## A Report on

**Greedy Algorithm for**

**Outlier Mining**

BACHELOR OF TECHNOLOGY in

COMPUTER SCIENCE AND ENGINEERING

By

SUDHANSHU SHEKHAR

Admission Number: 16JE001971

SESSION: 2017-2018

UNDER THE GUIDANCE OF

**Dr. Rajendra Pamula**

**Assistant Professor**



DEPARTMENT OF COMPUTER SCIENCE

INDIAN INSTITUTE OF TECHNOLOGY

(INDIAN SCHOOL OF MINES)

DHANBAD-826004

**Acknowledgements**

It gives me immense pleasure to extend my thanks to those who provided inestimable support for the help in my work.

In the first place I would like to record my gratitude to Dr. Rajendra Pamula for his supervision, advice, guidance from the very early stage of this project and for patiently guiding me. I gratefully acknowledge him for his crucial contributions, which made him a backbone of this dissertation.

I am also thankful to Mr. Manmohan for his invaluable support and encouragement as well as all faculty member of the department for providing time to time guidance during this period.

I would like to thank my parents, for their faith in me and allowing me to be as ambitious as I wanted.

Further, I would like to thank all the people who have directly or indirectly helped me.

**SUDHANSHU SHEKHAR**

**16JE001971**

**Computer Science & Engg.**

**IIT(ISM) Dhanbad**

**Table of Contents**

1. **Abstract………………...(pg-4)**
2. **Introduction………………..(pg-5)**
3. **Background…..(pg 6-7)**
4. **Related work…..(pg 8-10)**
5. **Experimental Results……………………….(pg 11-12)**
6. **Conclusion…………………..(pg 13)**
7. **References…………………...(pg 14)**

## 1. Abstract:

**An Outlier is an observation which is different from the others in a sample.**

**Usually an Anomaly occurs in every data due to measurement error.**

**Anomaly detection is identifying anomalous data for given dataset that does**

**Not show normal behaviour. Anomaly Detection can be classified into three**

**Categories: Unsupervised, Supervised and Semi-Supervised anomaly**

**Detection. Anomaly detection is used variety of domains like fault Detection,**

**Fraud detection, health monitoring system, intrusion detection.**

**In this report we implement a very fast greedy algorithm for mining outliers**

**(For single attribute).**

**Two major observations of the presented algorithm are :**

**(1).Our Algorithm has comparable performance with state of art outlier**

**Detection algorithms.**

**(2)Our Algorithm can be of order of magnitude faster than LSA algorithm.**

**Keywords:** Outlier, Greedy algorithm, Entropy.

# 2. Introduction:

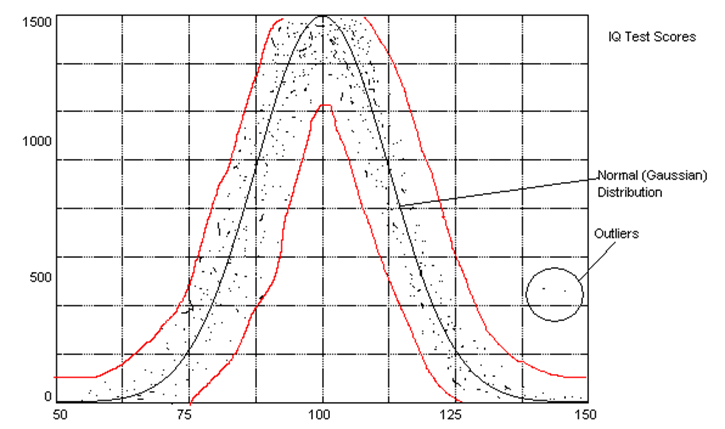
An Outlier is an observation which is different from the others in sample. Outliers arise due to many reasons like malicious activity. Example credit card fraud etc. Outlier detection is different from noise removal though it is very much similar to it. A key difference between noise removal and outlier detection is data in noise removal that is not of interest to analyst, but acts as a hindrance to data analysis whereas in outlier detection the data are interesting to the data analyst. A well quoted definition of outliers is firstly given by Hawkins , “An outlier is an observation that deviates so much from the different observations as to arouse suspicions that it was generated from a different mechanism.” However conventional approaches do not handle categorical data in a satisfactory manner and most existing techniques lack for a theoretical foundation or assume underlying distributions that are not well suited for exploratory data mining applications. To fulfil this void the problem of outlier detection in categorical data is defined as an optimisation problem as follows: finding k objects such that the expected entropy of the resultant dataset after removal of this subset is minimised.

In the above optimisation an exhaustive search through all possible solutions with outliers for the one with minimum value is costly as there are (n,k) possible values and thus the above algorithm is still very time consuming , so we present a very fast greedy algorithm for mining outliers under the same optimisation model .

