UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

Business Intelligence Appraisal of Augmented Data Based on Existing Customers' Dataset Obtained by Genetic Algorithm using Multiple Correlation Technique

Nethravathi P. S[1], K. Karibasappa[2]

Department of Master of Computer Applications, Shree Devi Institute of Technology, Mangaluru, Karnataka, India[1]

Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bengaluru, Karnataka, India[2]

Abstract: The major issues in the business intelligence is to predict the customer behavior and understanding them for the betterment of the business. This understanding and behavior are to be analyzed on the basis of the information given by the customer. It is important to determine customer preferences in formulating market strategies that is taken from the concept of Business Intelligence (BI) and analysis. This analysis is based on the customer profile. For these kinds of analysis, a large number of datasets are required. Having analyzed the data set in the paper [1], it was realized that more number of dataset will give maximum precession. To get these kinds of data set formulated for the work, is very difficult. This is the reason, the work has been concluded to go for genetic algorithm to generate new groups of datasets to augment the existing one, with a proper validation. Also, it has been decided to give more priority for the correlation by effectively using multiple correlation, to get an appropriate result in this work.

Keywords: Business Intelligence (BI), Multiple Correlation Technique, Augmented Data, Genetic Algorithm.

1. INTRODUCTION

The aim of this paper is to predict the customer purchase pattern based on their credentials provided by the customer. Its main objective is the establishment of a relationship to meet customer needs. When a customer visits a particular place based on his / her credentials one need to see what type of customer would be interested in what type of products. Then extracting the opinion of the customer about the product they have already purchased and using it. This information is needed for innovation in the business intelligence. This information also leads to achieve a higher degree of productivity in the company as per the requirements. To process with these kinds of work the main requirement is an abounded dataset. The kind of dataset working on, in this work is distinct from the commonly available groups of data, hence to get the sets of these types, it is difficult, and hence, it is preferable, if possible to generate these sets, from the available sets. Also, the generated set should satisfy the properties of the available sets. To get these categories of sets it is decided to enhance the groups of data.

Moheb R. Girgis [1] presents an automatic test data generation technique using Genetic Algorithm guided by the data flow dependencies in the program to search for test data. To fulfill the all-uses criterion, it accepts population size, maximum number of generations, and probabilities of the crossover and mutation as a parameter. Experiments have been carried out to evaluate the effectiveness of the proposed GA. Two techniques are applied and compared. In the first one a new random testing and selection of parent's technique where is made randomly, so that every effective member of the current population has an equal chance of being selected for recombination.

And the other roulette wheel method [D.E. Goldberg] [2] where the selection of a new population is done with respect to probability distribution based on fitness values. The results of these experiments showed that the proposed random testing technique method produced better results than the roulette wheel method.

M. Anbarasi et. al. [3] attempts to predict the presence of heart disease with reduced number of attributes using Genetic Algorithm. The algorithm determines the attribute contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. Naive Bayes, clustering classification and decision tree classifiers are used to predict the diagnosis of patients. The accuracy is measured before and after reduction of number of attributes. The observations exhibit that the decision tree outperforms other two data mining techniques after incorporating feature subset selection with relatively high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering performs poor compared to other two methods.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

Amit Kumar Sharma [4] proposes a GA-based software test data generator to demonstrate its feasibility. GAs show good results in searching the input domain for the required test sets. Genetic Algorithms may not be the answer to the approach of software testing, but do provide an effective strategy.
Nuwan I. Senaratna [5] is one of the most popular evolutionary-based algorithms. It has been successfully applied to numerous problems. Genetic algorithm is a search method utilizing the principles of natural selection and genetics. Genetic algorithm generates a sequence of populations by using a selection mechanism, and use crossover and mutation as search mechanisms.
Silvia TRIF [6] demonstrates the use of genetic algorithms for training neural networks used in secured Business Intelligence Mobile Applications. He assesses the use of genetic algorithm by the comparison between classic back-propagation method and a genetic algorithm based training. A comparative study is realized for determining the better way of training neural networks, from the point of view of time and memory usage. His study reveals that genetic algorithms are a solution that can be used on mobile devices to solve optimization problems like training a neural network. The obtained solutions are good and the resources used to obtain the solution are reasonable compared to classic training methods.
Nidhi Bhatla and Kiran Jyoti [7] aims at analyzing the various data mining techniques for heart disease prediction. Various techniques and data mining classifiers are defined for efficient and effective heart disease diagnosis. The analysis shows that Neural Network with 15 attributes has shown the highest accuracy compared to Decision Tree and Genetic Algorithm.
Nethravathi P. S, K. Karibasappa [8] demonstrates an approach to cluster the records on the basis of correlation index. This paper tried to inter-relate the hobby and other important credentials of the customers which plays an important role in purchasing the products. The experiment summarizes that the purchase behavior of a person is related to hobby to the real-world purchase pattern. The result and analysis shows the highlights of the results obtained by the clustering of records on the basis of correlation. Also indicates how hobby is directly related to the purchase pattern and the satisfaction levels. The results of purchase of an item related to hobby and satisfaction level, few records are tested with the results of correlations. It is also observed that there is satisfaction level and the advocacy level (recommendations) plays an important role in purchasing a product.
It is inferred from the earlier work [8] that, more number of dataset, it is possible to get more accurate mapping between the entities of the record set and the recommendation of a person's purchase pattern and the satisfaction levels can be improved. Collecting a very large amount of such data manually is a very cumbersome process. Hence it is decided in the earlier work (9) to enrich the dataset by using Genetic Algorithm.

