A New Approach to Outlier Detection

Lancang Yang[1], Bing Shi[1], Xueqin Zhang[2], and Lei Qiao[1]

[1]School of Computer Science and Technology
Shandong University, Jinan 250061, P.R. China
yanglancang@163.com
[2]School of Information Science and Engineering
Jinan University, Jinan 250022, P.R. China
zhangxueqin_2005@163.com

Abstract. For many data mining applications, finding the rare instances or the outliers is more interesting than finding the common patterns. At present, many automated outlier detection methods are available, however, most of those are limited by assumptions of a distribution or require upper and lower predefined boundaries in which the data should exist. Whereas a distribution is often unknown, and enough information may not exist about a set of data to be able to determine reliable upper and lower boundaries. For these cases, a new dissimilarity function was defined, which can be viewed as fitness function of genetic algorithm, and a GA-based outlier detection method was formed in this paper. This method allows for detection of multiple outliers, not just one at a time. The illustrations show that the improved approach can automatically detect outliers, and performs better than GLOF approach.

1 Introduction

Outlier detection in large data sets is an active research field in data mining, it has many applications in all those domains that can lead to illegal or abnormal behavior, such as fraud detection[1], network intrusion detection, insurance fraud, medical diagnosis, marketing, or customer segmentation, etc. Outlier detection has become an important branch of data mining[2].

There exist a large number of outlier detection methods in the literature. Traditionally, these can be categorized into three approaches: the statistical approach, the distance-based approach[3], and the deviation-based approach. But many of those are limited by assumptions of a distribution or limited in being able to detect only single outliers. If there is a known distribution for the data, then using that distribution can aid in finding outliers. Often, a distribution is not known, or the experimenter does not want to make an assumption about a certain distribution[4].

Genetic Algorithm (GA) was introduced in the mid 1970s by John Holland and his colleagues and students at the University of Michigan. GA is inspired by the principles of genetics and evolution, and mimics the reproduction behavior observed in biological populations. GA employs the principal of "survival of the fittest" in its search process to select and generate individuals that are adapted to

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 615–620, 2007.

their environment. Therefore, over a number of generations, desirable traits will evolve and remain in the genome composition of the population over traits with weaker undesirable characteristics. GA is well suited to and has been extensively applied to solve complex design optimization problems because it can handle both discrete and continuous variables and nonlinear objective and constrain functions without requiring gradient information[5].

In this paper, a new dissimilarity function had been defined, which can be viewed as the fitness function of genetic algorithm, and then genetic algorithm was used for outlier detection. This method can detect multiple outliers at a time, and what we should do is nothing but specifying the number of outliers we want. Extensive experiment results on synthetic and real data revealed that the GA-based outlier detection approach can detect outliers automatically and efficiently.

2 GA-Based Outlier Detection

2.1 Outlier Detection

Outlier mining can be described as: Given a set of n data points or objects, and k, the expected number of outliers, find the top k objects that are considerably dissimilar, exceptional, inconsistent with respect to the remaining data. The outlier mining problem can be viewed as two subproblems:

a. Define what data can be considered as inconsistent or exceptional in a given data set; and

b. Find an efficient method to mine the outliers so defined.

The computer-based outlier detection methods can be categorized into three approaches: statistical approach, distance-based approach, and deviation analysis approach. Notice also that many clustering algorithms discard outliers as noise. However, they can be modified to have outlier detection as a byproduct of their execution[6].

2.2 Genetic Algorithm

Genetic algorithm is a stochastic search technique that guides a population of solutions towards an optimum using the principles of evolution and natural genetics. In recent years, genetic algorithm has become a popular optimization tool for many areas of research, including the field of data mining.

The algorithm starts with a randomly generated initial population consisting of sets of "chromosomes" that represent the solution of the problem. These are evaluated for the fitness function or one of the objective functions, and then selected according to their fitness value[7]. To perform its optimization-like process, the GA employs three operators to propagate its population from one generation to another. The first operator is the selection operator,which mimics the principal of "Survival of the Fittest". The second operator is the crossover operator, which mimics mating in biological populations. The crossover operator

propagates features of good surviving designs from the current population into the future population, which will have better fitness value on average. The last operator is the mutation operator, which promotes diversity in population characteristics. The mutation operator allows for global search of the design space and prevents the algorithm from getting trapped in local minima[5].