# 3. Background:

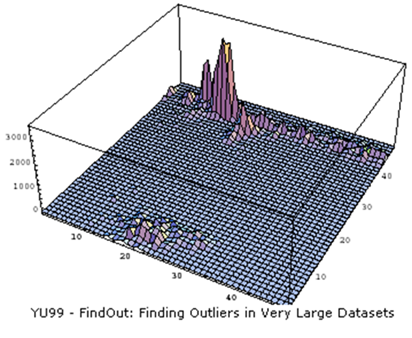
**Other methods for Outlier detection broadly fall into the following categories:**

**(1) Distribution Based : Data is assumed to be part of a working hypothesis (working hypothesis). Each data object in the dataset is compared to the working hypothesis and is either accepted in the working hypothesis or rejected as discordant into an alternative hypothesis (outliers).**

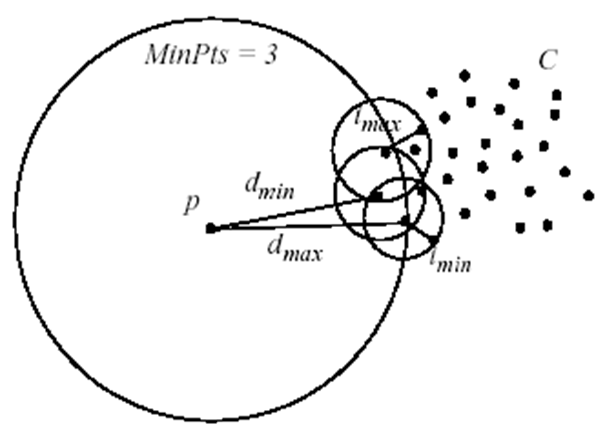
****

**(2)Depth based : Data is organized into layers according to some definition of depth . Shallow layers are more likely to contain outliers than deep layers.**

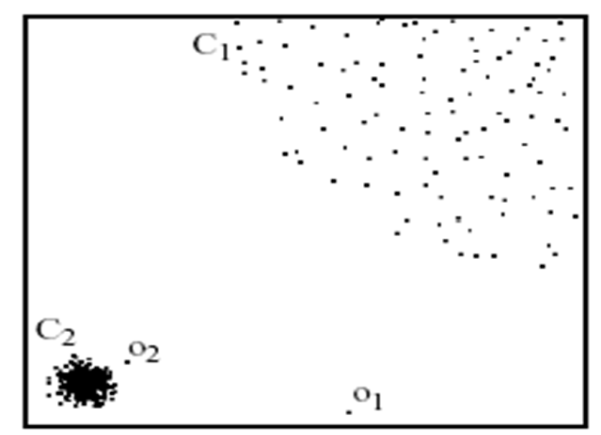
**This method can efficiently handle computation for k < 4.**

****

**(3)Distance based Outlier detection : Distance-based: An object O in a dataset T is a DB(p,D) outlier if at least fraction p of the objects in T are >= distance D from O. A point O in a dataset is an outlier with respect to parameters k and d if no more than k points in the dataset are at a distance of d or less from O.Relative measurement: Let Dk(O) denote the distance of the kth nearest neighbour of O. It is a measure of how much of an outlier point O is.**

****

**(4) Density based outlier detection :** **A density-based outlier detection (OD) method is presented by measuring the local outlier factor (LOF) on a projected principal component analysis (PCA) domain from real world spatial-temporal (ST) traffic signals. Its aim is to detect traffic data outliers which are errors in data and traffic anomalies in real situations such as accidents, congestions and low volume.**

****

# 4. Related Work:

# **Entropy Calculation:**

## 

# Approach Used:

**The problem we are trying to solve can be formulated as follows. Given a dataset D of n points p1,….pn, where each point is a multidimensional vector of a single categorical attribute . We would like to find a subset O of D with size k such that E(O-D) is minimum. To make the computations more simple we assume independence of the record such that the joint probability of combined attribute values becomes the product of the probabilities of each attribute and hence the entropy can be computed as the sum of entropies of the attributes.**

**E(X) = E (X1) + E (X2) +…..+ E (Xn)**

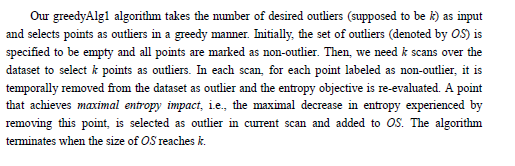
**In Our Source code written in C++ and compiled using GCC 4.9.2 we have taken the input of the individual values of the first attribute and used a hash map to store the frequency of the occurrence of that attribute for each vector.**

