining software that is simple to use, especially in terms of the user interface and how to use it. The second purpose of TANGARA is to build an architecture that would allow users to simply add their own data mining algorithms and compare their results [4].

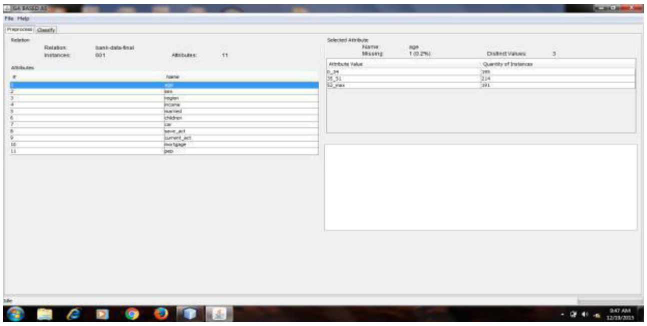
II. WORKING OF GABASS

a) Using java code, generate random subsets of attributes from bank data. There are 2n subsets of data D with n attributes. In general, the cost of computing data set D is O(n x |D| x log(|D|), where n is the number of characteristics and D is the number of instances. For m characteristics and n instances, the number of comparisons necessary is m*n2 [1].

A. GUI

As mentioned above that the Weka 3.7.0 is supported. Here we have to deal with datasets provided by the bank. The human-computer interface (GUI) employs windows, icons, pull-down menus, and a pointer, all of which can be controlled with a mouse. The utility has a graphical user interface (GUI) as depicted in Figure 1, with three command buttons labeled, (1) File, (2) Preprocess (3) Classify.

An ARFF format dataset file is viewed and taken as input by clicking the file button. An ARFF (Attribute-Relation File Format) report is an ASCII file that delineates a once-over of properties and shares a game plan.Figure 1 depicts one such list of characteristics. When you click on a specific attribute, you'll see its value and the number of times it's been used.



Relation
Relation: bank-data-final
Instances: 601
Attributes: 11

Attributes

#	Name
1	age
2	sex
3	region
4	income
5	married
6	children
7	car
8	save_act
9	current_act
10	mortgage
11	pep

Selected Attribute
Name: age
Missing: 1 (0.2%)
Distinct Values: 3

Attribute Value	Quantity of Instances
0_34	195
35_51	214
52_max	191

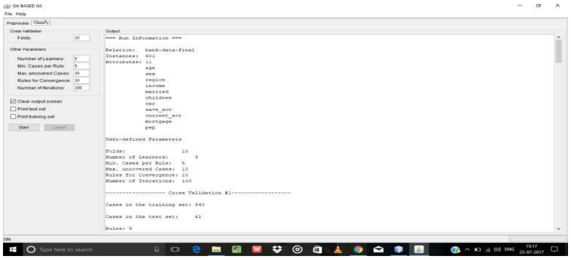
Figure 1 Attribute list

b) In figure 1, the initial population 11 is created by using the subsets generated as chromosomes.

c) The Naive Bayes classifier method is applied on individual subsets to compute the accuracy taken asa fitness value. This classifier has a minimum error rate. It performs consistently before and after reductions of numbers of attributes [5]. The fitness function is F= a-cnr+nr/2, where **"**a**"** is **"**accuracy**"**, **"**cnr**"** is **"**cases not covered**"** and **"**nr**"** is **"**number ofrules**"**.

A. Classification

After pressing the Classify button, Figure 2 presents a number of input boxes, checkboxes, and two buttons. User-defined values for various parameters are captured via input boxes. All checkboxes are optional and used to let the user make decisions depending on their preferences.



Cross Validation
Folds: 10

Other Parameters
Number of Learners: 5
Min. Cases per Rule: 5
Max. uncovered Cases: 10
Rules for Convergence: 10
Number of Iterations: 100

☑ Clear output screen
□ Print test set
□ Print training set

```
=== Run Information ===

Relation:   bank-data-final
Instances:  601
Attributes: 11
            age
            sex
            region
            income
            married
            children
            car
            save_act
            current_act
            mortgage
            pep

User-defined Parameters

Folds:                  10
Number of Learners:        5
Min. Cases per Rule:    5
Max. uncovered Cases:   10
Rules for Convergence:  10
Number of Iterations:   100

------------------ Cross Validation #1------------------

Cases in the training set: 540

Cases in the test set:     61

Rules: 9
```

