International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Implementing Outlier Detection using Greedy Based Information Theoretic Algorithms and its Comparison with PSO and ACO Optimization Techniques

Amandeep Kaur†* and Kamaljit Kaur†

†Dept. of C.S.E., SGGSWU, Fatehgarh Sahib, Punjab, India

Accepted 10 Jan 2015, Available online 01 Feb 2015, **Vol.5, No.1 (Feb 2015)**

Abstract

Outlier is defined as an observation that deviates too much from other observations. The identification of outliers can lead to the discovery of useful and meaningful knowledge. Outlier detection has been extensively studied in the past decades. However, most existing research focuses on the algorithm based on special background, compared with outlier detection approach is still rare. Most sophisticated methods in data mining address this problem to some extent, but not fully, and can be improved by addressing the problem more directly. The identification of outliers can lead to the discovery of unexpected knowledge in areas such as credit card fraud detection, calling card fraud detection, discovering criminal behaviors, discovering computer intrusion, etc. The greedy approach to develop two efficient algorithms , ITB-SS,ITB-SP that provide practical solutions to the optimization problem for outlier detection. For more optimized data in this paper a new work, which is used both algorithms with genetic algorithm which provide more accurate results as compare to previous results.

Keywords: *Information theoretic algorithms, Genetic algorithm, Ant colony Optimization, Particle Swarm Optimization.*

1. Introduction

Over the years, the technique of outlier detection has obtained the widespread concern with the increasing application of data mining. It can be used in many applications such as intrusion detection, fraud detection, medical examination and so on. The key point of outlier detection is to define what kind of data is anomalous. There does not exist a uniform definition of outlier since it has been researched. The most accepted definition is given by Hawkins, "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"(Hawkins D *et al*,1980).

An outlier is an observation point that is distant from other observations. An outlier is due to variability in the measurement or it may indicate experimental error the latter are sometimes excluded from the data set (Amandeep Kaur *et al*, 2014).

Outlier detection encompasses aspects of a broad spectrum of techniques. Many techniques employed for detecting outliers are fundamentally identical but with different names chosen by the authors. For eg, authors describe their various approaches as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection or exception mining. Outlier detection is a critical task in many safety critical environments as the outlier indicates abnormal running conditions from which significant performance degradation may well result, such as an aircraft engine rotation defect or a flow problem in a pipeline (Hodge *et al*, 2004).

2. Related Work

Outlier detection for categorical data sets. This problem is especially challenging because of the difficulty of defining a meaningful similarity measure for categorical data. A formal definition of outliers and an optimization model of outlier detection, via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model, two practical 1-parameter outlier detection methods, named ITB-SS and ITB-SP, which require no user-defined parameters for deciding whether an object is an outlier. Users need only provide the number of outliers they want to detect. Experimental results show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail (W. Lee *et al*, 2011). Distributed method for detecting distance-based outliers in very large data sets. This approach is based

*Corresponding author: **Amandeep Kaur**

on the concept of outlier detection solving set, which is a small subset of the data set that can be also employed for predicting novel outliers. The method exploits parallel computation in order to obtain vast time savings. Indeed, beyond preserving the correctness of the result, the proposed schema exhibits excellent performances. From the theoretical point of view, for common settings, the temporal cost of our algorithm is expected to be at least three orders of magnitude faster than the classical nested-loop like approach to detect outliers. Experimental results show that the algorithm is efficient and that its running time scales quite well for an increasing number of nodes. A variant of the basic strategy which reduces the amount of data to be transferred in order to improve both the communication cost and the overall runtime (Fabrizio Angiulli *et al*, 2013).

Stephen et al. propose modifications to a simple, but quadratic, algorithm for finding distance-based outliers, and show that it achieves near linear time scaling allowing it to be applied to real data sets with millions of examples and many features (Stephen D Bay *et al*, 2003).