Nethravathi P. S, K. Karibasappa [9] focuses on augmentation of the customer dataset using Genetic Algorithm. The dataset consists of the different factors inherent in each situation of the customer to understand the market strategy. Every new record is generated by picking up two random records from the existing dataset. This final new record is evaluated through a validation procedure to validate its authenticity to make it look like a real dataset. The validation procedure involves checking the new record parameters for their valid datasets as well as checking their valid combination. From the result and analysis of the paper [9], it is observed that, the generation of the new record is possible by picking up two random records from the existing dataset. However, from the result of multiple attribute crossovers and mutation will lead to a good combination for the generation of the new genome.
With the data augmentation approach, a large number of datasets are enhanced. It is also observed from previous work [8] that there is a scope to improve the relation between the satisfaction levels to the advocacy level in purchasing a product. For the best and accurate results, it has been proposed in this work to use multiple correlation technique with the enhanced datasets using genetic algorithm.

The rest of this paper is organized as follows: Section 1 already shows, introduced the work with a brief survey of related work on genetic algorithm and new generation, section 2 explains the concepts and the methodology based on multiple correlation. Result and Analysis are presented on the live dataset in the section 3. The conclusion of this paper is reported in section 4.

2. METHODOLOGY

The principle behind Genetic algorithm is that they create and maintain a population of individuals represented by chromosomes (essentially a character string analogous to the chromosomes appearing in DNA). These chromosomes are typically encoded solutions to a problem. The chromosomes then undergo a process of evolution according to rules of selection, mutation and crossover. Each individual in the environment (represented by a chromosome) receives a measure of its fitness in the environment. Reproduction selects individuals with high fitness values in the population, and through crossover and mutation of such individuals, a new population is derived in which individuals may be even better fitted to their environment. The process of crossover involves two chromosomes swapping chunks of data. Mutation introduces slight changes into a small proportion of the population and is representative.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

For correlation analysis, the dataset used in the work [8] is too meager, for the analysis. This is enhanced with the help of GA for further improvements and for the best and accurate results it has been proposed in this work to go with large dataset.

After getting sufficiently enhanced data sets generated by genetic algorithm using the original datasets these datasets are taken further for the analysis. However, it has been observed from [8] that the clustering of the data sets of 10% of highest ranking taken for the further process. This 10% ranking will not predict for the data set which falls in the categories of below 10%. This problem can be addressed by a technique called multiple regression method, which is elaborated in the next section.

The relationship between two variables and also the prediction of one from the other is estimated by considering the ratio of the covariance between two variables, has been seen in [9]. This solution may be generalized and extended to the problem to predict a single variable from the weighted linear sum of multiple variables (multiple regression) or to measure the strength of this relationship (multiple correlation) [10]. In statistics, the coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables.
Many applications of regression analysis involve situations in which there are more than one regresses or variable. A regression model that contains more than one regress or variable is called a multiple regression model.

