iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8219

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

Outlier Detection in High Dimensional Data Based on the Anti-Hub and Regression Technique

Golla. Hemalatha[1], T. Suresh[2]
[1]*M.Tech Student,* [2]*Assistant Professor, Department of Software Engineering*
LakiReddy BaliReddy college of Engineering (Autonomous), Mylavaram, Andhra Pradesh, India.

Abstract: Outlier detection refers to find patterns that do not fit in to normal behaviour. Outlier detection plays an important role in data mining. Most of real world datasets are using outlier detection. Outlier detection is useful in many fields like Network intrusion, Credit card fraud detection, stock market, and wireless sensor network data. The distance based outlier detection techniques be unsuccessful to will increases the dimensionality of data because in high dimensionality distance between the two points is less. In existing system using Anti-hub method in reverse nearest neighbors. Anti-hubs are few points are regularly comes in K-nearest neighbors list of another points and few points are an irregularly comes in k-nearest neighbors list of different points. In Anti-hub method propose high computational cost and time requirements for finding outliers. To overcome this problem we can use new method in this paper that is the advanced variety of Anti-hub is Anti-hub2.which is for reconsider the outlier score of a data point obtained by the Anti-hub method. The goal of this paper is locate the inconsistent objects in data which has high dimension through reduced computation time, cost and increase the accuracy. We apply logistic regression rule on the results of Anti-hub dataset then obtained combination of data, prevention measures and Anti-hub calculation. It increase the efficiency of remove out irrelevant, redundant feature.

Keywords: outlier detection, High dimensional data, Anti-hub, Anti-hub2, Logistic regression

I. INTRODUCTION

Data Mining means the process of extracting the knowledge for huge data sources. The general objective of the data mining process is to extract information from a data set and transform it into a reasonable structure for further utilize. In the Data Mining have four techniques that are Clustering, Classification, Association, and Outliers. In this paper we discuss the Outlier detection techniques definition of Outlier location (otherwise called inconsistency discovery) is the way toward discovering data objects with practices that are altogether different from desire. Such objects are called **outliers** or anomalies. Outlier detection is important in many applications such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance, and intrusion detection.

In general, outliers is classified into 3 varieties, those are global outliers, conditional outliers, and collective outliers. To discover global outliers, an essential issue is to search for out an acceptable mensuration of deviation with relation to the appliance in question. Global outlier detection is very significant in several applications. Take into account intrusion detection in laptop networks, as an example. If the statement behaviour of a laptop is extremely completely changed from the conventional designs (e.g., an oversized range of packages is broadcast in a very short time), this behaviour is also thought-about as a worldwide outlier and also the consistent laptop could be a suspected target of hacking. Next one is conditional outlier the item if it deviate considerably supported selected context. Attributes of information objects must to be 2 teams i.e.; discourse attributes: defines the context. E.g. Time and site. And second is activity attributes: features of object, employed in outlier analysis e.g. temperature. Third kind of outliers is Collective outliers a set information of knowledge of information object jointly deviate considerably from the complete data set, even individual knowledge object might not be outliers. Detection of collective outliers might take into account not solely individual knowledge object however additionally take into account cluster knowledge objects.

There are several outlier's recognition strategies here, we show two approaches to sort exception location techniques. Initially, we sort outlier discovery techniques as indicated by whether the example is information for investigation is given with area expert–provided marks that can be utilized to assemble an exception identification demonstrate. Outlier can got: Supervised, semi-supervised, unsupervised strategy.

A. Supervised Outlier location would then be able to be displayed as a grouping issue. The errand is to take in a classifier that can perceive exceptions. The example is utilized for preparing and testing. Difficulties to directed exception location incorporate the associated: imbalanced class i.e. exceptions uncommon Boost the outlier class and make up some fake

©IJRASET (UGC Approved Journal): All Rights are Reserved

outliers, and catch however many outliers as could be expected under the conditions i.e. review is more imperative than precision.

