An Effective Genetic Algorithm for Outlier Detection

P. Vishnu Raja
Assistant Professor(SrG)/CSE
Kongu Engineering College
Perundurai, Erode

Dr. V. Murali Bhaskaran
Principal
Pavaai College of Engineering
Pachal, Namakkal

ABSTRACT

The main objective the outlier detection is to find the data that are exceptional from other data in the data set. Detection of such exceptional data's is an important issue in many fields like fraud detection, Intrusion detection and Medicine . In this paper we are proposing an algorithm to detect outliers using genetic algorithm. The proposed method was exceptionally accurate in identifying the outliers the datasets that we have tested. The result analysis is done on some standard dataset to view accuracy of the algorithm.

General Terms

Database, Data storage, Information retrieval.

Keywords

Outliers, Genetic algorithm, Anomalies, Exceptional objects Optimization.

1. INTRODUCTION

Detecting outliers is an important issue in many of the common applications like fraud detection, intrusion detection, medicine, network robustness analysis and so on. Finding the Outliers or the rare instances will be more interesting compared to identifying the common data of usual form. Outliers in a dataset is defined informally as an observation that is considerably different from the remainders as if it is generated by different mechanism which are exceptional from the remaining data in a dataset[1][2].

In many of the data mining applications identifying the outliers or rare events discovers some new interesting and unexpected knowledge in many areas. It has been examined that in most of the algorithms that are developed to detect anomoly are not accurate [2]. It may detect the false data or an additional data which are not outliers which leads to false result. The results thus produced are also not optimized.

In this paper, we proposed an generalized genetic algorithm for identifying the exceptional objects from the dataset which also includes outliers. This is due to the fact that Genetic algorithm are very simple and easy to use and also computationally powerful. Many of the searching and optimization algorithms are not adaptive[10]. In the sense that they generally solve only the given problem. Since the algorithm is designed for their problem alone. But Genetic algorithm are adaptive and robust in nature[9][10], they can be applied to any domain and to any type of problem with slight modifications in the representation, fitness value or with the choice of the genetic operators. But the behavior of the genetic algorithm remains same. So we had chosen Genetic algorithm as our algorithm to solve outliers. In our approach the outliers are identified based on the fitness value that is generated. The fitness value that are lower are considered to be outlier.

The remainder part of this paper is described as follows: in section 2 we discussed about the related work done and the proposed work in detail and section 3 describes genetic algorithm in detail and section 4 shows the experimental results of the proposed algorithm. Finally section 5 concludes the paper with future work.

2. RELATED WORK

There is no generic approach is done to detect outliers. Many approaches have been proposed to detect or identify the outliers based on density based, distance based, distribution based and clustering based approaches.

In the Density based approach they compute the data with density of regions in which low density regions are identified as outliers. In (Breunig 2000, Papadimitriou 2003) assigned LOF(Local Outlier Factor) as an outlier score for to any given data point based on the distance from its local neighborhood.

In the Distance based approach[(knorr 2000, angiulli 2005) the outliers are detected by a distance measure on the feature space. In ramasamy(2000), the outliers are identified by using k-nearest neighbor method to rank the outliers. The problem with this approach is that it is very difficult to find a particular value in a dataset[3].

Distribution based approach (Rosseeuw 1996) had developed statistical methods from the given data and applied statistical test to find the object belong to a particular model or not. The object with low probability are identified as outliers in the statistical model. Because the distribution based approaches are univariate in nature they cannot be applied in multidimensional data space

Clustering based approach (Achuna and Rodriguez 2004) identified outliers as clusters of small sizes. The advantage of this approach is that it may not be supervised. Hierarchical based approach(Loureiro, 2004 and Almeida 2006) was used to identify the outliers by using the resultant clusters as an indicator to identify the outliers.

Many algorithms have been proposed to identify the outliers but optimized solution has not been defined. In this paper we proposed Genetic algorithm based outlier detection to have effective optimum result.

3. GENETIC ALGORITHM

Compared to other searching algorithms Genetic algorithms are adaptive heuristic and robust in nature which implies that they can be applied problems of any domain with slight modification of the representation, fitness evaluation and the choice of the genetic operators but the basic operation of the

algorithm remains same. Unlike other search algorithms, GA does not require any additional information about the problem.

