Augmentation of Customer's Profile Dataset Using Genetic Algorithm

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

[1]Department of Master of Computer Applications, Shree Devi Institute of Technology, Mangaluru, Karnataka, India
[2]Department of Computer Science and Engineering, Dayanand Sagar College of Engineering, Bengaluru, Karnataka, India

***Abstract:* - Data is the lifeblood of all type of business. Clean, accurate and complete data is the prerequisite for the decision-making in business process. Data is one of the most valuable assets for any organization. It is immensely important that the business focus on the quality of their data as it can help in increasing the business performance by improving efficiencies, streamlining operations and consolidating data sources. Good quality data helps to improve and simplify processes, eliminate time-consuming rework and externally to enhance a user's experience, further translating it to significant financial and operational benefits [1] [2]. All organizations/ businesses strive to retain their existing customers and gain new ones. Accurate data enables the business to improve the customer experience. Data augmentation adds value to base data by enhancing information derived from the existing source. Data augmentation can help reduce the manual intervention required to develop meaningful information and insight of business data, as well as significantly enhance data quality. Hence the business can provide unique customer experience and deliver above and beyond their expectations. The Data Augmentation is immensely important as it helps in improving the overall productivity of the business. It is also important in making the most accurate and relevant information available quickly for decision making.**

This work focuses on augmentation of the customer dataset using Genetic Algorithm(GA). These augmented data are used for the purpose of customer behavioral analysis. The data set consists of the different factors inherent in each situation of the customer to understand the market strategy. This behavioral data is used in the earlier work of analyzing the data [13]. It is found that collecting a very large amount of such data manually is a very cumbersome process. It is inferred from the earlier work [13] that the more number of data may give accurate result. Hence it is decided to enrich the dataset by using Genetic Algorithm.

I. INTRODUCTION

In today's competitive business environment it is much tougher to understand the opinion of the customer towards the purchase of a product. People are more mobile oriented and better informed. The personalized, individualized, and relevant information of the customers are required for business intelligence appraisal. In the previous work [13] the data is collected manually for the customer behaviour analysis. The experiment summarizes that the purchase behaviour of a person is purely related to his/her credentials (e.g. Hobby). From the result and analysis, it is observed that, with the huge dataset it is still possible to improve the advocacy level of the customer. As the manual data collection is tedious and time consuming, it is decided to generate the data by data augmentation using Genetic Algorithm.

Introduction to Genetic Algorithm (GA): **GA is** inspired by the process of that belongs to the larger class of Evolutionary Algorithm (EA). Genetic Algorithms are commonly used to generate high-quality solutions to and by relying on bio-inspired operators such as mutation, crossover and selection [4].

Functionality of GAs: Three basic operators responsible for GA are (a) selection, (b) crossover and (c) mutation [8].

Crossover performs combination of different solutions to ensure that the genetic information of a child life is made up of the genes from each parent. *Figure* 1 Illustrates the process of generating a dataset using GA. The reason behind selecting GA for the augmentation of data is due to its benefits [6]. (1) Generality and Versatility [6]: GA applied in a wide variety of settings and can be easily moulded to particular problems. (2) Robust and Online Problem Solving [6]: The decisions will be made automatically in run-time to cater to dynamic channel parameters indicating it is a faster process. (3) Support for Global Optimization [6] GAs is suited to find the global optima due to a number of properties

- Search by means of a population of individuals.
- Work with an encoding of the multiple parameters.
- Use a fitness function that does not require the calculation of derivatives.
- Search probabilistically.

(4) GA is computationally simpler compared to other complementary Artificial Intelligence techniques [6]. (5) GAs use evolutionary techniques to test and improve the solutions by using techniques such as mutation, crossover, selection, and recombination [8]. The important benefit of enhancing data in this paper are: (1) The data collection efforts are reduced by enhancing datasets (2) Easy to generate any number of data in future as the enhanced data is accurate. (3) To achieve accurate results.

