Using Genetic Algorithm for Efficient Mining of Diabetic Data

Mr. S. Kovalan[1], Mr. S. Mugilan[2], Mr. S. Balamurugan[3]
[1,2]PG Student, Sri ManakulaVinayagar Engineering College,Pondicherry-605106
[3]Assistant Professor, Sri ManakulaVinayagar Engineering College,Pondicherry-605106

***Abstract*– The paper we discuss about the drawing out of diabetic data by means of the Genetic Algorithm from the repeated datasets.The Knowledge Discovery in Database(KDD) process will be used and on the datamining point the Genetic Algorithm.The Genetic Algorithm useful in capturing corresponding patterns in data when the comparative magnitudes of data items are more important than their accurate values.The KDD process will be included with Selection, Pre-processing, Fitness Function, Data Mining. This algorithm will improve with analyzing of data easily from the large database with the minimal time and higher accuracy.**

Key Terms: KDD, Data Mining, Gene Selection, Genetic Algorithm, Fitness Function

1. INTRODUCTION

Diabetes is a collection of sicknesses marked by high levels of blood glucose, also named blood sugar, causing from defects in insulin creation, insulin accomplishment, or both. Diabetes can clue to serious complications and early death. The serious complications diabetes can be associated with include heart infection and stroke, high blood pressure, loss of sight, kidney infection, nervous system syndrome, deletions, dental ailment, and complications of pregnancy. Diabetes was the seventh importantsource of death recorded on U.S. Overall, the danger of death among people with diabetes is about double that of people short of diabetes of a comparable age.

Data mining is the process of choosing, discovering, and exhibitinghugevolumes of data to determine unknown patterns or associations useful to the data analyst. Nonetheless the common accepting that data mining smears to data analysis problems with huge amounts of data and is solved by resorting to effective mixture of investigative methods, the term data mining is still undefined, mostly due to its process-based nature. A hugerange of methods are considered part of data mining, together withcomputer science methods, such as multidimensional databanks, machine learning, soft figuring and data picturing, and statistical-based methods, including hypothesis testing, grouping, arrangement, and regression. The objectives of data mining can be classified into two tasks: description and prediction. While the purpose of description is to mine understandable forms and relations from data, the goal of prediction is to forecast one or more variables of interest.

2. BASIC GENETIC ALGORITHM

Genetic algorithm is adaptive heuristic search method premised on the evolutionary ideas of natural selection and genetics. The basic concept of Genetic Algorithm (GA) is to simulate processes in natural systems necessary for evolution, particularly those that follow the principles of survival of the fittest. It is usually applied in situations where the search space is relatively large and cannot be traversed efficiently by classical search methods. This is mostly the case with problems whose solution requires evaluation and equilibration of many apparently unrelated variables. As such they represent an intelligent exploitation of a random search space within a defined search space to solve a problem.

3. GENETIC ALGORITHM FOR DIABETIC MEASUREMENTS

The diabetic quantities are calculated with various features such as Pregnancies, PG Attentiveness, Diastolic BP, Tri Fold Thick, Serum Ins, BMI, DP Function, and Age. The diagnosis for the diabetic patients is all based on the calculation from the quantities.

With the genetic algorithm, all the measurements are processed with the initial population which provides the data for the fitness function. The fitness function will be calculated with the current population. The selection probability is the step of genetic algorithm for processing the fitness values. From the fitness values new or next generation values are developed. The selected values are accessed from the current population to produce offspring. These offspring processed from the arithmetic operators.

3.1 Initialization

The genetic algorithms are generally specified with an opening population that is generated randomlyusing special practices to produce a superior quality opening population. Such method is intended to give the GA a good start and speed up the evolutionary process.

In a sample data of diabetic problem, we can use a non-binary bit string demonstration to signify the chromosome because it is easy to understand and represent. We use six positions representing six exams with each location's value as the time slot allocated to the exam. We can create the population randomly to allocate each data.