In Genetic algorithm, the chromosomes are generated with a set of population strings which encode candidate solutions called individuals. Each individual led to an optimization problem which evolves toward better optimization solution. In general the solutions are represented as strings of 0's and 1's in the binary form, but other encodings like value, permutation encoding are also possible. Here, the evolution starts from the initial population with the randomly selected individuals. Then it seeks to optimize the function by an continuous iterative process of selection, crossover and mutation until global optimum population is obtained.

In each generation, every individual's fitness is evaluated by selecting multiple individuals based on their fitness value from the current population and recombined to form new population. This obtained new population is used in the next iteration of the algorithm. The termination of the algorithm takes place when a maximum number of generation has been reached from the population.

The genetic algorithm employs the following process to produce the best individuals from the initial set of populations:

Input: Set on N Chromosomes in the search space.

Output: Outliers with lowest fitness value.

1. **[Start]** Generate random population of N individuals.

2. **[Fitness]** Fitness function f(x) for each chromosome is evaluated.

3. **[New Population]** Repeat the following steps to create new population

   i) **[Selection]** Select two parents from the population according to their fitness.

   ii) **[Crossover]** With the crossover probability crossover the parents to form new offspring. If no crossover is performed the offspring is resulted as parents.

   iii) **[Mutation]** With the mutation probability mutate the offspring at each locus.

   iv) **[Accept]** Place new offspring in the population.

4. **[Replace]** Use new generated population for the next iteration.
5. **[Test]** If the termination condition is satisfied, return the best solution.
6. **[Result]** Sort the fitness value in descending order, the lower value are identified as outliers.
7. **[Loop]** Go to step 2 for next iteration.

3.1 Genetic Operators

In Genetic algorithm the representation of genes as chromosomes is done by encoding. The primary steps that are involved in the genetic algorithm are initialization, selection, reproduction (crossover and mutation) and termination.

The first step in the genetic algorithm is the representation of the chromosomes in the problem space with suitable encoding techniques. Various techniques like binary encoding, value encoding and permutation encoding are available.

3.1.1 Fitness function In each problem the fitness function is formulated in way to solve it by genetic algorithm. A fitness function is a problem dependent objective function that quantifies the optimality of an individual in a chromosome.

3.1.2 Selection Parent chromosomes are selected from the problem space using some standard selection mechanisms like Roulette Wheel selection, Tournament selection, Rank based selection, Truncate selection and Boltzmann selection. The result of all these selection mechanisms is to produce the best individual (chromosome) from the search space for the next iteration. The worst chromosomes are replaced by the best individuals in the next iteration which has the lowest probability.

3.1.3 Recombination The genetic operators of GA include reproduction, crossover and mutation are commonly applied to problems of GA.

3.1.4 Reproduction This operator is applied to an individual yields an offspring that is identical as the parent chromosome. There is no change in the genetic traits of the individual that is to be considered for the next generation.

3.1.5 Crossover This operator is also called as reproduction or recombination operator which is the primary operator which helps in generating the new offspring for the next generation. Generated offspring will not be identical with any of its parents. Every pair of individuals is not used in crossover. Generally single point and two point crossover is usually performed. Crossover is done with two parent individuals to form a new offspring for the further generation. The new offspring is produced with the combination of the parental traits. For the parent chromosomes in a single point crossover parent 1 and parent 2 are given below.

Parent 1 : 1111000011110000

Parent 2 : 1100110011001100

After crossover the resultant offspring will be

Offspring 1 : 1111000**11001100**

Offspring 2 : **11001100**11110000

In two point crossover, two cut points are chosen in both the parents and the offspring's are produced by exchanging the genetic materials as segments between the cut points. Suppose the crossover points randomly occur after the fourth and the thirteenth bit of the parent chromosome, then the offspring produced after the two point crossover are:

Offspring 1 :1111**11001100**0000

Offspring 2 :**1100**000011111**1100**

Similarly, Multipoint crossover is also used in which several crossover points are used for exchanging the genetic materials.

3.1.6 Mutation : New genetic traits are included into the existing individuals in the mutation operator. The genes value is randomly changed with in the chromosome. . Mutation is done to have diversity among chromosomes without changing the characteristics of the parent.

In the single point mutation, the gene is randomly chosen and it is mutated to produce the offspring.

Parent : 1111000**0**

Offspring : 1111000**1**

The last bit of the parent chromosome is mutated to generate the new offspring.