For the previous work [13] the data collected from the respondent directly to understand the market strategy and the different factors inherent in each situation of the customers. It is inferred from the earlier work that the more number of data may give accurate result. Hence it is decided to enrich the dataset by using Genetic Algorithm.

Mehboob, Junaid Qadir, Salman Ali, and Athanasios Vasilakos [6] provided a detailed survey of applications of GA using different kinds of GA techniques in wireless networking. They have also highlighted pitfalls and challenges in successfully implementing GAs in wireless networks and open issues of GA.

Moheb R. Girgis [14] presents an automatic test data generation technique using Genetic Algorithm. The GA technique presented in this paper is guided by the data flow dependencies in the program to search for test data to fulfil the all-uses criterion. The algorithm produces a set of test cases, the set of def-use paths covered by each test case, and a list of uncovered def-use paths. Experiments have been carried out to evaluate the effectiveness of the proposed GA compared to the random testing technique, and to compare the proposed random selection method to the roulette wheel method. The results of these experiments showed that the GA technique outperformed the random testing technique in 12 out of the 15 programs used in the experiment. The experiments also showed that the proposed selection method produced better results than the roulette wheel method [15].

M. Anbarasi et. al. [16] attempt 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.

Amit Kumar Sharma [17] 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.

Silvia TRIF [18] 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 [19] 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.

II. METHODOLOGY

For this work the practical data was collected from various public and private sectors from various regions to achieve the uniformity and consistency in data. Following block diagram *Figure* 1 illustrates this Methodology: This has got three processes mentioned as follows

1. Data acquisition
2. Data cleansing
3. Data Transforming and moulding

These processes are explained in the next section

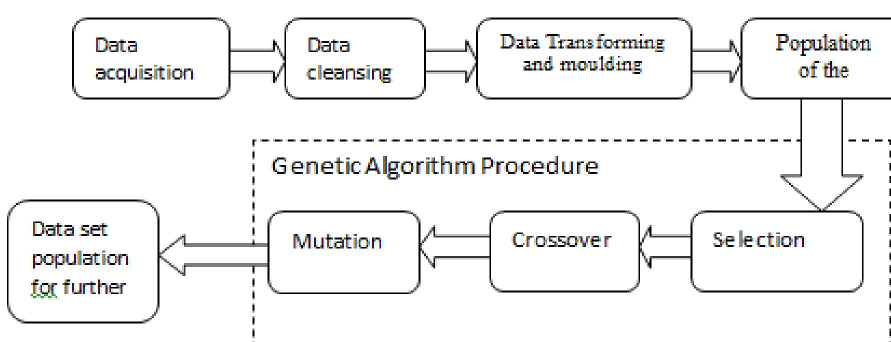

Figure 1: Schematic block diagram of the generation of a new data set using genetic algorithm.

Data Acquisition: The reason behind selecting behavioural for the purpose of this work is to understand the market strategy and the different factors inherent in each situation. These factors play a major role on habits, profession and opinion of the customers. Decisions and buying behaviour are obviously also influenced by the characteristics of each customer. A consumer does not buy the same product or service at 20 or 70 years. His lifestyle, values, environment, activities, hobbies, age group and consumer habits evolve throughout his life. The factors influencing the buying decision process may also change. The lifestyle of a person will influence on his behaviour and purchasing decisions. For example, a human with a healthy and balanced lifestyle will prefer to eat organic products and go to specific grocery stores, will do some jogging regularly (and therefore will buy shoes, clothes and specific products), etc. The occupation and economic situation of a person also has significant impact on his buying behaviour. For example, a marketing manager of an organization will try to purchase business suits, whereas a student tries to purchase books or stationary and a housewife try to purchase household items. Hobbies reflect the inner most desires of people, help them fulfil their needs. So, it is obvious that all these factors influence the purchase pattern. By identifying and understanding these factors, purchasing of the product can be predicted for a new customer.

