Global High Dimension Outlier Algorithm for Efficient Clustering and Outlier Detection

Nidhi Nigam
M.Tech (CSE)
Laxmi Narain College of technology
Bhopal, India

Tripti Saxena
Assistant Professor (CSE)
Laxmi Narain College of technology
Bhopal, India

ABSTRACT
In this digital era most of the knowledge kinded on the market in digital form. For several years, individuals have command the hypothesis that exploitation phrases for square measure presentation of document and topic ought to perform higher than terms. During this paper we have a tendency to square measure examine and investigate this reality with considering many states of art data processing strategies that offers satisfactory results to boost the effectiveness of the pattern. Here we have a tendency to implementing pattern detection methodology to resolve downside of term-based strategies and improved result that useful in info retrieval systems. Our proposal is additionally evaluated for many well distinguish domain, providing all told cases, reliable taxonomies considering preciseness and recall in conjunction with F-measure. For the experiment, we'll use massive dataset and therefore the results ought to show that we have a tendency to improve the discovering pattern as compared to previous text mining strategies. The results of the experiment setup ought to show that the keyword-based strategies not offer higher performance than pattern-based methodology. The results additionally indicate that removal of vacuous patterns not solely reduces the price of computation however additionally improves the effectiveness of the system

Keywords
KDD, DBSCAN, Noisy data, Distributed solving set, Lazy distributed solving set.

1. INTRODUCTION
Larger and bigger amounts of knowledge area unit collected and hold on in databases, increasing the requirement for economical and effective analysis ways to create use of the {data} contained implicitly within the data. Information discovery in databases (KDD) has been outlined because the non-trivial method of characteristic valid, probably helpful, and ultimately intelligible information from the information. Most studies in KDD concentrate on finding patterns applicable to a substantial portion of Objects during a dataset. However, for applications like police work criminal activities of varied sorts (e.g.

Electronic commerce), rare events, deviations from the bulk, or exceptional cases are also a lot of attention-grabbing and helpful than the common cases. Finding such exceptions and outliers, however, has not nevertheless received the maximum amount attention within the KDD community as another topics have, e.g. association rules.

Sample Application and scope of outlier detection scheme:

A. *Fraud detection*

Purchasing behaviour of a MasterCard owner sometimes changes once the buying behaviour of a MasterCard owner sometimes changes once the cardboard is taken abnormal shopping for patterns will characterize MasterCard abuse.

B. *Medicine*

Unusual symptoms or take a look at results could indicate potential health issues of a patient whether or not a specific take a look at result's abnormal could depend upon different whether or not a specific take a look at result's abnormal could depend upon different characteristics of the patients (e.g. gender, age …)

C. *Public health*

The prevalence of a specific unwellness, e.g. tetanus, scattered across The prevalence of a specific unwellness, e.g. tetanus, scattered across varied hospitals of a town indicate issues with the corresponding vaccination program therein town whether or not an occasion is abnormal depends on completely different aspects like vi frequency, abstraction correlation, etc.

There are unit varied applications and approaches that area unit raised and employed in the sector of outlier detection

D. *Sports statistics*

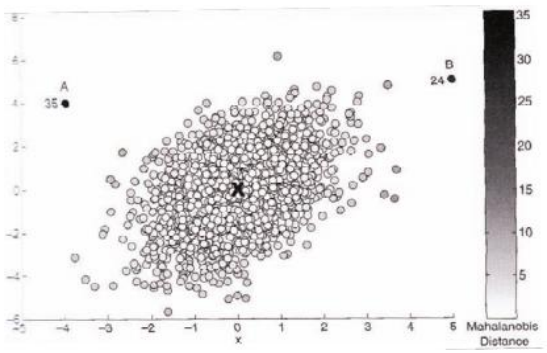
In several sports varied parameters are recorded for players so as to In several sports, varied parameters are recorded for players so as to gauge the players' performances may be known as having p abnormal parameter values typically, players show abnormal values solely on a set or a special combination of the recorded parameters. Outstanding (in a positive likewise as a negative sense)