3. Information theoretic-based approach

Several information-theoretic methods have been proposed in the literature. For anomaly detection in audit data sets, (W. Lee *et al*, 2011) present a series of information-theoretic measures, i.e., entropy, conditional entropy, relative conditional entropy, and information gain, to identify outliers in the univariate audit data set, where the attribute relationship does not need to be considered. The work of (Z. He *et al*,2005) employs entropy to measure the disorder of a data set with the outliers removed. In these methods, heuristic local search is used to minimize the objective function. The methods proposed in (Barnett *et al*, 1994) and (K. Das *et al*,2008) set a threshold of mutual information and obtain a set of dependent attribute pairs. Based on this set, an outlier factor for each individual object is defined. In general, information-theoretic methods focus either on a single entropy-like measurement or on mutual information, and require expensive estimation of the joint probability distribution when the data set is shrunk following elimination of certain outliers.

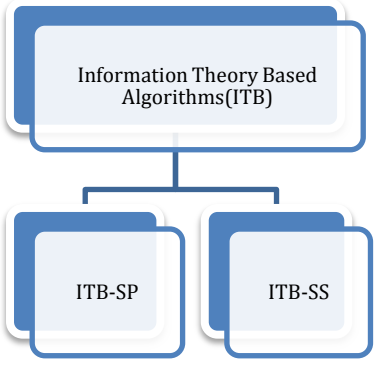


Figure 1: Information Theory based Algorithms

The two algorithms for outlier detection of Information-Theory- Based approach.One is named ITB-SS for Information-Theory- Based Step-by-Step (or SS for short), the other one is named ITB-SP for Information-Theory-Based Single-Pass (or SP for short). Both algorithms detect outliers one by one (Amandeep Kaur *et al*, 2014).

In both algorithms, search is conducted only within the anomaly candidate set AS.

3.1 Information-Theory-Based Single-Pass Algorithm

In Single Pass, the outlier factors are computed only once, and the o objects with the largest OF(ox) values are identified as outliers. In both algorithms, search is conducted only within the anomaly candidate set AS (anomaly candidate set), although this does not make any difference for the algorithm ITB-SP since the initialization of AS requires computation of the outlier factors of all the objects. In ITB-SP, the attribute weights $w_x(y_i)(1\leq i\leq m)$, the outlier Factor $OF(x_i)$ of all the objects, initialization of AS and the heapsort search to find the top-o outlier candidates are computed . The time complexity of ITB-SP O(nm). The upper bound on outliers (UO) is to estimate an upper limit on the number of outliers in a data set.

Algorithm 1. ITB-SP single pass

1: Input: data set X and number of outliers requested o
2: Output: outlier set OS
3: Compute $W_x(y_i)$ for$(1\leq i\leq m)$
4: Set OS =0
5: for i= 1 to n do
6: Compute $OF(x_i)$ and obtain AS
7: end for
8: if o > UO then
9: o= UO
10: else
11: Build OS by searching for the o objects with greatest $OF(x_i)$ in AS using heapsort
12. endif.

3.2 Information-Theory-Based Step-by-Step Algorithm

At each step of Step-by-Step, the object with the largest $OF(x_o)$ is identified as an outlier and is removed from the data set. Following this removal, the outlier factor OF(x) is updated for all the remaining objects. The process repeats until o objects have been removed.

Algorithm 2. ITB-SS Step-by-Step

1: Input: data set X and number of outliers requested o
2: Output: outlier set OS
3: Set OS =0
4: Compute $W_x(y_i)$ for$(1\leq i\leq m)$
5: for i=1 to n do
6: Compute $OF(x_i)$ and obtain AS
7: end for
8: if o > UO then

9: o = UO
10: else
11: for i = 1 to o do
12: Search for the object with greatest OF(x_o) from AS
13: Add x_o to OS and remove it from AS
14: Update all the OF(x) of AS
15: end for
16: end if Algorithm

In this algorithm, Considering that o(UO) is usually larger than n, it is possible to say that the final complexity of ITB-SS is O(om(UO)). Compared with ITB-SP, the time complexity of the ITB-SS method is a little higher (Amandeep Kaur *et al*, 2014).

4. Optimization Techniques

In optimization of a design, the design objective could be simply to minimize the cost of production or to maximize the efficiency of production. An optimization algorithm is a procedure which is executed iteratively by comparing various solutions till an optimum or a satisfactory solution is found.

(1) Genetic Algorithm

Genetic algorithms is the part of the class of evolutionary algorithms (EA), in which solutions are to be generated for optimization problems inheritance, mutation, selection, and crossover (K. F. Man *et al*,1996).