Based on the following data given in Table1 from different customers, the whole analysis is prepared. Table 1 List the credentials of the customers used for the analysis of the Business Intelligence. This study requires large set of live data to obtain the accurate results. Since, online customer websites like Amazon, Flipkart have large amount of such data however they are not available for research and other purposes due to confidentiality reason. Moreover, those data may not have all the parameter that we are planning to capture as mentioned above. This work takes the data from different sources required for evaluation of the prediction. Since, the type of data required depends purely on behavioural aspects of the respondents, further proceedings of data collection were done by collecting the live data from various public and private sectors. The typical data set contains gender, age group, hobbies, profession, product's usage and opinion on used products. The data collection was done from various sources to obtain various categories like different age group, different occupation, etc.

The data collected by the Google forms are shown in the following table below.

Table 1. A Typical dataset generated by the Google form

Gender	Occupa tion	Age gro up	Income gro up	Hobby1	Hobby2	Product category	Elec-tronics	Satisfact-ion Level	Advo-cacy Level	Feeling
Male	Engi neer	40-49 yrs	5 Lakh To 8 Lakh	Read ing	Photogra phy	Electronic	Laptop	6-very satisfied	3- Neutral	Neutral
Female	Doctor	20-29 yrs	2.5 L to 5 L	Reading	Music	Books	Books	6-very satisfied	4-Likely	Happy
Male	Athlete	30-39 yrs	2.5 L to 5 L	Sports	Adv sports	Sports	Men sports cloth	5- satisfied	4-Likely	Happy
Female	House maker	20-29 yrs	---	gardenin g	Singing	Household	Refriger ator	7-Ex. Satisfied	4-Likely	Happy
Male	Student	20-29 yrs	---	Arts & crafts	Playing with pet	household	Painting kit	6-very satisfied	5-Most likely	Happy

The issues while collecting data manually by interacting with the respondents are:

- Manual data collection consumed more time because challenges faced by the respondents due to poor knowledge of English.
- Time constrains to respond to the survey during office hours. They were unreachable in their post during office hours.
- Entry related access restrictions to certain offices; need for prior permission in such cases.
- Issues related to geographical spread of people in reaching them.
- Manual data collection includes inconsistencies. Hence some percentage of data will be invalid and become waste.

To avoid these problems the questionnaire is distributed through *Google* forms. Only 20 percent of people responded for the Google forms.

Hence this work is taken up to generate a large dataset for further improvements and for the better and accurate results. As this work required more number of dataset, GA is proposed to enrich the dataset to improve the accuracy of the result.

*Data Cleansing***:** Data Cleansing is the process of identifying and correcting inaccurate data from a data set. With reference to customer profile data, data cleansing is the process of maintaining consistent and accurate customer data through identification and removal of incorrect, incomplete, out-of-date data. Data cleansing help in the creation of a clean customer datasets which offers multiple benefits across functions and serves as a critical factor in the growth of business.

Real-world datasets are highly susceptible to missing and inconsistent data, lacking certain attributes of interest. Low-quality data or un-cleansed will lead to low-quality results. As this paper collected the real world data, which has a mixture of raw data with the datasets. This has to be filtered by manually as well as by machine in order to improve the quality of the data. This stage includes filling missing values, identify or remove outliers, and resolve inconsistencies. Missing values are filled and resolved the inconsistencies from the original dataset. It has been rejected records of these kinds as the information is very less and inconsistent for the processing.

The purpose of this work is to explore the views, experiences of customers with different hobbies, professions, gender etc., on specific products which they have used. For this study, sufficient large number of datasets with different verities are collected. The data is collected by distributing the questionnaire to the respondents directly and also collected through the Google Forms from various locations. As this work is related to behavioural data, more importance is given to parameters such as hobby and profession. These parameters play a vital role in purchase behaviour of a customer.