In the Multi point crossover any number of genes are randomly selected from the parent chromosome and mutated. The offspring produced after Multi point crossover is:

Parent : 11**1**10**0**0**0**

Offspring: 11**0**10**1**0**1**

After performing all the genetic operators in the dataset, the fitness value obtained after last iteration will be of optimum value. Obtained value is then sorted in the reverse order from which lower fitness value can calculated easily. The gene with lowest fitness value is calculated as an outlier.

4. EXPERIMENTAL RESULTS

The evaluation of our algorithm is done by comprehensive performance study. In this section the experimental results and the performance of the algorithm is shown. The dataset was obtained from standard UCI Machine Learning Repository and synthetic datasets.

We had evaluated the proposed algorithm for the ability to find the number of outliers present the dataset. We had used Breast cancer dataset which consist of 699 data's with 10 attributes each. Yeast dataset contains 1484 data's with 8 features and the last dataset is Liver disorder dataset which has 345 data's with 7 features. The first Dataset contains 241, the second contains 470 and the third data set 78 data as outliers. The experiment also identified the outliers accordingly and is listed in the Table.1. The algorithm is run with above mentioned dataset and the result is summarized as follows.

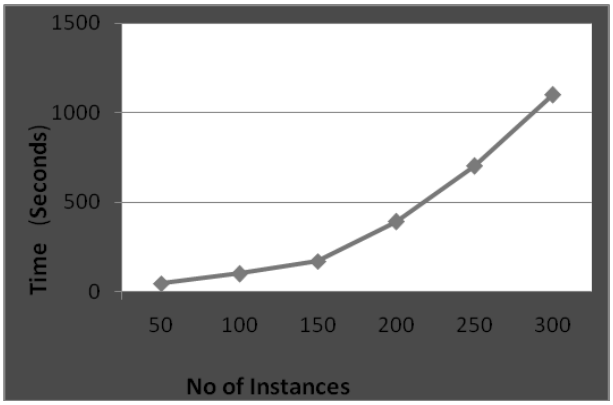
Table.1. Experimental results of various datasets used

Dataset	Data size	Number of attributes	Number of Outliers Detected
Wisconsin Breast Cancer	699	10	238
Yeast	1484	8	462
Liver Disorders	345	7	78

The Efficiency of our experiment is also done. Here we consider a set of dataset and the influence of factors such as size of the dataset and the number of outliers identified and the similar factors are experimentally studied with the same dataset.

The efficiency of our experiment is show in the fig 1. in which the time efficiency is calculated based on the number of instances used and time consumed for the particular instance is shown.

Fig.1. Efficiency of GA with no of Instances



In our experiment we had also calculated the number of outliers obtained from particular number of instances used in our experiment. Number of outliers obtained is shown in the fig 2. Which shows the number of outliers obtained in our experiment.

4.1 Impact of Variation of Crossover Operators

Here , the influence of various crossover operator on the proposed algorithm is studied. Generally one point and two point crossovers are applied and the results are Examined. The impact of one point crossover and two point crossovers on the profit of the proposed algorithm is Listed in the table.2.

Table.2. Effect of Various crossover operators on GA

No of Executions	Single point Crossover	Two point Crossover
1	368767	368739
2	352543	376700
3	300320	367938
4	367610	368601
5	333974	341777
6	355391	374541
7	293888	307481
8	318918	319498
9	393925	389609

Fig.2. Results of Outliers Identified

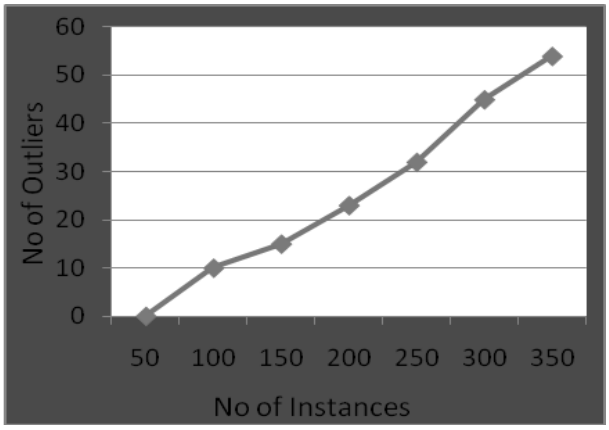
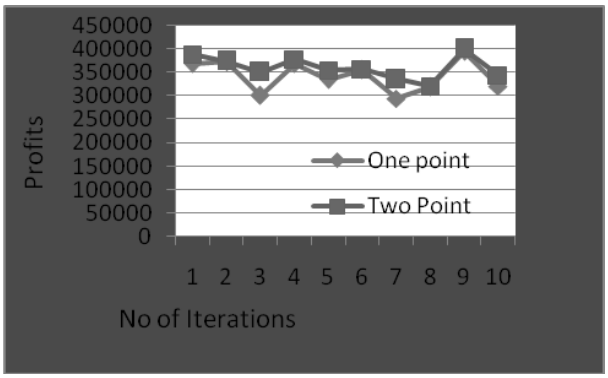