Patient BP BMI

Patient1 (140/90mm) e1 (30kg/m$^{2)}$ e3

Patient2 (140/90mm)e5 e6 (30kg/m$^{2)}$ e2, e4

If we randomly generate six numbers 3, 8, 4, 8, 6, 7 as six timeslots for *e1-e6,* then the chromosome is *3 8 4 8 6 7*.

If the population size is 5, an initial population can be generated randomly as follows:

index	chromosome	Fitness
1	*3 8 4 8 6 7*	0.005
2	7 3 7 6 1 3	0.062
3	5 3 5 5 5 8	0.006
4	7 6 7 7 2 2	0.020
5	1 7 4 5 2 2	0.040

3.2 Parent Selection Mechanism

The effect of selection is to return a probabilistically designated parent. Although this selection technique is stochastic, it does not imply GA employ a meaningless search. The chance of each parent being nominated is in some way related to its fitness standard, unique method for parent selection is Roulette Wheel selection or fitness-based selection. In this kind of parental selection, each chromosome has accidental of selection that is openlyrelational to its fitness. The effect of this depends on the sort of fitness values in the current population. The unique tournament selection is to choose K parents at random and returns the fittest one of these.

Fitness-based selection

The standard, unique method for parent choice is Roulette Wheel selection or fitness-based choice. In this kind of parentalselection, each chromosome has accidental of selection that is openlyproportionate to its fitness. The result of this depends on the series of fitness values in the current population.If fitness ranges from 5 to 10, then the fittest chromosome is double as probable to be selected as a parent than the least fit.

If we apply fitness-based selection on the population given in sample, we choice the second chromosome 7 3 7 6 1 3 as our first parent and 1 7 4 5 2 2 as our second parent.

3.3 Crossover

The crossover operator is the most significant in GA. Crossover is a process acquiescent recombination of bit strings via anreplace of segments between pairs of chromosomes. The process of one-point crossover is to randomly produce a number (less than or equal to the chromosome length) as the crossover location. Then, keep the bits before the number unmoved and swap the bits after the crossover location between the two parents.

Parent Parent1: 7 3 7 6 1 3

Parent2: 1 7 4 5 2 2

Children Child 1: 7 3| 4 5 2 2

Child 2: 1 7| 7 6 1 3

3.4 Mutation

Mutation has the effect of safeguarding that all possible chromosomes are manageable. With crossover and uniform inversion, the quest is inhibited to alleles which exist in the opening population. The mutation operator can overwhelm this by simply randomly selecting any bit location in a string and changing it. This is useful since crossover and reversal may not be able to produce new alleles if they do not appear in the initial generation.

New string: 7 3 4 5 1 3

Mutation rate is 0.001

For the first bit 7

Generaterandomly between 0 and 1

First bit 7 needs to mutate

Generate another between 1 and 8

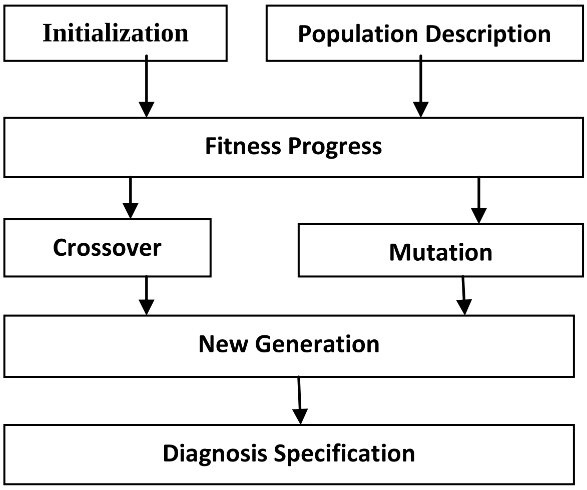
First bit mutates to 2

Repeat the same procedure

New chromosome 2 3 4 5 1 3

3.5 Specification of Diagnosis

All the data from the previous modules with their fitness progress, new generation and provides the inter cluster distance, intra cluster distance, accuracy for diabetic measurement results and specification of the diagnosis.