III. FLOW OF A TYPICAL GENETIC ALGORITHM

Three basic operators responsible for GA are (a) selection, (b) crossover & (c) mutation. Crossover performs recombination of different solutions to ensure that the genetic information of a child life is made up of the genes from each parent. The Figure1 above illustrates the process of generating a dataset using Genetic Algorithm.

As a first step the existing data sets are randomly populated. Out of these *n* record sets, a record *R1* is selected and all the chromosomes of *R1* that is *P0, P1…Pt* will be copied to the New record *Rng1*. The record set *Rng1,* is taken for further process of cross over and mutation. In this Rng1 one or a few chromosomes (*Pi, Pj, Pl*) are selected, for crossover and mutation. This process of cross over is explained in the next section.

Cross over:

In this work, crossover process uses random operation to generate new record from two parent records. As explained in the previous section after copying the record into new record *Rng1*, one more record *R2* is selected from *n* record set and taken for crossover. Since it is preferable to carry out the crossover for the chromosomes, randomly some chromosomes *Pi* have been selected from the new record set for the process of crossover. In the selected (*Pi)*sets of chromosomes random bits are selected for the crossover. These selected bits from the chromosomes (*Pi's*) of *Rng1* are replaced from those of *R2* chromosome.

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.

Example: A record R*ng*1 is selected which is having chromosomes *P0, P1…Pt*. Let for example one chromosome *Pi* is selected randomly that contains the bits as shown in the Figure2 bellow. Similarly, same set of chromosomes is selected from R2 as shown in Figure 3.

Y0	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10

Figure 2. Bit patterns of pi^{th} Chromosome of Rng1

X0	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10

Figure 3. Bit patterns of pi^{th} Chromosome of R2

In the above chromosome of R1 any random bits are selected, for example 1^{st} 4^{th} 7^{th} and 10^{th} bits these bits are replaced by the bits of the same place from the chromosome selected from R2, which forms a new chromosome of category Pi as shown in *Figure* 4.

Y0	X1	Y2	Y3	X4	Y5	Y6	X7	Y8	Y9	X10

Figure 4. Bit patterns of pi the newly generated Chromosome after crossover

This crossover chromosome is copied in to a new data set or else the same chromosome is processed further for mutation.

The Mutation:

The above crossover chromosome shown in *Figure* 4 is further taken for the mutation. The process of mutation is as shown in *Figure* 4. In this process of mutation, the most significant bit is written as it is as Y0. Zeroth bit and first bit are XORed the resultant is pleased in the first bit of the resultant chromosome that is resultant of Y0 X1. First and second bits are XORed to get the result as second bit, that is X1 Y2 leads to the resultant of the third bit. Similarly, all the bits are calculated as shown in the Figure 5, for the process of mutation. The advantage of this method is there is no over flow and data loss.

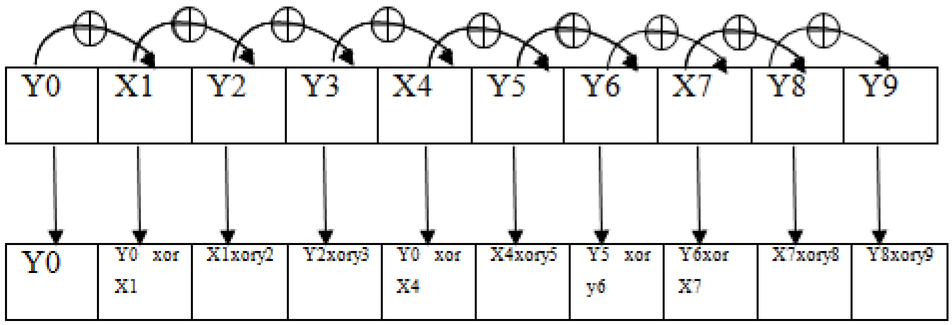

Figure 5. Example of Method of Mutation for a given datasets.