2.3 Genetic Algorithm Operations

In this paper, for the given set of objects located in the space, genetic algorithm was used to detect the outliers. There are five primary elements in the genetic algorithm, and the parameter setting of genetic algorithm was shown as following in details:

In the approach, the number of outliers was specified firstly, and a random population of chromosomes was created representing the solution space. Each member of this random population represents a different possible solution for the genetic algorithm. The genetic algorithm proceeded to find the optimal solution through several generations.

A. Parameter Encoding

The population representing the solution space consists of many chromosomes. Each chromosome consists of k genes, where k is the number of outliers given. These genes represent the serial number of objects in the data set, which are viewed as outliers. A chromosome can have any combination of these gene values.

B. Fitness Function

The genes, which represent the serial number of outliers, are updated with each new population created. The random population is sorted based on the least fitness. The top chromosome with the least fitness is considered to be the elite chromosome within the population.

The fitness function used in the approach is the dissimilarity function, which can be any function that, if given a set of objects, returns a low value if the objects are similar to one another. The greater the dissimilarity among the objects, the higher the value returned by the function[6].

In our research, a new dissimilarity function was defined, which can be used to evaluate the degree of the outliers. According to the function, the lower the value returned after removing some objects, the greater the degree of these objects being outliers is.

Definition 1. Given a set S of n objects, S' is a subset of S, which contains k objects and denotes the set of outliers in S. Let S"be the complement of S'. A dissimilarity function can be expressed as:

$$\frac{1}{n-k}\sum_{i=1}^{n-k}(x_i-\overline{x})^2 \tag{1}$$

Where n denotes the number of objects contained in set S, k denotes the number of objects contained in set S', namely, the number of outliers, x_i denotes the object in the set S", and $\overline{x}$ denotes the mean value of all objects in the set S".

C. Selection Operator

The Selection operator that mimics the principal of "Survival of the Fittest" selects the chromosomes with the least fitness as the elite chromosome within the population. Among many selection operators, stochastic tournament selection model is used in our approach.

D. Crossover Operator

The Crossover operator that mimics mating in biological populations propagates features of good surviving designs from the current population into the future population, which will have better fitness value on average. Among many crossover operators, two-point crossover model is used in our approach.

E. Mutation Operator

Mutation operator that promotes diversity in population characteristics allows for global search of the design space and prevents the algorithm from getting trapped in local minima. Among many mutation operators, what we used is basic mutation operator in our approach.

3 Experiments

A comprehensive performance study has been conducted to evaluate our algorithm. Our algorithm was implemented in VC++6.0. We ran our algorithm on some real life data sets (Wisconsin breast cancer data set and Boston housing data set) obtained from the UCI Machine Learning Repository[8], and demonstrated the effectiveness of our method against other algorithms. Experiment results are shown as follows.

3.1 Wisconsin Breast Cancer Data Set

The Wisconsin breast cancer data set has 699 instances with nine attributes. Each record is labeled as benign (458 or 65.5%) or malignant (241 or 34.5%). We followed the experimental technique of Harkins et a1.[9] by removing some of the malignant records to form a very unbalanced distribution, the resultant data

Table 1. Comparison between GLOF and our approach. Where N denotes the total number of "outliers" identified, $N1$ denotes the number of "true" outliers (malignant), $N2$ denotes the number of "pseudo" outliers (benign), and *gen* denotes the number of generations in genetic algorithm.

Approachs	Parameter	N	$N1$	$N2$
GLOF ($\|\xi - \mu\| \geq 2 \bullet \sigma$)	x_1 =-1,x_2=1	39	26	13
	x_1=1,x_2=1	39	26	13
	x_1=1,x_2=-1	40	35	5
OURS ($k = 39$)	gen=2000	39	31	8
	gen=3000	39	33	6
	gen=4000	39	35	4