B. Unsupervised strategies are all the more broadly connected, in light of the fact that alternate classifications require exact and agent marks that are regularly restrictively costly to get. Unsupervised strategies incorporate separation construct techniques that essentially depend in light of a measure of separation or closeness keeping in mind the end goal to recognize outliers.

C. Semi supervised outlier recognition watched uses of semi regulated learning. In numerous applications, in spite of the fact that getting some marked cases is possible, the quantity of such named illustrations is frequently little. On the off chance that lone some named outliers are accessible, semi-managed exception identification is delicate. Few named outliers are probably not going to speak to all the possible exceptions. To enhance the nature of exception location, we can get assistance from models for typical items gained from unsupervised strategies.

We can classify outlier detection methods into three types: statistical methods, proximity-based methods, and clustering-based methods. Statistical approaches (also known as model-based methods) type suspicions of data regularity. They accept that typical information objects are produced by a measurable (stochastic) display, and that information not following the model are outliers. Proximity-based techniques accept that a objection is an outlier if the nearest neighbors of the question are far away in highlight space, that is, the closeness of the question its neighbors essentially deviates from the nearness of the vast majority of alternate items to their neighbors in similar informational collection**.** Clustering-based techniques accept that the ordinary information objects have a place with vast and thick groups, though exceptions have a place with little or scanty groups, or don't have a place with any groups.

Previously so many methods are used for finding outliers.in this we can discuss some methods those are hub and Anti-hub both methods are using the nearest neighbour. K-nearest neighbor of purpose N is K points whose distance to point N is a smaller amount than all different points. Reverse nearest neighbors (RNN) of purpose N is that the points that N is in their k nearest neighbor list. Some points are oft comes in k-nearest neighbor list of different points referred as hubs and a few points square measure sometimes comes in k nearest neighbor list of another points are referred as Anti-hubs. Recently, the development of hubness was determined that affects reverse nearest-neighbor counts, i.e. k-occurrences (the range of times purpose x seems among the k nearest neighbors of all different points within the knowledge set). Hubness is manifested with the rise of the (intrinsic) spatial property of information, inflicting the distribution of k-occurrences to become skew, additionally having exaggerated variance. As a consequence, some points (hubs) terribly oft become members of k-NN lists and, at constant time, another points (antihubs) become sporadic neighbors. Use of antihubs for outlier detection is of high procedure task. Computational complexity depends on data dimensionality as dimensionality of data increases the complexity of computation increases. Because of this nearest neighbor introduced to remove irrelevant and redundant data.

We are examine above just presentation of exceptions yet in this section we talk about relapse rules due to in this paper can confirm the identifying outlier by utilizing regression rules. Regression can be data mining system that predicts determination. Benefit, deals, contract rates, house estimations, e.g. film, temperature, or separation may all be normal misuse relapse methods. An illustration, a regression model may be wont to foresee the value of a house upheld area, scope of rooms, parcel estimate, and distinctive elements. In factual displaying, regression investigation could be a connected math strategy for assessing the connections among factors. It incorporates a few strategies for displaying and examining numerous factors, once the primary target is on the association between a variable and one or a lot of autonomous factors. The only and oldest kind of regression is linear regression wont to estimate a relationship between 2 variables. This system uses the mathematical formula of a line (y = Mx + b) this merely means, given a graph with a Y associated an X axis, the connection between X and Y could be a line with few outliers. Advanced techniques, like multiple regression, calculate a relationship between multiple variables a lot of variables significantly will increase the complexes of the calculation. There are many sorts of multiple regression techniques together with standard, hierarchical, setwise and stepwise, every with its own application.

Nonlinear regression may be a style of multivariate analysis in that during which within which data-based knowledge area unit modeled by perform which may be a nonlinear combination of the model parameters and depends on one or additional independent variables. The graph of a nonlinear perform isn't a line. Linear functions have a constant slope, thus nonlinear functions have a slope that varies between points. Algebraically, linear functions area unit polynomials with the highest exponent adequate one or of the shape y = c wherever c is constant.

