ISSN:0975-9646

Genetic Algorithm for Outlier Detection

Maria Afzal
Department of Computer Science
Jamia Hamdard
New Delhi

S M Arif Ashraf
Department of Computer Science
Jamia Hamdard
New Delhi

Abstract-The primary objective of the outlier detection is to find the data that are noticeably different from other data in a particular dataset. Outlier detection has been a key research area in data mining since it has wide application in various areas like Intrusion Detection System, Fraud Detection, Bio Science and Drugs Research. Many approaches have been proposed for efficient and accurate outlier detection.

In this paper authors propose a soft computing technique; Genetic Algorithms for outlier detection.

The proposed algorithm was tested on some standard dataset for performance evaluation .The result was found to be very encouraging.

Keywords: Genetic Algorithm, Outliers, Fitness function.

INTRODUCTION

Outlier detection is an important field in data mining which is the discovery of data that deviates a lot from other data patterns. D. Hawkins, define outliers as: *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism* [1]. An outlier may be considered as an abnormal data object. These outliers are also known as *deviants* or *anomalies.* Outlier detection have role in significant fields- Credit card frauds, Medical Diagnosis, Earth Science etc. Outlier detection techniques are basically categorized into three groups: *Supervised methods, unsupervised methods and Semi-Supervised methods. Supervised methods* use trained data to detect outlier. An unsupervised outlier detection method does not follow any trained set of data. The central idea is to find cluster first, and then the data object not belonging to any cluster are detected as outlier.

In *semi- Supervised methods,* you may encounter cases where only a small set of the normal or outlier objects are labelled but most of the data is unlabelled [2].

Researchers are still doing research on determining outliers using Genetic Algorithms (GAs). GA is an adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and survival of the fittest [3].

RELATED WORK

Various techniques have been proposed for outlier detection and most of these work basically used statics measurement. These works involves density based, distance based, distribution based and cluster based approaches [2] [5].

In distance based approaches detection is done by measuring the distance of data points with a centre data point. If a data point has larger distance than a specified threshold then that data point will be considered as outlier. This kind of approach was proposed which used distance based and clustering based approach together; in this approach noise was measured in noise clustering, based on preservation of hyper volume of the feature space [6].

A distance based approach was also proposed that utilized particle swarm optimization and clustering based approach together. In this approach, ratio of number of particles under threshold range k, to the ratio of threshold distance, r can be used as a measure to detect outliers [7].

Distribution based methods depends on data distribution, where distributions like Normal Distribution, Poisson distribution, etc was used. The major problem of this method is to find whether a data is following normal distribution, gamma distribution or any other kind of distribution methods [5].

Genetic Algorithm for the purpose of outlier detection has not been used significantly. Detecting outliers one by one in previous approaches is time taking as well as costly. An approach was proposed where it is possible to detect all possible outliers at a same time in a data, in linear regression using Genetic Algorithm. This approach overcomes the problem of single outlier detection [8].

PROPOSED WORK

In proposed work we use Genetic Algorithms for outlier detection. The reason to use GA apart from its simplicity is its optimizing nature. Genetic algorithms being adaptive heuristics and robust in nature can be applied in problems of any domain with slight context based modification.