Example for Mutation :

This will be demonstrated as the new chromosome after mutation of 11010110 is illustrated next. To maintain the trace earlier chromosome, x1 will be written as it is in the new chromosome. x1 is XORed with x2 written in the position of x2. X2 is XORed with x3 written in the position of x3. Similarly, it will continue till the end that is x8. The new chromosome generated in this example is 10111101. After this process of mutation, the resultant chromosome is introduced in the data set and the dataset is tested for the validation. The next section explains about the process of validation.

Validation rules for the enhanced dataset

To ensure the resultant data set is in line with the existing data set, which is nearer to the real life and correctness of the dataset is validate during the association rule. This validation procedure is applied to the enhanced data set. Data validation is intended to provide certain well-defined guarantees for fitness, accuracy, and consistency for any of various kinds of user input. It also confirms that the following rules that have been established for the applications to validate data prior to sending updates the underlying database of the data sets. The resultant data set is 'rejected' in case if it satisfies any of the following association rules:

1. Gender=male and Occupation=Home maker.
2. Occupation=student and Age group between 30 to 39 or above 50 years
3. Occupation=student and Income group between 2.5 to 5 Lakhs or above 12 lakhs
4. Occupation=office assistant and Income group above 12 lakhs
5. Occupation=Home maker and Age group is under 19
6. Age group=under 19 and Occupation ≠ student.
7. Age group=above 50 years and Hobby1=Adventurous sports.
8. Age group=above 50 years and Hobby2=Adventurous sports
9. Age group= under 19 and Occupation other than student
10. Satisfaction level=Extremely dissatisfied and advocacy level=Likely/ Most Likely.
11. Satisfaction level=Extremely satisfied and advocacy level=Unlikely/ Most Unlikely.
12. Satisfaction level=very Dissatisfied and advocacy level=Likely or Most Likely.
13. Satisfaction level=very satisfied and advocacy level=Unlikely/Most unlikely
14. Satisfaction level=Satisfied and advocacy level=Most Unlikely
15. Satisfaction level=Neutral and advocacy level=Most likely
16. Satisfaction level=Dissatisfied and advocacy level=Most likely

IV. RESULTS AND ANALYSIS

Every new record is generated by picking up two random records from the existing dataset. The hobby parameters of these two records along with age group, occupation, income group and satisfaction are crossed over correspondingly to get the new record. Random 3, 4, 5, 6, and 7 bits are used for crossing over of parameters. The newly generated record is further mutated to modify its characteristics as per the genetic algorithm procedure. This final new record is vetted through a validation procedure to validate its authenticity to make it look like a real survey response. The validation procedure involves checking the new record parameters for their valid values as well as checking their valid combination. For example, a record with age group under 19 years with a high-income group is considered invalid. Similarly, records with conflicting values for Satisfaction and Advocacy, for e.g. Very Dissatisfied with the product but Most Likely Advocating it to others.

The Figure 6 below shows the rejection rate of newly generated records in the dataset of 5000 records. As can be seen in the graph below in Figure 6, the rejection rate dips when the crossover between two records happens with higher number of bits (6 or 7) with a combination of parameters. However, when the crossover is made using hobby and one other parameter, the dip is highest in 5 bits.

It is obvious from the graph that for these attempted data the drop is high. However, by the observation of the following graph shown below in Figure 6, it is evident that in case of multiple attribute crossovers the drop of the generated data is very less compared to the data of the attribute alone.

