Genetic Algorithm Based Bayesian Classification Algorithm for Object Oriented Data

Dr. Vipin Saxena
Department of Computer Science
B.B. Ambedkar University (A Central University)
Rae Barely Road, Lucknow-25, U.P. (India)

Ajay Pratap
Department of Computer Science
B.B. Ambedkar University (A Central University)
Rae Barely Road, Lucknow-25, U.P. (India)

Abstract— **Object-Oriented database System (OODBMS) is the fusion of two technologies which are: Database Management System and Object-Oriented System. OODBMS is available in the market as an option because the modern databases are large and complex. A real case study of Indian Postal Services is discussed to solve the problem of learning and object oriented data classification. Naïve Bayesian classifier is a classification algorithm based on Bayes theorem which is guided by genetic algorithm. Investment based survey is carried out and it is divided into five different classes. Proposed GA based classifier has predicted the grade of given test data on the basis of learning data. Implementation of proposed algorithm is also presented with the help of a case study.**

Keywords- OODBMS, Naïve Bayesian Classifier, Genetic Algorithm, Classificaton Framework, Training Object Set .

I. INTRODUCTION

The business organizations rely on digital data, which has become very large and complex these days. In spite of using traditional relational database system, these business organizations can use object oriented database system. Naïve bayesian classifiers are the simple but effective statistical classifiers which can be applied to many complex domains. They present a probabilistic approach of classification which is based on Bayes Theorem. Posterior and prior probabilities of hypothesis and the object under discussion is calculated with the help of Bayes theorem.

The $P(C_i \mid X)$ is calculated using the Bayes Theorem.

$$P(C_i \mid X) = \frac{P(X \mid C_i)\, P(C_i)}{\sum_i P(C_i \mid X)\, P(C_i)}$$

After the calculation of P(Ci | X) following check is performed:

$$P(C_i \mid X) > P(C_j \mid X) \quad \text{for all } i \neq j$$

Among all available classes the above check is carried out and the class having high probability is selected to have the proposed object.

Genetic algorithms are being accepted as a method for various optimization problems. Genetic algorithm based Bayesian classification is proposed in this research paper in order to improve the classification accuracy and minimize the response time. Genetic algorithm (GA) is an abstraction of biological evolution and its theoretical framework is adapted as an algorithm.

II. RELATED WORK

Various research papers related to object oriented database, naïve Bayesian classification and genetic algorithms are explored. Bauer [1] has proposed genetic algorithms and investment strategies as most powerful weapon. The speed, power and flexibility of genetic algorithms help to get winning investment strategies. In the article, Domingos and Pazzani [2] have shown that the Bayesian classifier's probability estimates are only optimal under quadratic loss. The classifier itself can be optimal under zero-one loss if independence assumption holds even when this assumption is violated. Colorni et. al. [3] has presented the results of an investigation of the possibilities offered by genetic algorithms to solve the timetable problem. The outcomes of the utilization of the implemented system to the specific case of the generation of a school timetable are reported. Korczak and Roger [4], has presented a genetic algorithm based searching approach to find technical trading rules which gives buying and selling advices about individual stocks is proposed. The proposed approach is tested out of a sample of 24 French stocks among the most important stocks traded on the French market. In today's era, traders and investment analysts require faster response from financial marketplace. V´azquez and Whitley [5] have developed a variety of genetic algorithms for the static Job shop scheduling problem. They have implemented a hybrid GA, known as OBGT and this algorithm can solve the static job shop scheduling problem. A hybrid algorithm using a novel classifier based on the Bayes discriminant function is presented by Raymer et al. [6]. Chickering [7] described a convenient graphical representation for an equivalence class of structures. A set of operators which can be applied for the convenient graphical representation by search algorithm is also introduced. The proposed equivalence-class operators can score locally and share the computational efficiency of traditional operators defined for individual structures. The proposed algorithm employs feature selection and extraction to isolate salient features from large biological data sets. The effectiveness of

proposed algorithm is applied and demonstrated on various biological and medical data sets. Bernardo [8] has presented the concept of Bayesian statistics. Bayesian methods provide a complete paradigm for both statistical inference and decision making under uncertainty.