$$\text{Let } D_s = (n - R) \quad (1).$$

In the whole data set rejected / filtered data set in terms of percentage is as per equation (2)

$$R = \frac{D_s}{n} \times 100 \quad (2)$$

These set of Ds are sent to the next process called evaluation of correlations coefficient. Let a single data set be

$$D_{s_i} = x_{i1k}, x_{i2k}, x_{i3k}, x_{i4k} \dots x_{imk} \quad (3)$$

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).
In the above equation (3) x_{imk} is considered as dependent variable this is evaluated based on the other independent variables $x_{i1k}, x_{i2k}, x_{i3k}, x_{i4k} \dots x_{i(m-1)k}$
As per the multiple regression method the value of x_{imk} is evaluated by the following equation (4)

$$x_{imk} = \beta_0 + \beta_1 \times x_{i2k} + \beta_2 \times x_{i3k} + \beta_3 \times x_{i4k} + \cdots + \beta_{m-1} \times x_{i(m-1)k} + \varepsilon \quad (4)$$

Where ε, is the random error component and the coefficients $\beta_0, \beta_1, \beta_2, \beta_3 \dots \beta_{(m-1)}$ are also called as regression weights, β_0 is the y intercept and β_1 determines the contribution of the Independent variable x_{i2k}, β_j determines the contribution of the Independent variable x_{ijk} . These weights are calculated as follows:
Let us consider the data sets $D_{s_i} = x_{i1k}, x_{i2k}, x_{i3k}, x_{i4k} \dots x_{i(m)k}$ i from 1 to n as a matrix D_s. This matrix D_s has m columns and n rows. The value of x_{imk} is the independent variable, which will be evaluated on the basis of all other dependent variables $x_{i1k}, x_{i2k}, x_{i3k}, x_{i4k} \dots x_{i(m-1)k}$ i from 1 to n.
Let us consider the equation (4) as

$$x_m = D_s\beta + \varepsilon \quad (5)$$

The regression weights β can be estimated by the following equation (6)

$$[\beta] = \{[D_s{}^T] \times [D_s]\}^{-1} \times \{[D_s{}^T] \times [x_{imk}]\} \quad (6)$$

After evaluation of β matrix, the value of independent variable x_{imk} is evaluated for individual data set again let that be $\widehat{x_{imk}}$ the difference between x_{imk} and $\widehat{x_{imk}}$ gives the random error component.

In this paper, the average error is calculated for the evaluation of the new value of x_{imk} for the existing parameters are found. This processes of finding the error is repeated until the minimum average error. This minimum error is used in the equation (4) for the evaluation of a new dataset's x_{imk} .

Thus, adaptation of equation (4) will leads to a prediction of the new independent parameter on the basis of existing datasets.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

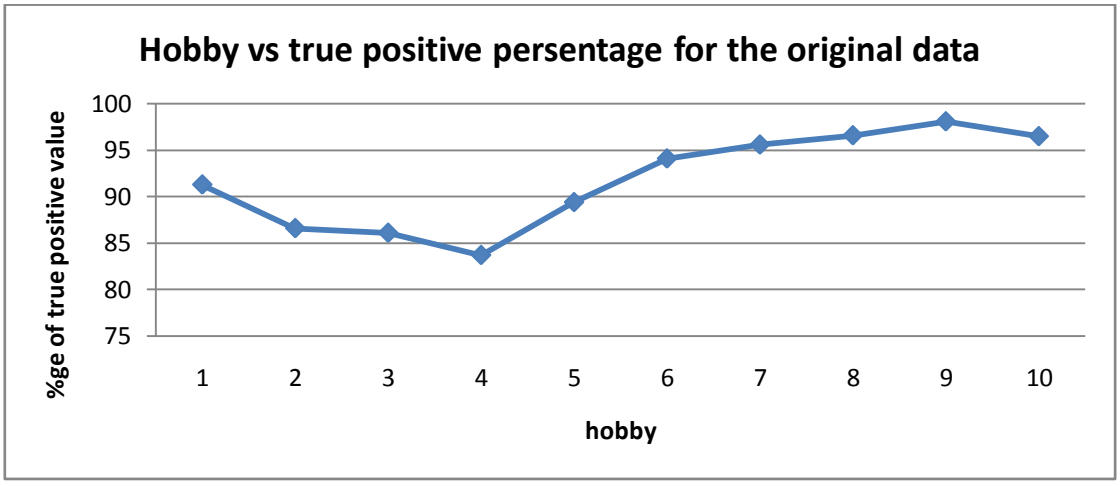
3. RESULT AND ANALYSIS

Since the aim of this research is to predict the customer purchase pattern based on their credentials provided by the customer. This indicates whether the customer is satisfied with the item he has purchased. Based on the above equation, this work tried to measure the satisfaction level. The below table 1 shows the results of the satisfaction level for various hobby considered. This result shows on an average the satisfaction level tallies with the existing one with a high percentage. Also, the figure 1 indicates the variation in the value of true positive for the prediction of happiness value. According to the figure 1, it varies from 83.7 to 98.1 this indicates the real true positive value is not below 83.7. This is done for the existing record set and also for the newly generated combined with the existing dataset.