Figure 2 Lists of Parameters

B. Parameters

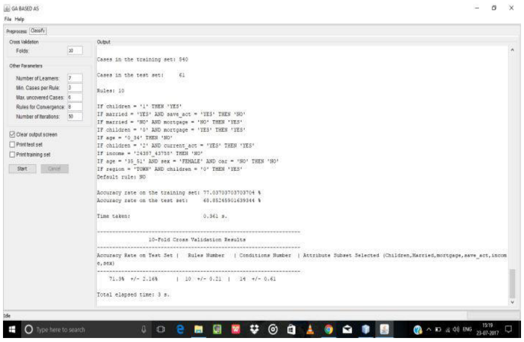
In figure 2, the first parameter is the number of learners. In this tool, 5 numbers of learners are taken and their accuracy compared to show best among them. "Forward Selection Multicross Validation, Bootstrap Backward Elimination, Relief, MIFS, and GABASS" are some of the techniques used in this study.

The four other parameters are "Minimum Cases per Rule," "Maximum Uncovered Cases," "Number of Rules for Convergence," and "Number of Iterations" for genetic algorithm optimization. Chromosomes that aregenerated using a genetic algorithm consist of 14 bits, out ofwhich the first two bits represent numbers of iterations, the nextthree bits represent minimum cases per rule, the next three bits represent maximum uncovered cases and the last four bits represents numbers rule convergence [5].

For example, Chromosomes samples: 010010110100
represents-
Number of Iterations- 200
Minimum Cases per Rule- 3
Maximum Uncovered Cases- 11
Number of Rule for convergence- 6

C. Genetic Algorithm

The genetic search begins with a population with zero characteristics and randomly created rules. The notion of survival of the fittest is used to generate a new population that will follow the laws of the fittest in the current population, as well as their children. The genetic operators cross over and mutations are used to create offspring. The generation process continues until a population P emerges, with every rule satisfying the fitness criterion [1]. By clicking on the start button below the parameters in figure 2, the aforesaid conditions are satisfied by the implemented tool. The output is shown in figure 3 which contains the number of rules as follow-



Cross Validation
Folds: 10

Other Parameters
Number of Learners: 7
Min. Cases per Rule: 3
Max. uncovered Cases: 6
Rules for Convergence: 8
Number of Iterations: 50

☑ Clear output screen
□ Print test set
□ Print training set

```
Cases in the training set: 540

Cases in the test set:     61

Rules: 10

IF children = '1' THEN 'YES'
IF married = 'YES' AND save_act = 'YES' THEN 'NO'
IF married = 'NO' AND mortgage = 'NO' THEN 'YES'
IF children = '0' AND mortgage = 'YES' THEN 'YES'
IF age = '0_34' THEN 'NO'
IF children = '2' AND current_act = 'YES' THEN 'YES'
IF income = '24387_43758' THEN 'NO'
IF age = '35_51' AND sex = 'FEMALE' AND car = 'NO' THEN 'NO'
IF region = 'TOWN' AND children = '0' THEN 'YES'
Default rule: NO

Accuracy rate on the training set: 77.03703703703704 %
Accuracy rate on the test set:     68.85245901639344 %

Time taken:                        0.361 s.

------------------------------------------------------------------
                10-Fold Cross Validation Results
------------------------------------------------------------------
Accuracy Rate on Test Set |  Rules Number   | Conditions Number  | Attribute Subset Selected (Children,Married,mortgage,save_act,incom
e,sex)
------------------------------------------------------------------
    71.3%  +/- 2.16%      |  10  +/- 0.21  |  14  +/- 0.61

Total elapsed time: 3 s.
```

Figure 3 User Input Parameters and its Resulting Rules with Accuracy

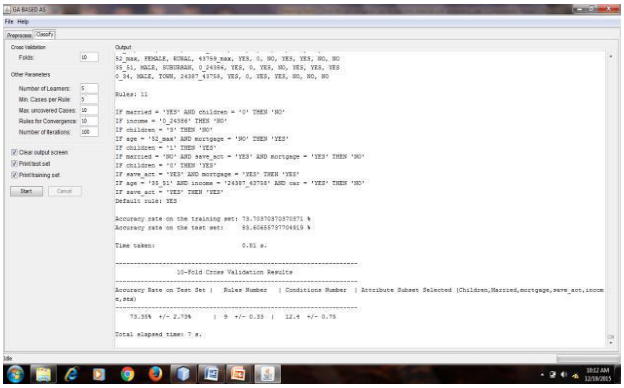
The roulette wheel selection is used to establish a new population over two chromosomes with a cross over the probability of 0.8 and a mutation probability of 0.001. Using this in a 10-fold cross validation until the requirements are met with the highest level of accuracy.