Margaritis [9] has addressed the important problem of the determination of the structure of directed statistical models. Bayesian network models are used to solve the problem. In this paper, Fenton and Walsh [10] presented the results of a fair comparison between messy genetic algorithms. A permutation based simple GA is applied to a job shop scheduling system. Whitley [11] described and to analyzed genetic algorithms and other evolution strategies in sufficient detail. Few model problems and case studies are discussed. Main focus is on performance and implementation related issues. Park and Cho [12] have discussed about the bayesian networks, which provide a robust formalism for probabilistic modeling. They have proposed an evolutionary optimization of attribute ordering in bayesian network to handle the problem using a genetic algorithm with medical knowledge. Wiggins et al. [13] has extracted the information from the electro-cardiograms (ECGs) of patients and presented a methodology to classify the patients on the basis of age. The evolved Bayesian network performed better than the greedy algorithm based classification. Phomasakha et. al. [14] has proposed a decision tree-based model for automatic assignments of IT service desk outsourcing used in banking business. The proposed model has made contribution of data preparation procedures which is proposed for text mining discovery algorithms. In [15], introduction of naive bayes classifier is given which is a simple probabilistic classifier based on applying Bayes' theorem. Naive Bayes classifiers can be trained very efficiently through supervised learning.

III. Problem Statement

Indian post offices are selling various non-banking products such NSC (National Saving Certificate) and KVP (Kisan Vikas Pattra). We have been studying the purchase pattern of these schemes in various locations of India based on different factors. We are trying to answer the business questions like:

- Can we classify the customers in some special category?
- Can we predict the category of given customer on basis of given training object?

IV. Investment Based Survey

A survey is carried out in different towns and cities of Uttar Pradesh to know the amount of investment they do with financial plans of Indian Postal Services. The initial criteria starts from Investment <= 5000 and at the interval of 5000, maximum investment is 25000. The survey is carried out among 100 customers and the statistics is presented with the help of grading table having columns as Serial Number, Grade, Criteria, Number of Customers and Percentage Value. Table 1 represents the grading table and Figure 1presentes the pictorial analysis of survey, are shown below:

TABLE 1: GRADING TABLE

SN	Grade	Criteria	Number	%
1	GRADE-1	Invest<= 5000	28	9.34
2	GRADE -2	5000< Invest<=10000	30	10
3	GRADE -3	10000< Invest<=15000	44	14.66
4	GRADE -4	15000< Invest<=20000	22	7.33
5	GRADE -5	20000< Invest<=25000	36	12