4. IMPLEMENTATION AND EVALUATION OF GENETIC ALGORITHM

In this section, we recommend the various methods and function of the genetic algorithm. Algorithm performs the following steps:

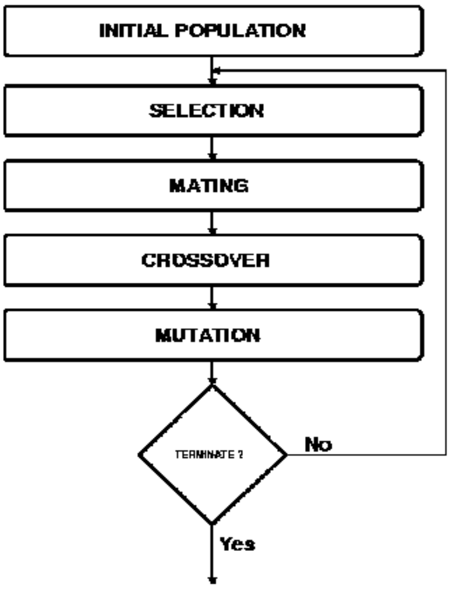
1. Generate an initial population , randomly or heuristically.

2. Compute and save the fitness for each individual in the current population.

3. Define selection probability for each individual so that it is proportional to its fitness.

4. Generate the next current population by probabilistically selecting the individuals from the previous current population, in order to produce offspring via genetic operators.

5. Repeat step 2 until a satisfactory solution is obtained.



Flow chart for Genetic Algorithm

As an initialization step, genetic algorithm generates randomly a set of solutions to a problem (*a population of genomes*). Then it enters a cycle where fitness values for all solutions in a current population are calculated, individuals for *mating pool* are selected (using the operator of *reproduction*), and after performing *crossover* and *mutation* on genomes in the mating pool, offspring are inserted into a population and some old solutions are discarded. Thus a new *generation* is obtained and the process begins again.

Genetic algorithm stops after the *stopping criteria* are met, i.e. the "perfect" solution is occured, or the number of generations has meet its maximum value .

The **first step** of a genetic algorithm is to define a search space and describe a complete solution of a problem in the form of a data structure that can be processed by a computer. Generally,Strings and trees are used, at same time any other representation could be relatively accessible, provided that the below procedures can be occured. Theresult is called to as *genome* or *individual*.

The **second step** is to define a convenient evaluation function (*fitness function*) whose task is to determine what solutions are better than others. Many problems require a specific definition of the fitness function which works best in that case.

The **third step** in the creation of a genetic algorithm is to define reproduction, crossover, and mutation operators are transforms the current generation into the next generation. Reproduction can be generalized, for each problem one can bring out individuals for mating accidentaly, or based on their fitness function. The harder part is to define crossover and mutation operators.

Crossover generates a new offspring by combining genetic material betweengiven parents. It incarnates with assumes that the result which has a high fitness value ows it to a gene combination. Combining genetic material from two given individuals, better results can be gained. Mutation provides some randomness with the population. Mutation randomly changes some genes in an individual, introducing varietywith the population and exploring a huge search space.

The **fourth step** is to define process ending criteria. Algorithm can ends after it has produced a definite number of generations, or when the improvement in accessible fitness over two generations is below a threshold. The nextmethod is better, yet the aim might be tough to reach, so the previous one is more valuable.

A fitness function is used to evaluate individuals, and reproductive success varies with fitness. An effective GA representation and meaningful fitness evaluation are the keys of the success in GA applications.