©IJRASET (UGC Approved Journal): All Rights are Reserved

II. RELATED WORK

As per the arrangement in [1], the extent of our hunt is to study: (1) point differences, i.e., singular focuses that can be considered as exceptions without considering logical or aggregate data, (2) unsupervised techniques, and (3) strategies that appoint an "anomaly score" to each point, creating as yield a rundown of anomalies positioned by their scores. The portrayed extent of our examination is the concentration of most exception identification inquire about [1].Distance construct anomaly discovery is based with respect to the distance of a point from closest neighbor. We rank each point on the premise of its separation to its closest neighbor and pronounce the best indicates in this positioning be exceptions. Cell based calculation whose multifaceted nature is direct in the measure of the database for high measurements [2]. The Angle based outlier detection (ABOD) utilizes the properties of the fluctuations to really exploit high dimensionality and has all the earmarks of being less touchy to the expanding dimensionality of an informational index than exemplary separation based strategies. Separation based anomalies which can't conquer the impacts of the dimensionality curse. Pairwise remove ends up noticeably undetectable as dimensionality builds at that point separate esteems are distinctive then it nearness superfluous properties.

A broadly utilized density based technique is the local outlier factor (LOF), which impacted numerous varieties, e.g., local correlation integral (LOCI), local distance-based outlier factor (LDOF), and local outlier probabilities (LOP). Additionally, there exists the influenced outlierness measure (INFLO), in light of a symmetric relationship that considers the two neighbors and turn around neighbors of a moment that assessing its density distribution. The principle concentrate of was on the effectiveness of processing INFLO scores. Three issues brought by the "curse of dimensionality" inside the general setting of inquiry, ordering, and information handling applications: poor discrimination of distance caused by fixation, nearness of uncertain properties, and nearness of excess qualities, the greater part of that loss the ease of use of past distance and similarity measures. For outlier location RNN thought is utilized however there's no hypothetical verification that investigates the connection between the exception natures of the focuses and inverts closest neighbors. The turnaround closest check is get influenced on the grounds that the spatial property of the data will increment, along these lines there's should explore however exception discovery systems bases on RNN get low with the spatial property of the data. Considering arbitrary information with iid organizes and Euclidean distance, fixation is reflected in the way that, as dimensionality builds, the standard deviation of the appropriation of distance stays consistent, while the mean regard keeps on developing.

RNN thought is used yet there is no theoretical proof which considers the association between the exception natures of the focuses and invert closest neighbors. RN check is get affected the dimensionality of the information builds, so there is need to investigate how anomaly acknowledgment systems bases on RNN get impacted by the dimensionality of the information.

III. EXISTING SYSTEM

From set of cases existing framework contains the strategy for discovering irregular examples and it goes for make the use of exception identification discover interruption discovery and outlier location in a few applications and honest sources. Existing framework examined the issue in outlier identification in high dimensionality and demonstrates that however unsupervised strategies is utilized for exception recognition in high dimensional data. It likewise explores however argumentative to Anti-hubs are related with outlier nature of the reason and supported the connection against Anti-hubs points and outlier, there are 2 routes that of utilizing k-event data are predicted for outlier discovery for high and low dimensional learning for demonstrating the outlierness of focuses, beginning with the procedure ODIN (Outlier Detection using indegree Number).

Definition 1 (k-occurrences): Let $D \subset R^d$ be a finite set of n points. For point $\mathbf{x} \in D$ and a given distance or similarity measure, the number of k-occurrences, denoted $N_K(X)$, is the number of times $\mathbf{X}$ occurs among the k nearest neighbors[2] of all other points in D. Equivalently, $N_K(X)$ is the reverse k-nearest neighbor count of $\mathbf{x}$ within D.