Fig.3. Effect of Crossover Operator



5. CONCLUSION AND FUTURE WORK

In this paper, we proposed an new algorithm for outlier detection using genetic algorithm. The experimental shows that the proposed algorithm is good in calculating the number of outliers in a particular period of time. The future work is done by implementing the proposed work for more dataset of various types with necessary changes in the algorithm to make it more efficient. It has also be planned to implement the proposed system in distributed environment, to improve the processing speed and performance of the algorithm.

6. REFERENCES

[1] Abe,n., Zadrozny., Langford,j: 2006 Outlier Detection by active Learning. SIGKKD-USA

[2] Charu C. Aggarwal and Philip S. Yu. 2001 Outlier detection for high Dimensional data.

[3] E.M.Knorr and R.T.Ng. 1998. Algorithms for mining Distance-Based Outliers in Large DataSets. In Proc-VLDB,pp.392-403

[4] M.M. Breunig , H.P.Kriegel, R.R.Ng, and J.Sander.2000 LOF : Identifying Density – Based Local Outliers. In Proc. SIGMOD conf.pp 93-104

[5] Angiulli, F.Basta, S., Pizzuti 2006. Distance- Based Detection and Prediction of Outliers. IEEE Transactions on Knowledge and Data Engineering.

[6] Provost, F., Fawcett, 2001. Robust Classification of Imprecise environments Machine Learning 42,203-231.

[7] Z.Michalewicz. 1996. Genetic algorithm and Data Structures- Evolution Programs. NY: Springer-verlag.

[8] S.Forrest. " Genetic Algorithms". 1996, ACM Computer society,sum., vol.28.,pp 77 – 80.

[9] W, Banzhar, P.Nordin, R.Keller, and E Francone 1998- Genetic Programming on the automatic evolution of computer program and its applications. Morgan Kaufmmn Publishers.

[10] David E. Goldberg -2005- Genetic algorithm in Search, Optimization and Machine Learning.

[11] C.Aggarwal and P.Yu -2001- Outlier detection for high dimensional data" In Proceedings of the ACM SIGMOD International Conference of management of data , volume 30, issue 2, pages 37-46.

[12] M.Breunig, H.P.Kriegel R.Ng and J.Sander, 2000 " LOF: Identifying Density Based Local Outliers". In Proceedings of the 2000 ACM SIGMOD International Conference of Management of Data Pages 93-104.

[13] M.Brito, E.Chavez, A.Quiroz and J.Yukich-1997- " Connectivity of the mutual K-Nearest Neighbor Graph in Clustering and Outlier Detection". Statistics and Probability Letters, volume 35, Issue 1, Pages 33-42.

[14] Z.He, X.Xu and S.Deng, June 2003 " Discovery Cluster based Local Outliers " Pattern Recognition Letters, Volume 24, Issue 9-10, Pages 1641-1650

[15] V.Hautamaki, I.Karkkainen and P.Franti, -2004- " Outlier Detection Using K-Nearest Neighbor Graph". In Proceedings of the international Conference on Pattern Recognition". Volume 3, Pages 430 – 433.

[16] M.Jaing, S.Tseng and C.Su, -2001- " Two Phase Clustering Process for Outlier Detection".

[17] Knorr, E.M., Ng,R.T.-1997- A Unified Notion of Outliers : Properties and Computation . In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining Proceedings .,PP 219-222.

[18] UCI Repository www.ics.uci.edu/mlearn/ MLRepositary .html

[19] V.Bamett and T.Lewis. Outliers in Statistical Data.