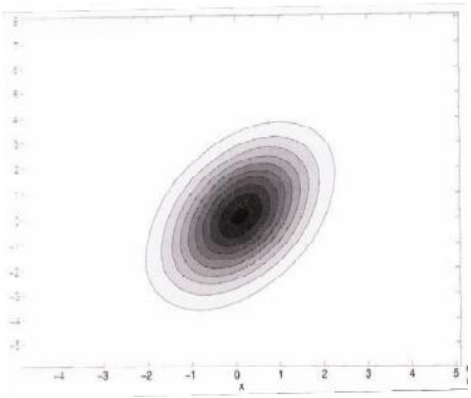
Detecting measurement errors

Information derived from sensors (e.g. during a given scientific experiment) could contain menstruation errors Abnormal values may give a sign of a menstruation error Removing such errors may be necessary in different data processing and information analysis tasks "One person's noise may be another person's signal." so as to figure with the outlier approach we are able to work with 3 totally different approaches for the outlier detection scheme:

- International versus Native outlier detection Considers the set of reference objects relative to that every purpose Considers the set of reference objects relative to that every purpose s' "outlines" is judged
- Labelling versus scoring outliers Labelling versus scoring outliers considers the output of an algorithm
- Modelling properties modelling properties Considers the concepts based on which "outlines" is modelled.

There can be following visualization if we discuss and talk about the outlier approach:





Motivation And Scope Of Work

Outlier detection task can be very time consuming and recently there has been an increasing interest in parallel/ distributed methods for outlier detection The work defined by us is having the various motivation obtain in order to progress further to the research topic chosen by us, on working with the current given title and scenario we can work further with the various security concern with the help of outlier detection technique in the large dimensional dataset.

The research is motivated to find:

- Outlier of user required data from large dataset.
- Working with various pattern mining approaches based on the outlier detection technique in the dataset.
- Intrusion detection with the help of treating them as outlier in the system intrusion and prevention system.
- We will solely target "unsupervised learning" techniques. Reduces to finding thin regions in massive three-D information sets. We are going to 1st summary applied mathematics ways and so give a summary of ways that have emerged from at intervals the DM community.

2. LITERATURE REVIEW

Rakesh Agrawal Johannes Gehrke Dimitrios Gunopulos Prabhakar Raghavan In this paper Author proposed a technique which is "CLIQUE technique" and the technique specifies a process of outlier detection in this technique named CLIQUE identify dense clusters in subspaces of maximum dimensionality, CLIQUE finds accurate clusters in large high dimensional datasets according to specification given in this clique technique and the technique outperform for determining the outlier from the various dataset provided for the training and testing purpose.

The proposed technique in this paper notify following disadvantage:

The problem is to automatically identify projections of the input data into a subset of the attributes with the property that these projections include regions of High density.

They address the problem of evaluating the quality of clustering's in different subspaces.

There can be following Application of the Clique technique:

- Data mining pattern recognition.
- Computational geometry.

Charu C. Aggarwal Philip S. Yu

The author proposed a paper in which in order to find a better outlier and to utilize them in the field of research the author have proposed Naïve Brute force Approach which is using in the paper , according to them the technique followed by them is to find out the pattern from the multiple dataset, the algorithm say that a naive brute force algorithm which is very slow at Finding the best patterns because of its exhaustive search of the entire space, and a much faster evolutionary algorithm which is able to quickly and hidden combinations of dimensions in

Which the data is sparse.

The algorithm having the following drawback:

- Entire space searching in high dimensional data.
- Slowness of algorithm.

  And the system can have following application in various areas:

- fraud detection,
- network robustness analysis,
- Intrusion detection thus the algorithm having a good approach and application it is proven in various secured aspects and data mining areas.

Charu C. Aggarwal The author of this paper have again proposed another paper in which they have stated the approach names the effects of generalizing low dimensional techniques to high dimensional application and the have work on generalization and their effect in high dimensional application and its data, the technique was showing the effect of low dimensional technique in high application.

In the scheme the merits and demerits are depending on the situations and the required scheme which is being on the parameters being given by the user.