Definition 2 (hubs and antihubs): For $q \in (0, 1)$, hubs are the $\lceil nq \rceil$ points $\mathbf{x} \in D$ with the highest values of $N_K(X)$. For $p \in (0, 1)$, $p < 1 - q$, antihubs are the $\lceil np \rceil$ points $\mathbf{x} \in D$ with the lowest values of $N_K(X)$

Algorithm 1: AntiHub*dist* oldest method (*D*, *k*) (based on ODIN)

Input:

Distance measure dist

Ordered data set $D = (X_1, X_2, ...X_M)$, where $\mathbf{x}_j \in R^d$, for $j \in \{1,2,...,m\}$

©IJRASET (UGC Approved Journal): All Rights are Reserved

No. of neighbors $k \in \{1, 2, ...$

Output:

Vector $\mathbf{s} = (s_1,s_2,...,s_m) \in R^n$, where s_i is the outlier score of $\mathbf{x}_j$ for $j \in \{1,2,...,m\}$

Temporary variables:

$t \in R$

Steps:

For each $j \in (1,2,...,m)$

$t := Nk(Xi)$ computed w.r.t. dist and data set $D\backslash\mathbf{x}_i$ 3) $s_j := f(t)$, where $f : R \rightarrow R$ is a monotone function

We compress the properties of the Anti-Hub (ODIN) strategy by considering diverse components: High dimensionality actuates great general connection between N_k scores and "ground truth" about outlierness, notwithstanding when k ≪ n. Notwithstanding, assuming one is keen on distinguishing exceptions that speak to just a little extent of the informational index (the typical situation), high dimensionality can cause issues in separating the scores, since the larger part of hopeful focuses will have comparable low N_k values. We will concentrate on the transaction between the quantity of information focuses (n) and dimensionality (d), and the topic of how data sparsity influences hubness. Let $X^{(i)}=(X_1^{(i)}……..X_d^{(i)}),i=0,…….n$ Give x a chance to be arbitrary vectors with a nonstop circulation F(X), $X =(X_1,...,X_d) \in Rd$ Let $N_1^{n,d}$ be the quantity of focuses from {x(1),...,x(n)} whose closest neighbor concerning Euclidean separation is x(0).
Although the limits the breaking points can't offer authoritative answers concerning what occurs in the limited case, they bring up fascinating issues of how data sparsity influences hubness: (1) Will dimensionality which rules the quantity of focuses actuate greatly solid hubness that will be difficult to oversee, and (2) Will expanding the quantity of focuses quick a decrease and removal of hubness from the data.