IV. RESULT

The best result is found by taking the value of k fold cross validation 10. When the parameters entered by the user are –

Number of Learners= 5

Minimum Cases per Rules= 5 Maximum Uncovered Cases=10 Rules for Convergence= 10 Number of Iterations= 100



Cross Validation
Folds: 10

Other Parameters
Number of Learners: 5
Min. Cases per Rule: 5
Max. uncovered Cases: 10
Rules for Convergence: 10
Number of Iterations: 100

☑ Clear output screen
☑ Print test set
☑ Print training set

Output
```
52_max, FEMALE, RURAL, 43759_max, YES, 0, NO, YES, YES, NO, NO
35_51, MALE, SUBURBAN, 0_24386, YES, 0, YES, NO, YES, YES, YES
0_34, MALE, TOWN, 24387_43758, YES, 0, YES, YES, NO, NO, NO

Rules: 11

IF married = 'YES' AND children = '0' THEN 'NO'
IF income = '0_24386' THEN 'NO'
IF children = '3' THEN 'NO'
IF age = '52_max' AND mortgage = 'NO' THEN 'YES'
IF children = '1' THEN 'YES'
IF married = 'NO' AND save_act = 'YES' AND mortgage = 'YES' THEN 'NO'
IF children = '0' THEN 'YES'
IF save_act = 'YES' AND mortgage = 'YES' THEN 'YES'
IF age = '35_51' AND income = '24387_43758' AND car = 'YES' THEN 'NO'
IF save_act = 'YES' THEN 'YES'
Default rule: YES

Accuracy rate on the training set: 73.7037037037071 %
Accuracy rate on the test set:     83.60655737704919 %

Time taken:                        0.81 s.

-------------------------------------------------------------------
                 10-Fold Cross Validation Results
-------------------------------------------------------------------
Accuracy Rate on Test Set |   Rules Number   | Conditions Number  | Attribute Subset Selected (Children,Married,mortgage,save_act,incom
e,sex)
-------------------------------------------------------------------
    73.35%  +/- 2.73%      |  9  +/- 0.33  |   12.4  +/- 0.75

Total elapsed time: 7 s.
```

Figure 4 Selected subset of Attribute

Where the number of instances for the training set is 540 and for the testing set is 61. The accuracy observed at the bottom in figure 4 is 73.35% +/- 2.73% i.e., 76% approx. a subset of selected attributes is shown in the last of the above figure i.e., (Children, married, mortgage, save_acc, income, sex).

The graph in Fig 5 depicts the comparison between Random and GABASS. This comparison was made based on classification accuracy and GABASS was discovered to be superior to other methods.

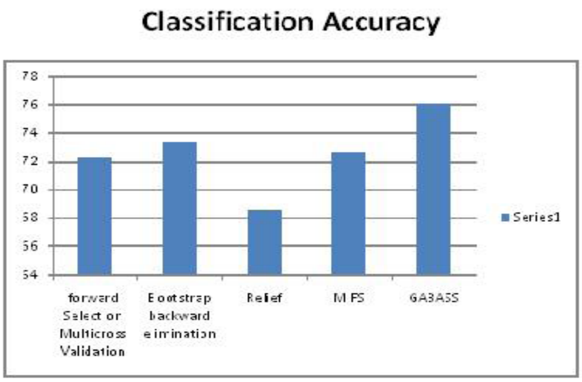


Figure 5 The Graph Shows the Comparison between Random and GABASS

[4] Ricco Rakotomalala, "TANAGRA: UN logicielgratuit pour l'enseignement et la recherche", in Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005.

[5] Data Mining: Concepts, Methodologies, Tools, and Applications, Volume 1 Management Association, Information Resources IGI Global, 30-Nov-2012.

ABOUT THE AUTHORS

Pradeep Gupta received his B.E. (CSE) in 2006, MTech (CSE) in 2011.He has 13 years of experience in teaching.He is currently employed as an Assistant Professor at Ghaziabad's Ajay Kumar Garg Engineering College. Artificial Intelligence, Machine Learning, Deep Learning, and Cyber Security are some of his research interests.

Jay Kant Pratap Singh Yadav completed his B. Tech. and M. Tech. (NIT, Surat) and currently working as an Assistant Professor at Ajay Kumar Garg Engineering College in Ghaziabad, Uttar Pradesh, in the Department of Computer Science and Engineering.He is a lifetime member of IAENG and other academic societies and published several research papers in reputed international journals and conferences. His area of interest is Machine Learning, Soft Computing, Digital Image Processing, Computer Vision.