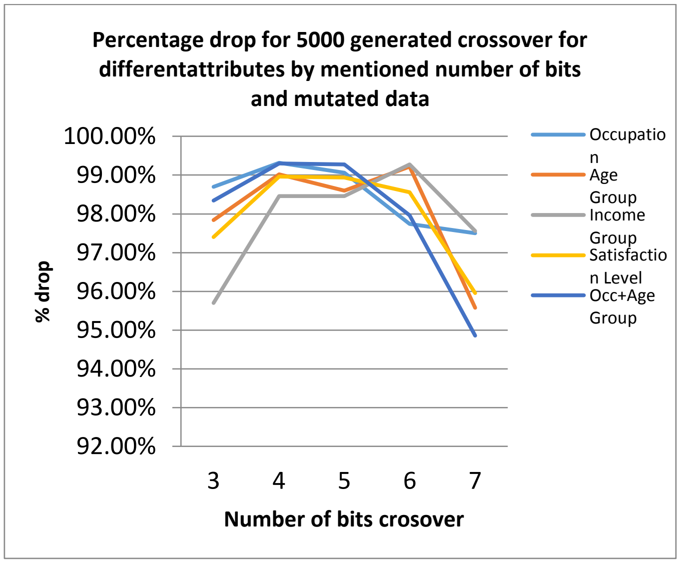

Figure 6. Graph of Percentage drop for 5000 generated crossovers for different attributes by mentioned number of bits and mutated data.

The negative peak in the graph shows the drop in the generated genome. This is very much helpful to decide about the attributes of the genome for the maximum generation during the new generation. It is possible to find well in advance the rejection and acceptance at the earlier stage, so as to improve the rejection rate.

In the second case, 5 bits are taken for the crossover operation. The percentage drop versus number of bits are plotted in the graph shown below in Figure 7. The observation is made on the basis of the following Figure 7. This graph takes percentage drop in the Percentage Drop axis and number of bits in the case of number of bits taken for the crossover axis. It is evident that the graph drops during the selection of 5bits for the crossover. This indicates clearly that

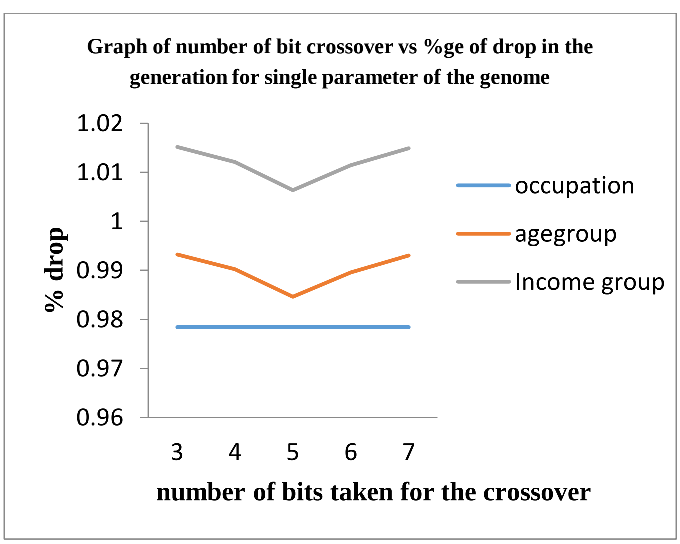

Figure 7. Graph of Percentage drop for 5000 generated crossovers for different attributes by mentioned number of bits and mutated data.

for the better performance, it is better to take 5 bits for the crossover. However, in this case the crossover is done for only one parameter of the genome. When two or more parameters are considered the result is shown as in the Figure 8 below.

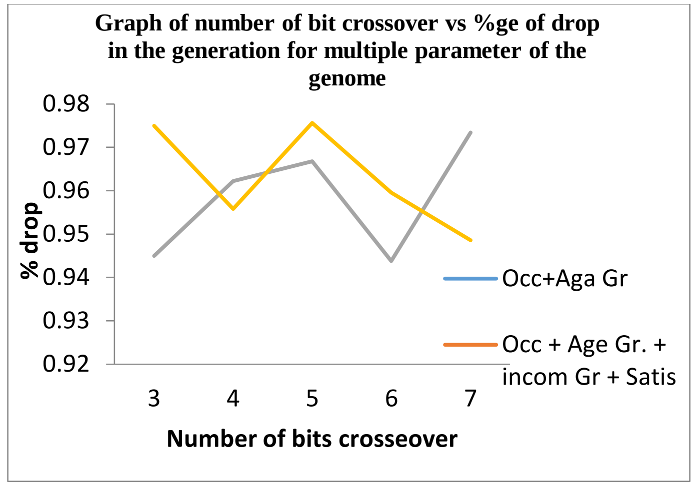

Figure 8: Graph of number of bit crossover vs percentage of drop in the generation for multiple parameters of the genome

In the Figure 8 it is obvious that more than 5 bits gives a good value of percentage drop. Considering the Figure 2 and Figure 4 it is apparent that more than 5 bits of crossover and mutation will leads to a good combination for the generation of the new genome.