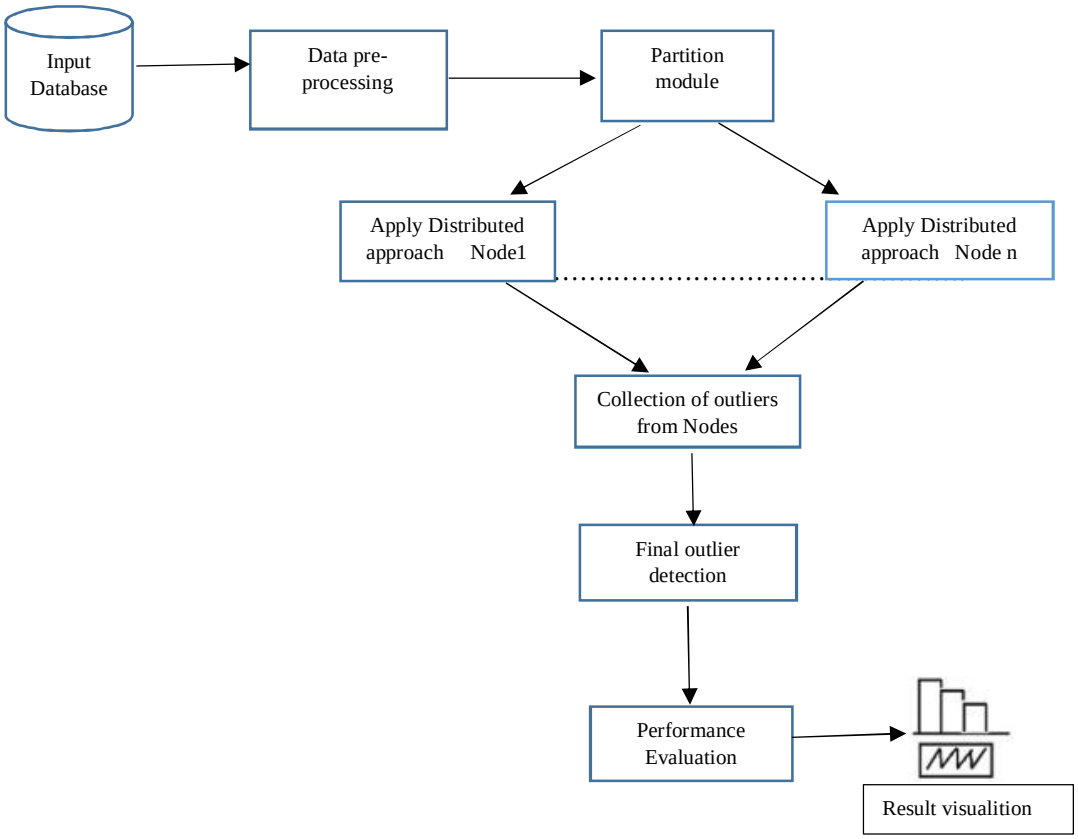
A. Disadvantages

1) Threshold values is utilized to separate outliers from ordinary object and lower outlierness threshold value will bring about high false negative rate for exception recognition.
2) Problem occurs when data instance is situated between two clusters, the inter distance between the object of k nearest neighbourhood increments when the denominator value builds reminders high false positive rate.
3) Necessities to enhance to compute outlier recognition speed.
4) Necessities to enhance the effectiveness of density based outlier recognition.

IV. PROPOSED SYSTEM

Description of the proposed system: An input of collection of large data set will be provided to the proposed system, as data is collected from standard data set repositories, data pre-processing will be applied before passing data to the next phase of the system. Further, this pre-processed input is being passed through to the partition module, where these datasets are been partitioned among many nodes from that one of the node is supervisor node and generate partition statistics and this statistical data is been visualized.

©IJRASET (UGC Approved Journal): All Rights are Reserved



A. Data Collection and Data Pre-Processing

In information gathering the underlying data sets for this framework will be gathered from standard dataset entrance i.e. UCI informational index storehouse. As proposed in framework, the standard dataset will be utilized for this framework incorporates Cover sort, IPS datasets. Gathered datasets might be accessible in their unique, uncompressed shape along these lines; it is required to pre-process such information before sending for future strides. To pre-process vast dataset substance, strategies accessible is information mining, for example, information joining, information change, information cleaning, and so forth will be utilized and cleaned, required information will be produced.

B. Data Partitioning

In this module, as expressed prior in framework execution design, the pre-processed information is partitioned into number of customers from focal director hub i.e. server according to the information ask for made by wanted number of customers. This divided information will be then prepared by singular customers to distinguish exceptions in view of connected calculation procedure.

C. Outlier Recognition

The method proposed for distinguishing outliers will be connected at first at circulated customers and their consequences of recognized exceptions would be coordinated on server machine at definite stage calculation of exceptions. To do this, the exception discovery procedures proposed are KNN Algorithm with ABOD and INFLO Method. The Distributed approach proposed with above Method in view of irregularity discovery procedures in view of closest neighbor .In this system supposition is that ordinary information cases happen in thick neighborhoods, while exceptions happen a long way from their closest neighbors. In this proposed work utilizing ideas of closest neighbor based inconsistency discovery techniques:(1) utilize the separation of an information example to its kth closest neighbors to register the exception score.(2) process the relative thickness of every information occurrence to figure its anomaly score. The proposed calculation consider the k-events characterized as dataset with limited arrangement of n focuses and for a given point x in a dataset, indicate the quantity of k-events in view of given likeness or separation measure as Nk(x),that the quantity of times x happens among every single other point in k closest neighbor and focuses those much of the time happened as a centers and focuses those happen rarely as an antihub.Uses invert closest neighbors for example , finding the cases to

©IJRASET (UGC Approved Journal): All Rights are Reserved

which question protest is closest. In this initially read the each trait in high dimensional dataset, at that point utilizing edge based exception identification strategy register the separation for each property utilizing dataset Set separation and contrast and separation from each case and allot the anomaly score. In light of that anomaly score utilizing reverse closest neighbor verify that specific example is an exception or not.