The steps in Genetic Algorithm are

1. **Initialization**: First chromosomes are randomly created. At this step, it is very necessary that the population is diverse. Otherwise, the algorithm does not produce good solutions.
2. **Evaluation**: At this step each chromosome is rated on how well the chromosome solves the problem. A fitness value is then assigned to each chromosome.
3. **Selection**: In this the fittest chromosomes are then selected for propagation into the future generation based on their fitness value.
4. **Recombination**: In this individual chromosomes and pairs of chromosomes are get recombined, modified and then put back into the population.

The evolution starts from a population of randomly generated individuals, and it is an iterative process, and the population in each iteration called a generation. In each step of generation, the fitness of every individual in the population is calculated and the fitness is usually the value of the objective function in the optimization problem being solved. The more fit individuals are selected from the current population, then each individual's genome is changed which is then recombined and get mutated to form a new generation. The new generation of resulting solutions is then used in the next iteration of the algorithm. The algorithm terminates when a maximum number of generations has been produced, or a good fitness level has been reached for the population (K. F. Man *et al*,1996).

A typical genetic algorithm requires

1. A genetic structure of the solution domain,
2. A fitness function to evaluate the solution domain.

(2) Ant Colony Optimization

In this, at initial stage ants moves randomly, and on finding food they return to their colony by laying down pheromone trails. If other ants do find a path, they do not travel randomly, but instead follow the trail, returning and reinforcing it if they find food (Marco Dorigoa *et al*, 2005).

However, the pheromone trail starts to evaporate and thus reduces its attractive strength. The more time taken by the ant to travel down the path and back again, the more time the pheromones get evaporate. A short path gets marched over more frequently, and the pheromone density gets higher on shorter paths other than longer ones. Pheromone evaporation avoids the convergence to a locally optimal solution. If there were no evaporation, the paths chosen by the first ants would tend to be getting excessively attracted to the following ones. So by this the exploration of the solution space would get constrained.

When one ant finds a good (i.e., short) path from the colony to a food source, the remaining ants are more likely to follow that same path, and positive feedback eventually leads to all the ants following a single path (Marco Dorigoa *et al*, 2005).

(3) Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population based stochastic optimization tool inspired by social behaviour of flocks of birds (and schools of fish, etc), as developed by Kennedy and Eberhart in 1995 (Papadimitriou *et al*, 2003). PSO consists of a population of particles that search for a solution within the search domain. Each particle (population member) in the swarm correspond to a solution in a high-dimensional space with four vectors, its current position, best position found so far, the best position found so far by its neighborhood and its velocity and adjusts its position in the search space based on the best position reached by itself (pbest) and its neighbor (gbest) during the search process.

Steps in PSO algorithm can be explained as below:

1. Initialize the swarm by assigning a random position.
2. Estimate the fitness function for each particle.
3. For each individual particle, compare the particle‘s fitness value with its pbest. When the current value is better than the pbest value, then set this as pbest and the current particle‘s position, xi, as pi.

Table 1 Comparison between existing with proposed method for dataset1

Algorithms	ITB-SS	ITB-SP	ITB-SS With GA	ITB-SP With GA
Precision	0.9	0.9	0.95	0.95
Recall	0.8	0.8	0.84	0.84
F-measure	0.847059	0.847059	0.872686	0.872606
Time	0.696004	1.13786	0.363186	0.363186

Table 2 Comparison between proposed method with PSO and ACO

Algorithms	ITB-SS With GA	ITB-SP with GA	ITB-SS with PSO	ITB-SS with ACO	ITB-SP with PSO	ITB-SP with ACO
Precision	0.92	0.91	0.87	0.87	0.87	0.87
Recall	0.83	0.82	0.71	0.71	0.70	0.71
F-measure	0.872686	0.862659	0.781899	0.781899	0.842686	0.781899
Time	0.615798	0.335201	1.12185	1.12185	0.361332	0.361332

4. Identify the particle that has the best fitness value. This fitness function identified as gbest, pg
5. Revise the velocities and positions of all the particles using (1) and (2).
6. Repeat steps 2–5 until a sufficiently good fitness value (Bing Xue *et al*, 2012).