**This is followed by calculation of entropy of each vector based upon their frequency.**

**Finally we have calculated out total entropy and used greedy algorithm to determine for which vectors the change in entropy is maximum.**

**These Vectors are then flagged as outliers.**

# Use of greedy Algorithm:



# Data structure used:

**Given a dataset D of n points p1….pn where each point is a multidimensional vector of a single categorical attribute, we need 1 corresponding hash table as our basic data structure. The hash table has values as keys and the frequencies of the attribute values as referred values. Thus in O(1) expected time we can determine the frequency of an attribute value in the hash table.**

# Time and space complexity:

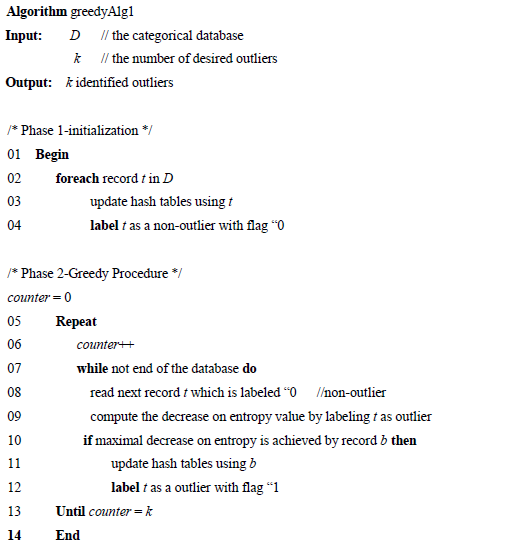
**Worst case: The time and space complexity depend on the size of data set(n),the number of attributes(m , in this project we have taken it a 1 for simplicity of code),the size of every hash table , the number of outliers(k).**

**We will assume that every attribute has the same number of distinct attribute values,p.Then in the worst case in the initialisation phase the time complexity is O(n\*m\*p). In greedy procedure since the computation of value change on entropy require at most O(m\*p) and hence this phase has time complexity O(n\*k\*m\*p).**

**The algorithm only needs to store m hash tables and the dataset in main memory so the space complexity of our algorithm is O((p+n)\*m).**

**(In this project we have taken m as 1 for simplicity of code.)**

# Algorithm Used:



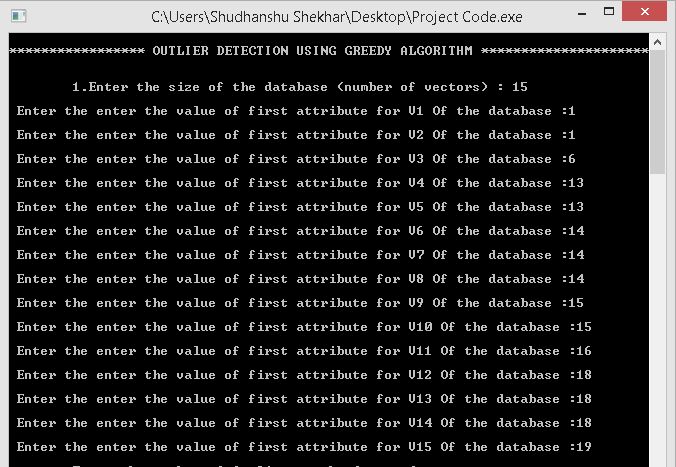
# 5. Experimental Results:

**Our code for the implementation of the above algorithm is written in C++ language. Sample test case depicted below is using only a single attribute for simplicity of explanations.**

# Test case for input:

**Data set of size 15 was taken i.e. N=15.**

**Values of individual attributes for each of the 10 vectors are given in the picture below:**



**Value of K (number of outliers) was taken as : 3.**

**Output detected was:**



# Explanation of the obtained result :

**The input parameters can be distributed on a line segment as:**

**1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19**

**Detected Anomalies**

# 6. Conclusion:

**Convectional outlier mining algorithms do not handle categorical data in a satisfactory manner. To fulfil this void, the very fast greedy algorithm for mining is provided. Experimental results show the high ability of this algorithm.**

**Furthermore our Algorithm solves the given problem in the time of O(n\*k\*m) , where n is the database size , k is the number of outliers to be detected and m is the number of attributes (we have taken this 1 in our code for simplicity ). This time complexity is far better than that of LSA ie O(n\*k\*l\*m) where l is the number of iterations in the LSA and is usually greater than 10.**

# 7. References:

**[1] . Zengyou He, Xiaofei Xu , Shegchung Deng . Paper on “A fast greedy algorithm on Outlier mining “.**

**[2]. .wikipedia.org/wiki/Anomaly\_detection**

**[3]. "Introduction to Data Mining" by P. Tan, M. Steinbach, and V. Kumar**