Table 1: Evaluation of true positive values for different hobbies

Sl. No	1	2	3	4	5	6	7	8	9	10
Hobby	1	2	3	4	5	6	7	8	9	10
*TPER	91.3	86.6	86.1	83.7	89.4	94.1	95.6	96.6	98.1	96.5
**TPENR	83.33	90.6	97.1	94.10	92.95	91.07	89.23	90.2	84.26	87.83

*True positive ratio of satisfaction level for the existing records is abbreviated as TPER
** True positive ratio of satisfaction level for the enhanced records are abbreviated as TPENR
The average true positive ratio crosses 91.8 % mark, considering 10 of the hobbies



. Fig 1. The variation of true positive values of prediction for 10 hobbies

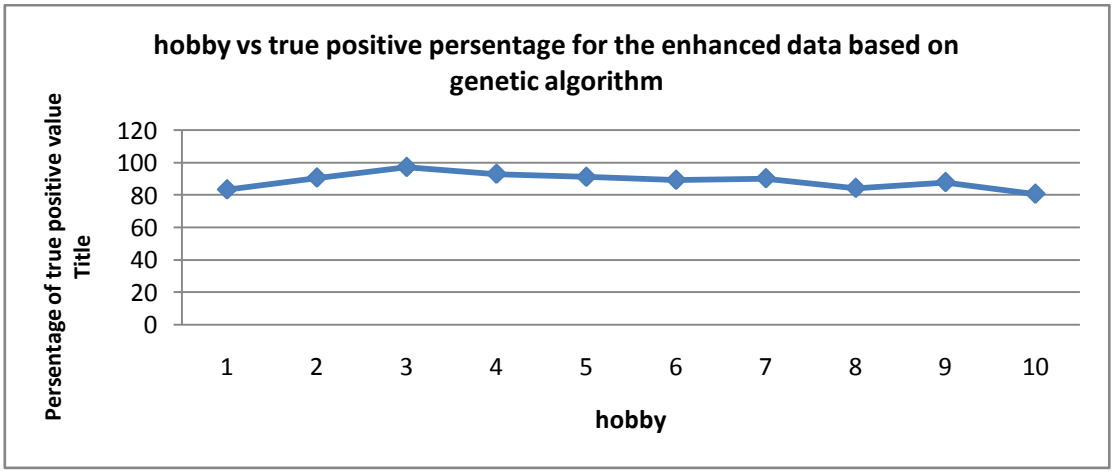


Fig 2. The variation of true positive values of prediction for 5 hobbies with extended data using genetic algorithm

Similarly, the other evaluations also can be processed.
This evaluation is based on the coefficient calculated based on multiple regression method. The error is calculated for a group of variables interested in. the error is evaluated on the basis of average error this average error is fed back to the system to evaluate the error again. This process is continued till the end of the process where the system takes deviation

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

from the downward error towards upwards indicating the minimum error in can be included for the system. The multiple iterations will give the values which is shown in the figure 3. This minimum error is evaluated and plotted for all the values considered. The following figure 3 shows the pattern for the minimum error for one set of data for the same hobby.