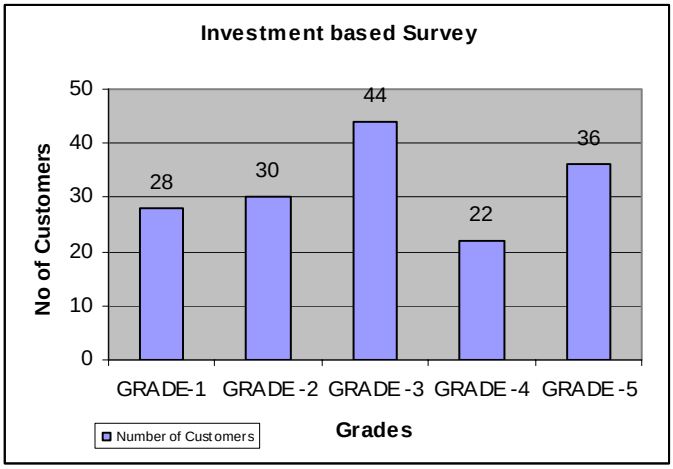

Figure 1: Investment Based Survey

V. Proposed Framework For Data Classification

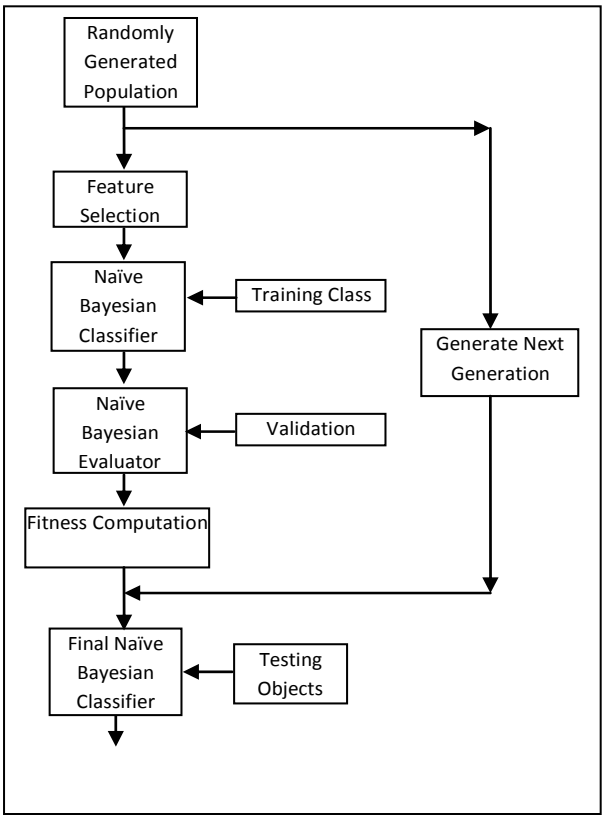

Figure 2: GA based Naïve Bayesian Classifier

A framework for GA based naïve bayesian classification for the object oriented data is presnted in this section. Genetic algorithm is an abstraction of biological evolution whose theoretical framework is adopted. Genetic algorithm is a method for moving from one population of chromosomes to another new population. Here chromosomes are the bit strings representing organisms or candidate solutions to any problem and consists of genes. The selection process chooses the chromosomes which can reproduce. They use selection together with the genetic operators of crossover and mutation.

Here GA Tool Box is being used to optimize the classification performance and objective is to find the population of best weight to minimize the classification error.

Rules of form: If A →C

Where A → Antecedent

C → Consequent

Here, A= {A1, A2, A3, ……..,An}

Let attribute are :-

Location , Age, Income , Source of income.

There values are:-

Location= {Village, Town, Metro}

Age = { Youth , Middle ,Senior}

Source of income= {Agriculture, Govt, Private , Business }

Income= {High, Medium, Low}

Hence Representation of different attributes values, in case of binary chromates are as follows:-

TABLE 2: BINARY GENE REPRESENTATON FOR LOCATON

Location	Bit-1	Bit-2	Bit-3
Village	1	0	0
Town	0	1	0
Metro	0	0	1

TABLE 3: BINARY GENE REPRESENTATON FOR AGE

Age	Bit-1	Bit-2	Bit-3
Youth	1	0	0
Middle	0	1	0
Senior	0	0	1

TABLE 4: BINARY GENE REPRESENTATON FOR INCOME

Income	Bit-1	Bit-2	Bit-3
High	1	0	0
Medium	0	1	0
Low	0	0	1

TABLE 5: BINARY GENE REPRESENTATON FOR SOURCE OF INCOME

Source of income	Bit-1	Bit-2	Bit-3	Bit-4
Agriculture	1	0	0	0
Govt.	0	1	0	0
Private	0	0	1	0
Business	0	0	0	1

Selection:

Here, individuals training data is given a probability which is calculated by following.

Let Xi be the training data then,

(Xi)=f(Xi)/m

Here f(xi) fitness of individual training data set Xi .

F(Xi) is probability of the data Xi for being selected.

Crossover:

Crossover operation is performed on the pairs which are selected .This function selects genes from parent chromosomes and create now off springs .The single point crossover is performed.

Crossover computation

TABLE 6: GRADING TABLE

Chromosomes 1	1001| 001001000
Chromosomes 2	0010| 101000001
Off springs 1	1001| 101000001
Off springs 2	0010| 001001000

Mutation:

Mutation is another genetic operator, which is applied to new chromosomes. Mutation causes the individual genetic representation to be changes according to rule. A random number is chosen for mutation using random () function.