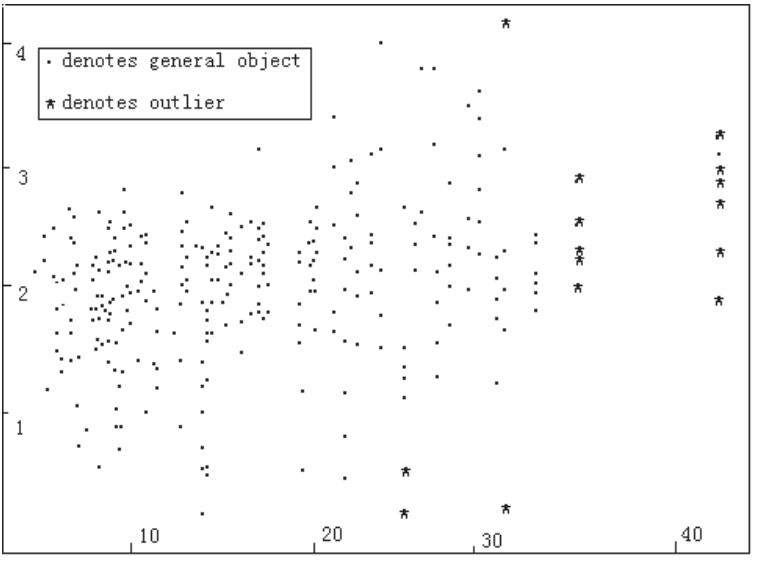

Fig. 1. The test result on boston housing data

set had 39 (8%) malignant records and 444 (92%) benign records. With the data set, we compared our approach with the GLOF approach[10]. The comparison between two approaches is shown in Table 1.

3.2 Boston Housing Data Set

The Boston housing data set was taken from the StatLib library which is maintained at Carnegie Mellon University. This data set has 506 instances with 13 continuous attributes. For the convenience of observation, two attributes NOX (nitric oxides concentration) and RM (average number of rooms per dwelling) were extracted to form a test data set, and the number of outliers was specified as 15. The test result is shown in Fig. 1.

4 Conclusions

In this paper, a new dissimilarity function was defined, which was viewed as the fitness function of genetic algorithm, and then genetic algorithm was used for outlier detection. In the approach, what we should do is nothing but specifying the number of outliers we want. Extensive experiments on synthetic and real data showed that the GA-based outlier detection approach can automatically detect outliers, and that the improved approach performs better than GLOF approach.

References

1. Park, L.J.: Learning of neural networks for fraud detection based on a partial area under curve. In: Advances in Neural Networks. Lecture Notes in Computer Science, Vol. 3497. Springer-Verlag, Heidelberg (2005) 922-927
2. Angiulli, F., Basta, S., Pizzuti, C.: Distance-Based Detection and Prediction of Outliers. IEEE Trans. Knowl. Data. Eng. 18 (2006) 145-160
3. Guha, R., Dutta, D., Jurs, P.C., Chen, T.: R-NN curves: An intuitive approach to outlier detection using a distance based method. J. Chem. Inf. Model. 46 (2006) 1713-1722
4. Amidan, B.G., Ferryman, T.A., Cooley, S.K.: Data Outlier Detection using the Chebyshev Theorem. In: Aerospace. IEEE AEROSPACE CONFERENCE PROCEEDINGS, IEEE, Piscataway NJ USA (2005) 3814 - 3819
5. Hassan, R., Cohanim, B., Weck, O., Venter, G.: A COMPARISON OF PARTICLE SWARM OPTIMIZATION AND THE GENETIC ALGORITHM. In: AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials. Collection of Technical Papers, Vol. 2. American Institute of Aeronautics and Astronautics, Reston (2005) 1138-1150
6. Han, J.W., Kamber, M.: Data Mining Concepts and Technique. San Francisco: Morgan Kaufmann, (2001)
7. Jerald, J., Asokan, P., Saravanan, R., Rani, A. Delphin Carolina : Simultaneous scheduling of parts and automated guided vehicles in an FMS environment using adaptive genetic algorithm. Int. J. Adv. Manuf.Technol. 29 (2006) 584-589
8. Merz C.J., Merphy P.: UCI repository of machine learning databases. URL: http://www.ics.uci.edu/mleaml MLRRepository.htm1. (1996)
9. Harkins, S., He, H., Willams, G.J., Baster, R.A.: Outlier detection using replicator neural networks. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds.): Data Warehousing and Knowledge Discovery. Lecture Notes in Computer Science, Vol. 2454. Springer-Verlag, Berlin (2002) 170-180
10. Jiang, S.Y., Li, Q.H., Li, K.L., Wang, H., Meng, Z.L.: GLOF: A NEW APPROACH FOR MINING LOCAL OUTLIER. In: Machine Learning and Cybernetics. Int. Conf. Mach. Learn. Cybern., Vol. 1. Institute of Electrical and Electronics Engineers Inc. (2003) 157-162