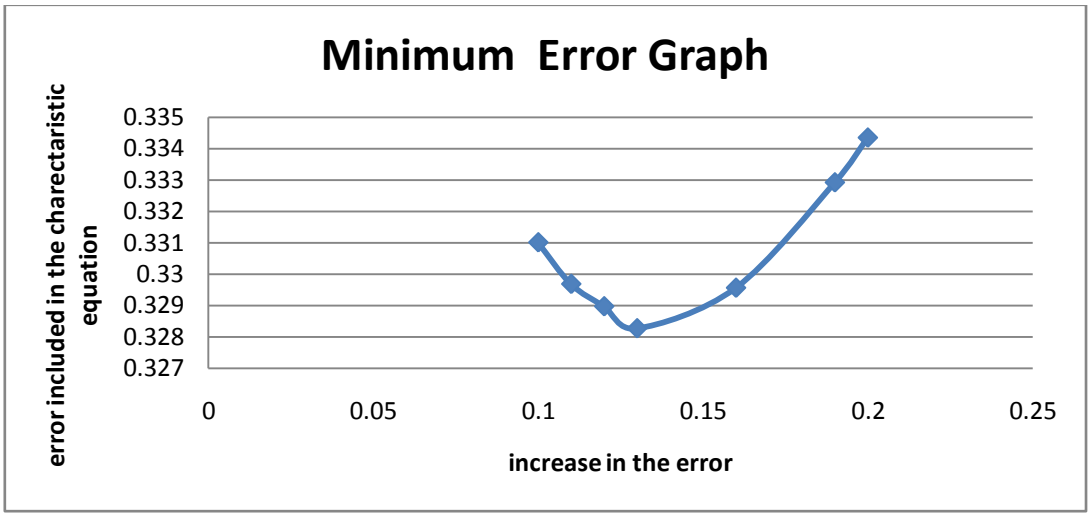


Figure 3. The evaluation of minimum error by multiple iterations in the equation of multiple regression method

4. CONCLUSION

It is concluded that by using the multiple regression method it is possible to evaluate the value of the satisfaction level to the accuracy level of around 89 to 90%. It is observed that the improvement in the data is evident after the enhanced data. This enhancement of the dataset is done based on the rule and genetic algorithm. Compared to single correlation, multiple correlation method gives more accurate results. A more focused result can be achieved by increasing the volume of the data set. This can be obtained using Genetic algorithm and rule of association. This work focuses on the same. In this work, the total volume of the dataset is increased by around 30% from the original set. This is evident from the table no.1 that the result of the newly generated set is more focused than original one. This result obtained from the multiple correlation can still be fine-tuned by minimizing the error iteratively as illustrated in the figure 3 above.

REFERENCES

1. Moheb R. Girgis, "Automatic Test Data Generation for Data Flow Testing Using a Genetic Algorithm", Journal of Universal Computer Science, vol. 11, no. 6 (2005), 898-915.
2. D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," Machine learning, vol. 3, no. 2, pp. 95–99, 1988.
3. M. Anbarasi et. al., "Enhanced Prediction of Heart Disease with Feature Selection Method by using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), 2010, 5370-5376.
4. Amit Kumar Sharma, "Optimized Test Case Generation Using Genetic Algorithm", International Journal of Computing and Business Research, Volume 4 Issue 3, NIMS University, Jaipur, 2013.
5. Nuwan I. Senaratna, "Genetic Algorithms: The Crossover-Mutation Debate", A literature survey submitted in partial fulfilment of the requirements for the Degree of Bachelor of Computer Science(Special) of the University of Colombo, November 15, 2005.
6. Silvia TRIF, "Using Genetic Algorithms in Secured Business Intelligence Mobile Applications", Informatica Economica, Volume 15, Romania.2011.
7. Nidhi Bhatla Kiran Jyoti, "International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October 2012.
8. Nethravathi P. S, K. Karibasappa, "Business Intelligence Appraisal of the Customer Dataset Based on Weighted Correlation Index", International Journal of Emerging Technology and Research2016.
9. Nethravathi P. S, K. Karibasappa, "Augmentation of the customer's profile dataset using Genetic Algorithm", International Journal of Research and Scientific Innovation (IJRSI) | Volume IV, Issue VIS, June 2017.
10. A.K. Sharma, "Text Book of Correlations and Regression", Discovery Publishing House, 2005.
11. B. Ghilic, M.Stoica and M. Mircea, "How to Succeed in Business Intelligence Initiative: A Case Study for Acquisitions in Romania Public Institutions", in Proc. WSEAS TRANSACTIONS on BUSINESS and ECONOMICS, Issue 6, Vol. 5/2008, pp. 298-309.
12. Asha Rajkumar and Mrs. G.Sophia Reena, "Diagnosis of Heart Disease Using Data mining Algorithm", GJCST, Vol. 10 Issue 10 Ver. 1.0 Sep 2010, pp. 38-43.
13. Catherine A. Peters, "Statistics for Analysis of Experimental Data", Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544.
14. H. D. Mills, M. D. Dyer, R. C. Linger, Cleanroom software engineering, IEEE Software, 4 (5), 19-25, 1987.
15. Grefenstette, J.J. and Baker J.E."How Genetic Algorithms Work: A Critical Look at Implicit Parallelism", In Schaffer, J.D. Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo, CA, 1989, pp. 20–27.