Fitness Function:

The rules discovered for classification must have high confidence and high completeness factor. A fitness function is to defined as a combination of these two and represented as:-

Fitness= $(\text{confidence})^n$* completeness

VI. PROPOSED ALGORITHM FOR CLASSIFICATION OF OBJECT-ORIENTED DATA

Input:

1. Class under discussion
2. Object O, which is represented by a n-dimensional vector, O= {V_1, V_2, V_3, ……..……, V_n},
   where V is the value from n attributes, A_1, A_2, A_3, ……..……, A_n

Assumption:

1. Let D be the training set of objects and D = {O_1, O_2, O_3, ……..……,O_n}

2. Let H be a hypothesis such that object O belongs to any specified category of grade.
3. Let the number of categories (C) in which the objects can be classified are m which are
   $C = \{C_1, C_2, C_3, \ldots\ldots\ldots, C_m\}$ and they are independent.

Algorithm:

Step-1: Start
Step-2: Calculate the posterior probability of each category with the help of Bayes' theorem.

$$P(H/O) = P(O/H).\ P(H) / P(O)$$

Here,

P (H/O) [Posterior Probability of H] is the probability that object O of class customer will belong to any grade (GRADE-X) given that we know the location, age, income and source of income of object.

P (O/H) [Posterior Probability of O] is the probability that object O has certain location, age, income and source of income and we know that it belongs to certain group.

P (H) [Prior Probability of H] is the probability that the given object will belong to certain grade category, regardless of any given information.

P (O) [Prior Probability of O] is the probability that an object from class customer fulfils the certain values of attributes.

Step-3: Calculate the posterior probability of object O {all categories are independent}

$$P(O/H) = \Pi\ P(V_k/H)$$
$$= P(V_1/H) \times P(V_2/H) \times \ldots\ldots P(V_n/H)$$

Step-4: Calculate and compare the multiplication of P (O/H) and P (H) for all possible categories

Step-5: The object under discussion belongs to the class having maximum P (O/H) x P (H)

Step-6: Stop

VII. Implementation: A Case Study of Indian Postal System

TABLE 7: TRAINING OBJECTS FROM CUSTOMER CLASS

OID	LOCATION	AGE	INCOME	SOURCE	CLASS: GRADE
1	Village	Youth	High	Agriculture	GRADE-5
2	Town	Youth	Medium	Government	GRADE-2
3	Town	Middle	Medium	Government	GRADE-4
4	City	Youth	Low	Private	GRADE-1
5	Village	Middle	High	Business	GRADE-4
6	City	Senior	Medium	Private	GRADE-3
7	Metro	Youth	Low	Private	GRADE-1
8	Town	Youth	High	Business	GRADE-5
9	Metro	Senior	Medium	Private	GRADE-2
10	City	Middle	Medium	Agriculture	GRADE-3
11	Village	Youth	Low	Government	GRADE-2
12	Village	Senior	High	Agriculture	GRADE-4
13	Town	Middle	Low	Agriculture	GRADE-2
14	City	Senior	Medium	Business	GRADE-3
15	Metro	Middle	High	Business	GRADE-5

We wish to classify the object O = {LOCATION=Town, AGE= Middle, INCOME= Medium, SOURCE OF INCOME= Agriculture}
All objects of the class customer are categorized in five grades.
P (GRADE=1) = 2/15 = 0.133
P (GRADE=2) = 4/15 = 0.266
P (GRADE=3) = 3/15= 0.20
P (GRADE=4) = 3/15= 0.20
P (GRADE=5) = 3/15= 0.20
P (LOCATION= Town/GRADE=1) = 0/2
P (AGE= Middle/GRADE=1) = 0/2
P (LOCATION= Town/GRADE=2) = 2/4
P (AGE= Middle/GRADE=2) = 1/4
P (LOCATION= Town/GRADE=3) = 0/3
P (AGE= Middle/GRADE=3) = 1/3
P (LOCATION= Town/GRADE=4) = 1/3
P (AGE= Middle/GRADE=4) = 2/3
P (LOCATION= Town/GRADE=5) = 1/3
P (AGE= Middle/GRADE=5) = 1/3