D. Performance Evaluation and Result Visualization

In this module, the exception recognized by above approach will be assessed on the premise of set assessment parameters for their execution assessment. The execution assessment will likewise give insights about actualized framework execution measurements, requirements and headings for future degree. With the assistance of appropriate representation of results, the framework execution will be made more justifiable and explorative for its evaluators.

E. Anti-Hub

The calculation Anti-Hub point is to get unimportant objects. For each object x in the requested informational index, discover the k number of turn around nearest neighbors Nk(x) for every last objection regarding distance measured by Euclidean. Locate the conflicting object score which is 1/ (Nk(x) + 1) for each protest. There might be a extra object if the score is higher. As indicated by the client positive threshold value, insignificant object is found. Anti-Hub, which expresses the conflicting object score of protest x from data set D as an element of Nk(x). The scores made by Anti-Hub point are unmistakable regardless of dimensionality. Distance of scores represents a dynamic quality of the Anti-Hub point calculation. Against hub point, which enhances outlier scores shaped by the Anti-Hub calculation by additionally considering the Nk scores of the neighbors of x, not with standing Nk(x) itself.

For each protest it discovers aNni which is the summation of exception score for each object. It finds the ct value by computing $(1 - \alpha) \bullet ai + \alpha \bullet aNni$ where ai is the counter hub point score additionally figures the cdisc. Cdisc is disc Score (ct, p) where $p \in (0, 1]$ yields the measure of restrictive things among ⌈np⌉ least individuals from ct, partitioned by ⌈np⌉. By looking at plate and cdisc value, relating ct and cdisc values are do led out to t and disc separately. Unessential object score is accomplished for each protest and there is a shot of finding unimportant objects if the score is higher.

Algorithm 2 AntiHub^2 *dist* (b, N, p, Q)

Input:
Distance measure *dist*
Ordered data set $D = (b_1, b_2, \ldots, b_n)$, where $b_i \in R^d$, for $i \in \{1, 2, \ldots, n\}$
No. of neighbors $N \in \{1, 2 \ldots\}$
Ratio of outliers to maximize discrimination $p \in (0, 1]$
Search parameter $Q \in (0, 1]$
Output:
Vector $s = (s_1, s_2 \ldots s_n) \in R^n$, where s_i is the outlier score of b_i, for $i \in \{1, 2, \ldots, n\}$
Temporary variables:
AntiHub scores $a \in R^n$
Sums of nearest neighbors' AntiHub scores $aNn \in R^n$
Proportion $\alpha \in [0, 1]$
(Current) discrimination score *cdisc*, *disc* $\in R$
(Current) raw outlier scores ct, $t \in R^n$
Local functions:
discScore(y, t): for $y \in R^n$ and $p \in (0, 1]$ outputs the number
of unique items among ⌈*np*⌉ smallest members of y, divided by ⌈*np*⌉
Steps:
$a := AntiHub_{dist}$ (D, N)
For each $i \in (1, 2, \ldots, n)$
$aNn_i := \Sigma_{j \in NNdist\,(k,i)}\, a_j$, where NN_{dist} (N, i) is the set
of indices of N nearest neighbors of b_i

©IJRASET (UGC Approved Journal): All Rights are Reserved

disc := 0
For each $\alpha \in (0, Q, 2* Q, \ldots, 1)$
For each $i \in (1, 2 \ldots n)$
$ct_i := \alpha' * a_i + \alpha * ann_i$
cdisc := discScore(ct, p)
If cdisc > disc
t := ct, disc := cdisc
For each $i \in (1, 2, \ldots, n)$
$s_i := f(t_i)$, where $f : R \rightarrow R$ is a monotone function

V. REGRSSION TECHNIQUE

Regression modeling has several applications in analytic thinking, commercial designing, marketing, monetary prediction, statistic prediction, medical specialty and medicine response modeling, and environmental modeling.