The paper having the following application: Distance based applications for high dimensional problems are a promising line of future research.

Dr. Mohammed Ali Hussain, 2Dr. R. Satya Rajesh, 3Md. Abdul Ahad

The author proposed an innovative approach in the respected area of clustering and outlier detection approach where they have proposed technique "DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is based on

centre-based approach". They have follow a novel approach for clustering with noise and large amount of dataset and they have proposed following steps in order to find outlier and to make new cluster:

The DBSCAN algorithm is given as follows:

- Step 1: Label all points as core, border, or noise points.
- Step 2: Eliminate noise points.
- Step 3: Put an edge between all core points that are within Eps of each other.
- Step 4: Make each group of connected core points into a separate cluster.
- Step 5: Assign each border point to one of the clusters with

As the algorithm having a good scope while dealing with the noisy data but at the same time DBSCAN scans many objects repeatedly with many times. Costing is much high. Which is a big drawback in the proposed algorithm proposed by the authors of this paper.

On working with the proposed algorithm in this paper as there are limitations but at the same time there are some advantages and application in the respected area which is by exploring the benefits of Gaussian-Means. DBSCAN can discover all clusters with arbitrary shape and separate noises.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander The paper and author proposed a technique which is efficient algorithm in order to produce the outliers the technique proposed by the is LOF Technique which is local outlier factor technique for the boundary or outlier detection.

The technique is progressive and its native in this the degree depends on however isolated the thing is with relation to the encircling neighbourhood. We have a tendency to provide an elaborate formal analysis showing that LOF enjoys several fascinating properties and also the properties area unit usage so as to search out out the simplest outlier in needed approach.

The algorithm having following limitation that it is limited to local outliers and depends on mts Pts. parameter.

Application also provided for the proposed algorithm of outlier LOF technique which is Efficient with local outlier and can be used with in exclusive dataset and also to work with large number of dataset it can work efficiently.

3. EXISTING SYSTEM

The authors approach is predicated on the thought of outlier detection determination set that could be a little set of the information set which will be extensively utilized for predicting novel outliers. The strategy exploits parallel computation so as to get huge time savings. Indeed, on the far side conserving the correctness of the result, the planned schema exhibits wonderful performances. They need planned a variant of the fundamental strategy that reduces the quantity of information to be transferred so as to boost each the communication price and therefore the overall runtime. Significantly, the determination set computed by our approach in a distributed environment has the same quality as that produced by the corresponding centralized method.

They have proposed a method which is distributed method (called Distributed Solving Set) to detect distance-based outliers, which is suitable to be used both in parallel and distributed scenarios.

Other than solving the distance-based outlier detection task in the distributed scenario, the method computes an outlier detection solving set of the overall data set.

They have presented a variant of the basic method, called Lazy Distributed Solving Set, which reduces the amount of data to be exchanged from the nodes with respect to Distributed Solving Set by adopting a strategy that leads to the transmission of a reduced number of distances while slightly increasing the number of communications.

The core computation executed at each node consists in the following steps:

1. Receiving this resolution set objects at the side of this edge for the burden of the highest ordinal outlier. Comparing them with the native objects.
2. extracting a replacement set of native candidate objects (the objects with the highest weights, in line with this estimate) alongside the list of native nearest neighbors with regard to the finding set and, finally.
3. Determinative the quantity of native active objects, that's the objects having weight not smaller than this boundary.

4. PROPOSED METHODOLOGY & ALGORITHM

The requirement is to conclude with the algorithm which provides the outlier which is efficient, without any

Repetition in values or outlier output and at the same time it give us less duration while executing and global search.

Algorithm Name: Global High Dimension outlier Approach (GHOA) algorithm for outlier detection.

STEPS:

Terms

All available dataset D (1-n).

Available density or hierarchy of dataset or relationship definition- RI

Relational Dataset – RD (1-n)

Threshold or peak for outlier - T

Outlier Detection on threshold-

Evaluation of data - Summation (RD) 1-n =SRD (i-n)

BEGIN-

Step 1-

For D (1-n)

Step 2-

Calculation for RD (1-n) based on RI.