P (INCOME= Medium/GRADE=1) = 0/2
P (SOURCE OF INCOME= Agriculture/GRADE=1) = 0/2
P (INCOME= Medium/GRADE=2) = 2/4
P (SOURCE OF INCOME= Agriculture/GRADE=2) = 1/4
P (INCOME= Medium/GRADE=3) = 3/3
P (SOURCE OF INCOME= Agriculture/GRADE=3) = 1/3
P (INCOME= Medium/GRADE=4) = 1/3
P (SOURCE OF INCOME= Agriculture/GRADE=4) = 1/3
P (INCOME= Medium/GRADE=5) = 0/3
P (SOURCE OF INCOME= Agriculture/GRADE=5) = 1/3

Obtaining P (O/H) = Π P (Vk/H) by using the above probabilities:

P (O/GRADE=1) = P (LOCATION= Town/GRADE=1) x P (AGE= Middle/GRADE=1) x P (INCOME= Medium/GRADE=1) x P (SOURCE OF INCOME= Agriculture/GRADE=1)
= 0/2 x 0/2 x 0/2 x 0/2
= 0

P (O/GRADE=2) = P (LOCATION= Town/GRADE=2) x P (AGE= Middle/GRADE=2) x P (INCOME= Medium/GRADE=2) x P (SOURCE OF INCOME= Agriculture/GRADE=2)
= 2/4 x 1/4 x 2/4 x 1/4

= 0.0156

P (O/GRADE=3) = P (LOCATION= Town/GRADE=3) x P (AGE= Middle/GRADE=3) x P (INCOME= Medium/GRADE=3) x P (SOURCE OF INCOME= Agriculture/GRADE=3)
= 0/3 x 1/3 x 3/3 x 1/3
= 0

P (O/GRADE=4) = P (LOCATION= Town/GRADE=4) x P (AGE= Middle/GRADE=4) x P (INCOME= Medium/GRADE=4) x P (SOURCE OF INCOME= Agriculture/GRADE=4)
= 1/3 x 2/3 x 1/3 x 1/3
= 0.0237

P (O/GRADE=5) = P (LOCATION= Town/GRADE=5) x P (AGE= Middle/GRADE=5) x P (INCOME= Medium/GRADE=5) x P (SOURCE OF INCOME= Agriculture/GRADE=5)
= 1/3 x 1/3 x 0/3 x 1/3
= 0

Applying Laplacian correction

Laplacian correction is applied for P (O/GRADE=1), P (O/GRADE=3) and P (O/GRADE=5) by assuming that if number of object used for the training are 5000 and at least one object for LOCATION= Town is present for GRADE=1 and GRADE=3, at least one object for AGE= Middle, INCOME= Medium and SOURCE OF INCOME= Agriculture is present for GRADE=1 and at least one object for INCOME= Medium is present for GRADE=5.
Corrected probabilities will be 1/5004 = 0.0001

P (O/GRADE=1) = 0.0001 x 0.0001 x 0.0001 x 0.0001
= 0.0000000000001
P (O/GRADE=3) = 0.0001 x 0.33 x 1 x 0.33
= 0.00001089
P (O/GRADE=5) = 0.33 x 0.33 x 0.0001 x 0.33
= 0.0000035937

P (O) is constant and now we need to compute P (O/H) P (H),
P (O/ GRADE=1) x P (GRADE=1) = 0.0000000000001 x 0.133 = 0.000000000000133

P (O/ GRADE=2) x P (GRADE=2) = 0.0156 x 0.266 = 0.0041496

P (O/ GRADE=3) x P (GRADE=3) = 0.00001089 x 0.20 = 0.000002178

P (O/ GRADE=4) x P (GRADE=4) = 0.0237 x 0.20 = 0.00474

P (O/ GRADE=5) x P (GRADE=5) = 0.0000035937 x 0.20 = 0.0000007187

Among all values of P (O/H) P (H), the value of P (O/ GRADE=4) x P (GRADE=4) is highest. So, the naïve Bayesian classifier predicts that the given object O will belong to GRADE=4.

VIII. CONCLUSION

IX. ACKNOWLEDGEMENT

The authors are thankful to Prof. B. Hanumaiah, Vice-Chancellor, B.B. Ambedkar University (A Central University) Lucknow for providing the excellent facility in the computing lab of B. B. Ambedkar University, Lucknow, India. I am also thankful to Mr. Martin Gogolla and Mr. Mark Richters for his support to understand the concepts used in the design and implementation of our project.