5. Experimental Evaluation and Results

All the implementations for the experiments were made in MATLAB programming language. The proposed method was tested under Microsoft Windows 7 Ultimate operating system. The hardware used in the experiments had 1 GB of RAM and Intel Core2Duo 1.60 GHz processor.

5.1 Parameters

For experimentation various parameters are used which confirms the results. In terms of Precision, Recall, F-measure and accuracy it compares the results with previous work which clears that proposed approach produces the more refined results as compared to previous approach.

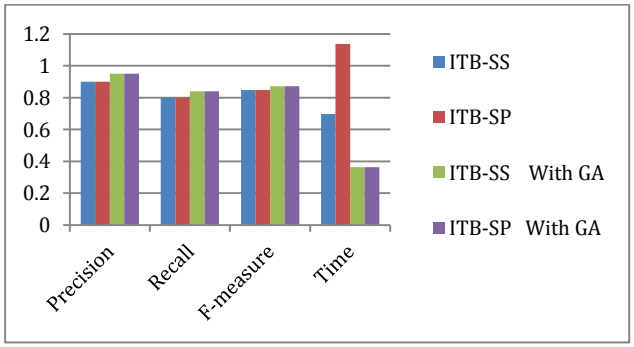
1. **F-measure**: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:
   F-measure = $\frac{2 \times Recall \times Precision}{Recall + Precision}$
2. **Precision:** In simple language precision is the fraction of retrieved documents that are relevant to the search, for a text search on a set of documents precision is the number of correct results divided by the number of all returned results.
   Precision = $\frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$
3. **Recall:** Recall is the fraction of the documents that are relevant to the query that are successfully retrieved, for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.
   Recall = $\frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$
4. **Time:** In this work time parameter is used which display the time of implementation of a particular algorithm. The time parameter represent the time in seconds in this work.

In comparison table 1, after implementation we have results for ITB-SS,ITB-SP and ITB-SS with GA ,ITB-SP with GA. The result is compared with PSO and ACO in next table .In ITB-SS and ITB-SP with GA detect more outliers than PSO and ACO as shown in table. The comparison is based upon four parameters as shown in Table. There are Automobile dataset is used which is named dataset1 in this work. This dataset is taken from UCI Machine Learning Repository.

In table 2 is used to display the result for proposed method and results of other optimization techniques for comparison as PSO and ACO. There are ITB-SS with GA and ITB-SP with GA results are shown in below table which is obtained on various parameters i.e. Precision, Recall and F-measure and Time. The results showed that the value of all the parameter are enhanced after applying both algorithms of proposed work as compare to others.

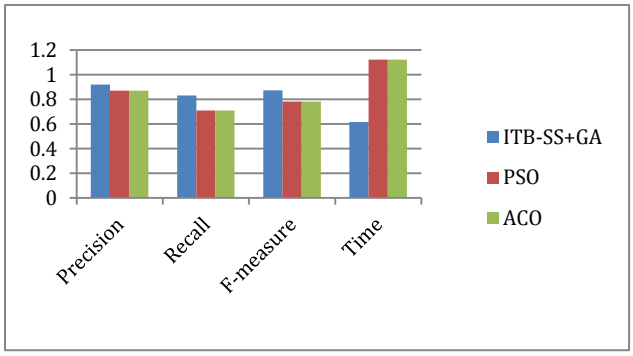
Experimental Graph

The graph for the results obtain on different steps is given below



Graph 1 – Compare ITB-SS and ITB-SP with proposed method

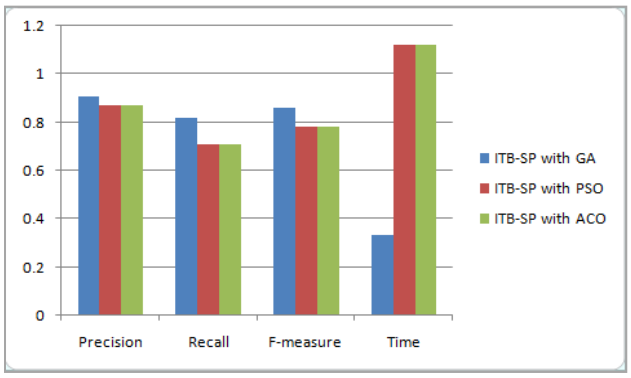
In above graph shown the Information theory algorithms as ITB-SS and ITB-SP with proposed algorithm .In graph there are ITB-SS with GA and ITB-SP with GA have better results as compare to previous algorithms of information Theory.