Finding union from relational dataset based on threshold value and storing into outlier.

Step 3 -

Outlier = SRD (i-n) U T

Step 4-

Reduplication on Outlier = redundancy removal =Unique (Outlier)

Final Outlier = Unique (outlier) END

5. ADVANTAGE OF PROPOSED SCHEME

1. On Providing Global search it will remove the disadvantage of one of our previous work which limit to local search.
2. DE duplication is applied with unique function which only considers unique values and remove redundancy.
3. Data redundancy is the efficient outcome from the defined approach by the proposed scheme.
4. Our research can be work for the global search and hence it can work with various depth search and external linked resources.

6. CONCLUSION

We presented the Global High Dimension Outlier Approach algorithm, a distributed method for computing an outlier detection solving sets. The work which is done by our survey paper is to elaborate and to find research on the various technique which are already been used in order to work with the outlier and also we have checked stated the various techniques, their advantages and limitation and we have kept our observation in different aspect and the application may possible with the help of stated approach.

In this work, we discussed a new technique for outlier detection which is especially suited to very high dimensional data sets. The method works by finding lower dimensional projections which are locally sparse, and cannot be discovered easily by brute force techniques because of the number of combinations of possibilities. This technique for outlier detection has advantages over simple distance based outliers which cannot overcome the effects of the dimensionality curse. A new defined approach is also mentioned in order to detect outlier which is enhancement of the outlier detection technique which is solving set given in the base paper and we have studied and proposed the algorithm to prove best in order to provide a best result in the field of data redundancy.

7. REFERENCES

[1] Fabrizio Angiulli, Senior Member, IEEE, Stefano Basta, Stefano Lodi, and Claudio Sartori "Distributed Strategies for Mining Outliers in Large Data Sets" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013.

[2] F. Angiulli, S. Basta, S. Lodi, and C. Sartori, "A Distributed Approach to Detect Outliers in very Large Data Sets," Proc. 16th Int'l Euro-Par Conf. Parallel Processing (Euro-Par), pp. 329-340, 2010.

[3] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, Feb. 2006.

[4] Rakesh Agrawal Johannes Gehrke_ Dimitrios Gunopulos Prabhakar Raghavan," Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications"

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast Discovery of Association Rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthu rusamy, editors, Advances in Knowledge Discovery and Data Mining, chapter 12, pages 307{328. AAAI/MIT Press, 1996.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Survey, vol. 41, no. 3, pp. 15:1-15:58,2009.

[7] H. Dutta, C. Giannella, K.D. Borne, and H. Kargupta, "Distributed Top-K Outlier Detection from Astronomy Catalogs Using the DEMAC System," Proc. SIAM Int'l Conf. Data Mining (SDM), 2007.

[8] A. Ghosting, S. Parthasarathy, and M.E. Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets," DataMining Knowledge Discovery, vol. 16, no. 3, pp. 349-364, 2008.

[9] S.E. Guttormsson, R.J. Marks, M.A. El-Sharkawi, and I. Kerszenbaum, "Elliptical Novelty Grouping for on-line Short- Turn Detection of Excited Running Rotors," Trans. Energy Conversion, vol. 14, no. 1, pp. 16-22, 1999.

[10] J. Han and M. Kamber, Data Mining, Concepts and Technique. Morgan Kaufmann, 2001.

[11] E. Hung and D.W. Cheung, "Parallel Mining of Outliers in Large Database," Distributed and Parallel Databases, vol. 12, no. 1, pp. 5-26,2002.

[12] S. Jakubek and T. Strasser, "Fault-Diagnosis Using Neural Networks with Ellipsoidal Basis Functions," Proc. Am. ControlConf., vol. 5, pp. 3846-3851, 2002.

[13] Advances in Distributed and Parallel Knowledge Discovery, H. Kargupta and P. Chan, eds. AAAI/MIT Press, 2000.

[14] E. Knorr and R. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," Proc. 24rd Int'l Conf. Very Large DataBases (VLDB), pp. 392-403, 1998.