V. CONCLUSION

It is observed that, it is possible to generate a valid new record by picking up two random records from the existing dataset. It is also observed from the result and analysis that in case of multiple attribute crossovers the drop in the generated data is very less compared to, the crossover when only one data of the attribute is considered. For the dataset of 5000, the rejection rate dips with 6 or 7 bits of crossover operation, and will get about 5% of datasets accepted. However, from the result of multiple attribute crossovers the drop of the generated data is very less compared to the data of the attribute alone and more than 5 bits of crossover and mutation will lead to a good combination for the generation of the new genome. However, it is also possible to enhance this percentage by evaluation of the data set for acceptance or rejection in advance.

REFERENCES

[1]. Erhard Rahm, Hong Hai Do, “Data Cleaning: Problems and Current Approaches”, University of Leipzig,Germany. http://dbs.uni-leipzig.de.

[2]. Heiko Müller, Johann-Christoph Freytag, “Problems, Methods, and Challenges in Comprehensive DataCleansing”, Humboldt-Universität zu Berlin zu Berlin, 10099 Berlin, Germany.

[3]. Grefenstette, J.J. and Baker J.E. How Genetic Algorithms Work: A Critical Look at Implicit Parallelism. In Schaffer, J.D. (ed.),

Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo, CA,1989, pp. 20–27.

[4]. Genetic algorithm [Online] Available: http://en.wikipedia.org/wiki/Genetic_algorithm

[5]. D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," Machine learning, vol. 3, no. 2, 1988, pp. 95–99.

[6]. Mehboob, Junaid Qadir, Salman Ali, and Athanasios Vasilakos, "Genetic Algorithms in Wireless Networking: Techniques, Applications, and Issues" arXiv :1411, CS.NI

[7]. Li, L Weinberg., Darden T.A. and Pedersen,L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 17, 2001, pp 1131–1142

[8]. Melanie Mitchell Santa Fe L. D. Davis, "Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold", Santa Fe., 1991.

[9]. Shruthi Rathnakar, K. Rajeswari, Rose Jacob, "Prediction of Heart Disease Using Genetic Algorithm For Selection of Optimal Reduced Set of Attributes "International Journal of Advanced Computational Engineering and Networking, Volume-1, Issue-2, April 2013.

[10]. K. F. Man, Member, IEEE, K. S. Tang, and S. Kwong, "Genetic Algorithms: Concepts and Applications", IEEE Transactions on Industrial electronics, VOL. 43, NO. 5, October 1996

[11]. Mitchell Melanie, "An Introduction to Genetic Algorithms A Bradford Book ", The MIT Press Cambridge, Massachusetts, England, Fifth printing, 1999.

[12]. 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, 2005.

[13]. Nethravathi P. S, K.Karibasappa, "Business Intelligence Appraisal of the Customer Dataset Based on Weighted Correlation Index", International Journal of Emerging Technology and Research. 2016.

[14]. 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

[15]. D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," Machine learning, vol. 3, no. 2, 1988, pp. 95–99.

[16]. 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.

[17]. A.K. Sharma, "Text Book of Correlations and Regression", Discovery Publishing House, 2005.

[18]. Silvia TRIF, "Using Genetic Algorithms in Secured Business Intelligence Mobile Applications", Informatica Economica, Volume 15, Romania.2011.

[19]. Nidhi Bhatla Kiran Jyoti , "International Journal of Engineering Research & Technology IJERT Vol. 1 Issue 8, October 2012.