REFERENCES

[1] R.J. Bauer, Genetic Algorithms and Investment Strategies, Wiley, 1994.

[2] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," Machine Learning, Vol. 29, No. 2-3, 1997, pp. 103-130. doi:10.1023/A:1007413511361

[3] A. Colorni, M. Dorigo and V. Maniezzo, "A Genetic Algorithm To Solve The Timetable Problem," Centre for Emergent Computing, Napier University, Edinburgh EH10 5DT, UK,2000.

[4] J. Korczak, P. Roger, "Stock Timing using Genetic Algorithms," Res. Rep. ULP, LSIIT, Illkirch, No 01/2000

[5] M. V´azquez and D. Whitley, "A comparison of genetic algorithms for the static job shop scheduling problem," Proc. 6th Parallel Problem Solving from Nature – PPSN VI, pp.303–312, 2000.

[6] M. L. Raymer, L. A. Kuhn, and W. F. Punch, "Knowledge discovery in biological datasets using a hybrid bayes classifier/evolutionary algorithm," In Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, pages 236–245, Washington, DC, USA, 2001. IEEE Computer Society.

[7] D. M. Chickering, "Learning Equivalence Classes of Ba- yesian-Network Structures," Journal of Machine Learning Research, Vol. 2, 2002, pp. 507-554.

[8] J. M. Bernardo, "Bayesian Statistics", Encyclopedia of Life Support Systems, Probability and Statistics, Oxford, UK, 2003.

[9] D. Margaritis, "Learning Bayesian Network Model Structure," Technical Report CMU-CS-03-153, 2003.

[10] P. Fenton and P. Walsh: "A comparison of messy GA and permutation based GA for job shop scheduling," Genetic And Evolutionary Computation Conference 2005, pp.1593–1594, 2005.

[11] D. Whitley, Artificial Intelligence: Genetic Algorithms and Evolutionary Computing, Van Nostrand's Scientific Encyclopedia, Wiley, 2005.

[12] H.S. Park and S.B. Cho, "An efficient attribute ordering optimization in bayesian networks for prognostic modeling of the metabolic syndrome," In ICIC (3), pages 381–391, 2006.

[13] M. Wiggins, A. Saad, B. Litt, and G. Vachtsevanos, "Evolving a bayesian classifier for ECG-based age classification in medical applications," Applied Soft Computing, 8(1):599 – 608, 2008.

AUTHORS PROFILE

[14] P. Phomasakha, N. Sakolnakorn and P. Meesad, "Decision Tree-Based Model for Automatic Assignment of IT Service Desk Outsourcing in Banking Business," Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, IEEE Computer Society, 2008, pp. 387–392.

Dr. Vipin Saxena is a Professor & Head, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India. He got his M.Phil. Degree in Computer Application in 1991 & Ph.D. Degree work on Scientific Computing from University of Roorkee (renamed as Indian Institute of Technology, Roorkee, India) in 1997. He has more than 16 years teaching experience and 19 years research experience in the field of Scientific Computing & Software Engineering. Currently he is proposing various software designs by the use of Unified Modeling Language for the research problems related to the Software Domains & Advanced Computer Architecture. He has published more than 91 International and National research papers in various refereed Journals and authored four books covering Software Engineering, E-Learning and Operating System. Dr. Saxena is a life time member of Indian Congress.
Mobile: 0091-9452372550
E-mail: vsax1@rediffmail.com

Ajay Pratap is a research student in Department of Computer Science, Babashaheb Bhimrao Ambedkar University, Lucknow, India. Earlier, he got his M.Phil. Degree in Computer Science in 2008 and Master of Computer Application from the said University & presently he is working on performance estimation of object oriented database through UML.
Mobile: 0091-9936-189152
E-mail: pratap_aj@yahoo.co.in

[15] Wikipedia, "Naive Bayesian classification," http://en.wikipedia.org/wiki/ Naive_Bayesian_classification. (Accessed on)