Graph 2 - ITB-SS with GA and compare with ACO and PSO

Now After implementation of ITB-SP with GA the result values are used to make a graph which display better performance of ITB-SP with GA than other algorithms.

In next graph shows the results of ITB-SS with GA with better results as compare to PSO and ACO based upon parameters as Precision, Recall and F-measure and Time.



Graph 3 - ITB-SP with GA and compared with PSO and ACO

Conclusion

Outlier detection is the data mining task whose goal is to isolate the observations which are considerably dissimilar from the remaining data. This task has practical applications in several domains such as fraud detection, intrusion detection, data cleaning, medical diagnosis, and many others. In this paper a hybrid algorithm is used to detect outliers in better way than previous algorithm. The algorithm based on information theoretic with genetic algorithm is better for detecting outliers and obtained optimized data. After implementation of this algorithm compare the result with other optimization techniques as ACO and PSO.

References

D Hawkins (1980),Identification of outliers,Chapman and Hall, London.

Amandeep Kaur, Ms. Kamaljit Kaur(2014),Different Outlier Detection Algorithms in Data Mining: A Review, IJSRD - International Journal for Scientific Research & Development, Vol. 2, Issue 04, 2014 , ISSN (online): 2321-0613.

Hodge, V.J. and Austin, J. (2004), A survey of outlier detection methodologies. Artificial Intelligence Review,22 (2). pp. 85-126.

Jiawai Han,Micheline Kamber,Jian Pei (2012) ,Data Mining-concepts and techniques ,morgan kaufmaan publishers,third edition.

Barnett, V. & Lewis, T. (1994),Outliers in Statistical Data, 3rd edition, (John Wiley & Sons,Chichester),pp-584,ISBN 0-471-93094-6.

Rousseeuw, P. & Leroy A. (1996),Robust Regression and Outlier Detection, 3rd edition , Copyright@1996 John Wiley & Sons,Inc.

S. Ramaswamy,R. Rastogi,K.Shim(2000),Efficient algorithms for mining outliers from large data sets, Proceedings of the International Conference on Management of Data, Dallas, Texas.

Z.Chen,a.Fu,J.Tang(2003), On Complementarity of Cluster and Outlier Detection Schemes,Springer Verlag,LNCS 2737, pp.234-243 .

M. O. Mansur , Mohd. Noor Md. Sap(2005),Outlier Detection Technique in Data Mining: A Research Perspective, Proceedings of the Postgraduate Annual Research Seminar 2005,23-31.

Papadimitriou, S., Kitawaga, H., Gibbons, P., Faloutsos,C.(2003), LOCI: Fast outlier detection using the local correlation integral, Proc. of the International Conference on Data Engineering.

W. Lee and D. Xiang(2001), Information-Theoretic Measures for Anomaly Detection, Proc. IEEE Symposium Security and Privacy,pp.130-143.

Z. He, X. Xu, and S. Deng (2005), An Optimization Model for Outlier Detection in Categorical Data,Proc. Int'l Conf. Advances in Intelligent Computing (ICIC '05).

K. Das, J. Schneider, and D.B. Neill(2008), Anomaly Pattern Detection in Categorical Data Sets, Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08).

K. F. Man, K. S. Tang, and S. Kwong (1996), Genetic Algorithms: Concepts and Applications, IEEE Transactions On Industrial , Vol.43,Issue:5 , pp-519-534,ISSN-0278-0046.

Marco Dorigoa, Christian Blumb (2005), Ant colony optimization theory: A survey, Theoretical Computer Science 344, pp.243-278.

Bing Xue, Member, IEEE (2012), Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach, 2168-2267/© 2012 IEEE

Fabrizio Angiulli, Stefano Basta, Stefano Lodi, and Claudio Sartori (2013),Distributed Strategies for Mining Outliers in Large Data Sets, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 7, July 2013.

Stephen D. Bay and Mark Schwabacher (2003), Near Linear Time Detection of Distance-Based Outliers and Applications to Security, epic.org/ privacy/airtravel/ nasa/study.pdf.