5. Pseudocode for Genetic Algorithm

i) Select initial population

ii) Calculate the fitness of each distinct

In the population

iii) Repeat

a) Select best-ranking individuals in the population

b) Sort new generation over crossover and mutation (genetic operations)

Give birth to offspring

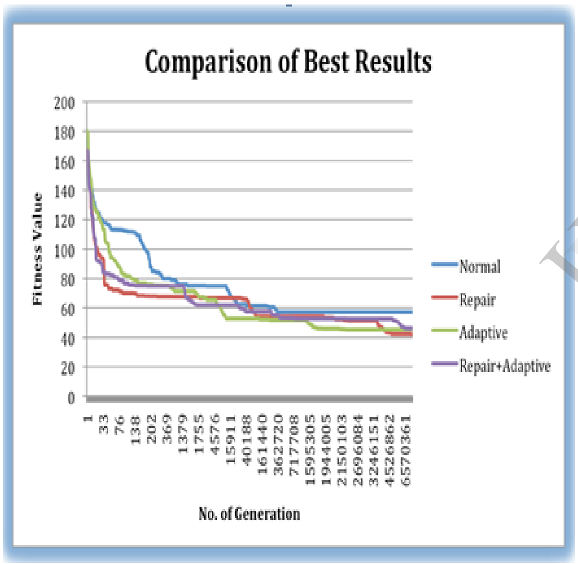
c)Calculate the distinct fitnesses of the offspring

d)Swapvilestordered part of population with offspring

iv) Until termination

v) Specification of Diagnosis

6. FITNESS PROGRESS FOR GENETIC ALGORITHM



7. CONCLUSION

The paper we provide a genetic algorithm move toward for classification of diabetic data problem. Genetic Algorithm has been used to analyze large datasets and establish useful classifications and patterns in the datasets. The effectiveness of the classification algorithm which could provide the examine of data from comparatively **larger databases** and**time consequences are lower** with previous algorithms. The efficient specification of the diagnosis is provided from the given diabetic measurements with the methods of the genetic algorithm.

8. REFERENCES

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rulesin large databases. In VLDB, pages 487–499, 1994.
2. R. Agrawal and R. Srikant. Privacy-preserving data mining. In SIGMOD Conference, pages 439–450, 2000.
3. Chen, M.S., Han, J., and Yu, P.S., 1996. Data mining: An overview from a database perspective. IEEE Transactions. Knowledge and Data Engineering, Vol. 8, Issue 6. pp. 866-883.
4. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., 1996. Advances in Knowledge Discovery and Data Mining. MIT Press.
5. Brachman, R., and Anand, T., 1996. The Process of Knowledge Discovery in Databases: A Human-CenteredApproach. Advances in Knowledge Discovery and Data Mining, pp. 37–58. AAAI Press.
6. Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading. MA: Addison-Wesley.
7. S. Forrest and G. Mayer-Kress, “Genetic algorithms, nonlinear dynamical systems, and models of intemational security,” in Handbook of Genetic Algorithms, L. Davis, Ed. New York: Van Nostrand Reinhold, 1991, pp. 166-185.
8. D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989.
9. C. K. Chui, B. Kao, K. Y. Yip, and S. D. Lee, “Mining order preservingsubmatrices from data with repeated measurements,”in Eighth IEEE International Conference on Data Mining (ICDM’08),2008, pp. 133–142.
10. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders,M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, “Comprehensiveidentification of cell cycle-regulated genes of the yeastsaccharomyces cerevisiae by microarray hybridization,” MolecularBiology of the Cell, vol. 9, no. 12, pp. 3273–3297, 1998.
11. A. Ben-Dor, B. Chor, R. M. Karp, and Z. Yakhini, “Discoveringlocal structure in gene expression data: the order-preservingsubmatrix problem,” Journal of Computational Biology, vol. 10, no.3-4, pp. 373–384, 2003.
12. L. Cheung, K. Y. Yip, D. W. Cheung, B. Kao, and M. K. Ng,“On mining micro-array data by order-preserving submatrix,”International Journal of Bioinformatics Research and Applications, vol. 3, no. 1, pp. 42–64, 2007.