A simple GA based outlier detection techniques in pseudo form can be written as –

```
Input: A dataset for outlier detection
Output: Outlier with lowest fitness value
1. Generate random population of N individuals.
2. 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 is identified as outliers.
7. [Loop] Go to step 2 for next

The crossover which we used in our algorithm is a GA is a variant of two parent uniform order based crossover (UOX). At each position in the input string, a value is randomly chosen from one of the two parents to form the new children. Duplicates value are removed from by selecting a new unique random numbers. For mutation, we randomly choose a point in the chromosome and exchange it with a unique random number value. In addition, we sort the point indices in chromosome for improved performance.

The Fitness Function is defined as

f(x)=a/(r*k)+ k/r + k/(n-k)

Where

n –>the size of the dataset

a->constants

The Value of "r" should be large enough to include some neighbouring point to be valid else some valid point may wrongly be detected as outlier.

K->is the no of points considered as the valid points (not outlier)

The ratio k/r can be used as a measure to detect outliers.

I. GENETIC ALGORITHM

Genetic Algorithms (GAs) are search methods based on the principle of Survival of fittest. GAs considers search variables as strings of finite lengths. These strings are referred as chromosomes, and individual elements or characters of these strings are known as genes and the value of genes are known as allele [4]. GA involves operators like crossover, mutation and selection.

A. Population Generation

The initial step of GA is population generation. A population is a collection of individuals, and can have number of individuals. The individuals in a population are known as chromosomes. The population size depends on the complexity of problem [3]. The individual chromosomes can be of binary or any other type of strings depending upon the nature of problem. In case of binary coded chromosome each bit is initialized to a random 0 or 1. In most of the cases the initial population is randomly chosen. Figure1 shows a simple population of four chromosomes, chromosomes are of 8 bits here.

Population	Chromosome 1	1 0 1 0 0 0 1 1
	Chromosome 2	1 0 1 0 1 0 1 0
	Chromosome 3	1 1 1 0 0 1 1 1
	Chromosome 3	1 1 1 0 0 0 1 0

Figure 1: Population

B. Selection

Selection is the process to choose two parents from the population for crossing. The purpose of selection is to choose fitter parents from the population so that their offspring would have higher fitness. In GA chromosomes with higher fitness value are selected as parents. Fitness values are decided by fitness function. There is various selection methods used in GA depending upon the nature of problem and data, some of them are: Roulette Wheel Selection, Random Selection, Rank Selection, Tournament Selection, Boltzmann Selection and Stochastic Universal Sampling.

C. Crossover

After the selection process, mating pool will have parent chromosomes with better fitness value. Crossing two randomly chosen parents from mating pool will produce better offspring. Cross over process can be summarized as:

1) Choose a pair of chromosomes randomly from the mating pool.
2) Select a cross site randomly along the length of string.
3) The position values are swapped between the two strings following the cross site.

Following are some crossover techniques:

Single-point crossover

In single point crossover two parents are cut once at corresponding points and the sections after the cut are exchanged. Single point crossover is illustrated in figure 2 where a single crossover point is selected randomly and the bits next to the crossover point are exchanged to produce children.

Parent 1	1 0 1 0 0	**0 1 1**
Parent 2	1 0 1 0 1	**0 1 0**

Child 1	1 0 1 0 0	**0 1 0**
Child 2	1 0 1 0 1	**0 1 1**

Figure 2: Single point crossover

Two-Point Crossover

In Two-point crossover two cross over points are chosen randomly in parent chromosomes and the bits between the crossovers are swapped to produce the children. Figure 2 shows two point cross over, cross over points are shown by dotted lines, and the contents between these points are exchanged among parents to produce new children.

There are some other types of crossover techniques as, Multi-Point crossover, Uniform crossover, three parent crossover, Shuffle crossover, Precedence preservation crossover, partially matched crossover. [3].

Parent 1	1 0	**1 0 0**	0 1 1
Parent 2	1 0	**1 0 1**	0 1 0

Child 1	1 0	**1 0 1**	0 1 1
Child 2	1 0	**1 0 0**	1 1 1

Figure 3: Two point crossover

D. Mutation

After cross over mutation plays the role of recovering lost genetic material. Mutation helps to introduce diversity in child chromosomes. It introduces new genetic structures in the population by modifying some of its building blocks. In case of binary representation, a simple mutation can be done by inverting the value of each gene with a small probability. There are many different forms of mutation for different kind of mutation. In case of binary representation, usually probability is considered about 1/L, where L is the length of the chromosomes. Mutation of a bit involves flipping a bit, changing 0 to 1 and vice-versa. Figure 4 shows that bit at position 4 in child was flipped, 0 was replaced by 1.

Child Chromosome	1 0 1 **0** 1 0 1 1
Mutated chromosome	1 0 1 **1** 1 0 1 1

Figure 4: Mutation

There are some other types of mutation techniques like Flipping, Interchanging, Reversing and Flipping.[3].

RESULTS

We tested our algorithms on some standard dataset for performance evaluation. The dataset we used for testing are breast-Cancer, supermarket and diabetes. These all are built-in dataset which are available in the WEKA (Waikato Environment of Knowledge Analysis version 3.6.12, University of Waikato, Hamilton New Zealand) software.

Following are the characteristics of the used data set.

Dataset Name	Instance	No. of Attributes
Breast-Cancer	286	10
Diabetes	768	9
Supermarket	4627	217

The algorithms were evaluated for the ability to find the number of outlier present in the dataset. Apart from performance evaluation we also tested the efficiency. This is done on ability of algorithms of finding the new outliers with an increase in dataset.

The results are shown as follows:

Dataset Name	Instance	No. of Outliers
Breast-Cancer	286	55
Diabetes	768	278
Supermarket	4627	1651

CONCLUSION AND FUTURE

The purpose of this paper was to use GA to detect outliers. Results obtained through the experiments shows positive encouragement. It was also observed that with an increase in dataset, the algorithm still performs well. To make it more efficient it will be desirable to check it on more diverse dataset having large no of attributes and records.

As cloud computing is being adopted at a rapid rate by both large and small scale companies; therefore it will be a future task to implement this algorithm on a cloud computing environment.

We will also look for the use of this algorithm for big data analysis on “Hadoop” Platform as well.

REFERENCES

[1]. D.Hawkins, “Identification of outliers”. Chapman and Hall, London 1980

[2]. Jaiwei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques, Elsevier, Morgan Kaufmann.

[3]. S.N. Sivanandam, S.N. Deepa, “Principles of soft computing”, Wiley.

[4]. Mitchell Melanie, “An Introduction to Genetic Algorithms”, The MIT press.

[5]. M. O. Mansur, Mohd. Noor Md. Sap, “Outlier Detection Techniques in Data Mining : A Research Perspective”, Proceedings of the postgraduate annual research seminar 2005, Universiti Teknologi Malaysia.

[6]. Frank Rehm , Frank Klawonn, Rudolf Kruse, “A novel approach to noise clustering for outlier detection”, Journal of soft computing (Springer), 489–494 vol 11 (2007).

[7]. Ammar W Mohemmed, Mengjie Zhang, Will Browne, “Particle Swarm Optimization for Otlier Detection”, Technical report: ECSTR10-07, Victoria university of wellington.

[8]. J.Tolvi, “Genetic algorithms for outlier detection and variable selection in linear regression models”, Soft Computing 8 (2004) 527–533 _ Springer-Verlag 2003