A. Need of Regression Technique

There are different purposes behind utilizing regression procedure in data mining. Some of these are recorded under:

1) A regression undertaking starts with data collection in which the objective values are known. For instance, a regression display that predicts youngsters' height could be produced in view of watched information for some kids over some stretch of time. The information may track age, stature, weight, formative points of reference, family history, et cetera. Stature would be the objective, alternate traits would be the indicators, and the information for every kid would constitute a case.

2) In the model form (preparing) process, a regression calculation assesses the estimation of the objective as a component of the indicators for each case in the assemble information. These connections amongst indicators and target are abridged in a model, which would then be able to be connected to an alternate data set index in which the objective values are unclear.

3) Regression models are tried by registering different insights that measure the contrast between the anticipated values and the normal values.

In regression techniques have seven types those are linear regression, logistic regression, polynomic regression, stepwise regression, ridge regression, lasso regression, elastic Net regression. In this paper I will explain logistic regression because detecting outliers using logistic regression rules. Logistic regression is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.
P (1-p) =probability of event occurrence/probability of not event occurrence

$$\text{Logit } (p) = B + B1X1 + B2X2 + B3X3 + \ldots\ldots\ldots\ldots + BiXi \quad \text{for } i=1, 2\ldots..n$$

B = shared parameter

X = independent given matrix

B. Importance of Logistic Regression

It is generally utilized for classification issues
Logistic regression doesn't require direct connection among dependent and independent factors. It can deal with different sorts of connections since it applies a non-direct log change to the predictable chances quantity
To keep away from over fitting and under fitting, we ought to incorporate every single critical variable. A decent way to deal with guarantee this training is to utilize a stage insightful strategy to appraise the calculated regression
It requires extensive sampling sizes since most extreme probability measures are less capable at low sample sizes than standard slightest square
The free factors ought not to be corresponded with each other i.e. no multi collinearity. Be that as it may, we have the choices to incorporate connection impacts of all out factors in the investigation and in the model.
In the event that the estimations of ward variable is ordinal, at that point it is called as Ordinal strategic regression

©IJRASET (UGC Approved Journal): All Rights are Reserved

In the event that reliant variable is multi class then it is known as Multinomial Logistic regression.

We apply logistic regression rule on the results of Anti-hub data set then obtained combination of data, prevention measures and Anti-hub calculation. It increases the efficiency of remove out irrelevant, redundant feature.in this we take the age attribute to find the which age group persons are work the hours per week, his or her marital status, education, relationship those attribute's basis will find the age of the person.

TABLE*:* Result of the regression rule on applying Anti-hub data

Rule	Attributes	Age	Class A (<50)	Class B (>50)
1	Hours per week>37.5, Marital status- married	41	A	
2	Hours per week>38.5, Work class-occupation	39	A	
3	Hours per week>35.5, Education ,work class	41	A	
4	Hours-per-week > 32.5, Education	44	A	
5	Hours per week>10.5 Fnlwgt <= 199729.5	56		B
6	Hours per week>24.5 Marital status- Widowed	42	A	
7	fnlwgt <= 182207 education-num <= 10.5	32	A	
8	Education-num <= 12.5 Marital status-unmarried	24	A	
9	Hours per week>34 Education	26	A	
10	Relationship-husband Hours per week>34	29	A	
11	Education Relationship-wife	21	A	
12	Work class, education	31	A	
13	Unmarried, educated	23	A	
14	Hours per week>11.5 widowed	53		B
15	Unmarried, education	20	A	
16	Hours per week>19 Relationship-wife	46	A	
17	Relationship-husband Hours per week<11	67		B
18	Married status-married	46	A	
19	Hours per week>38.5 Work class	38	A	

©IJRASET (UGC Approved Journal): All Rights are Reserved

20	Married status-widowed	45	A	
21	Hours per week>19	59		B

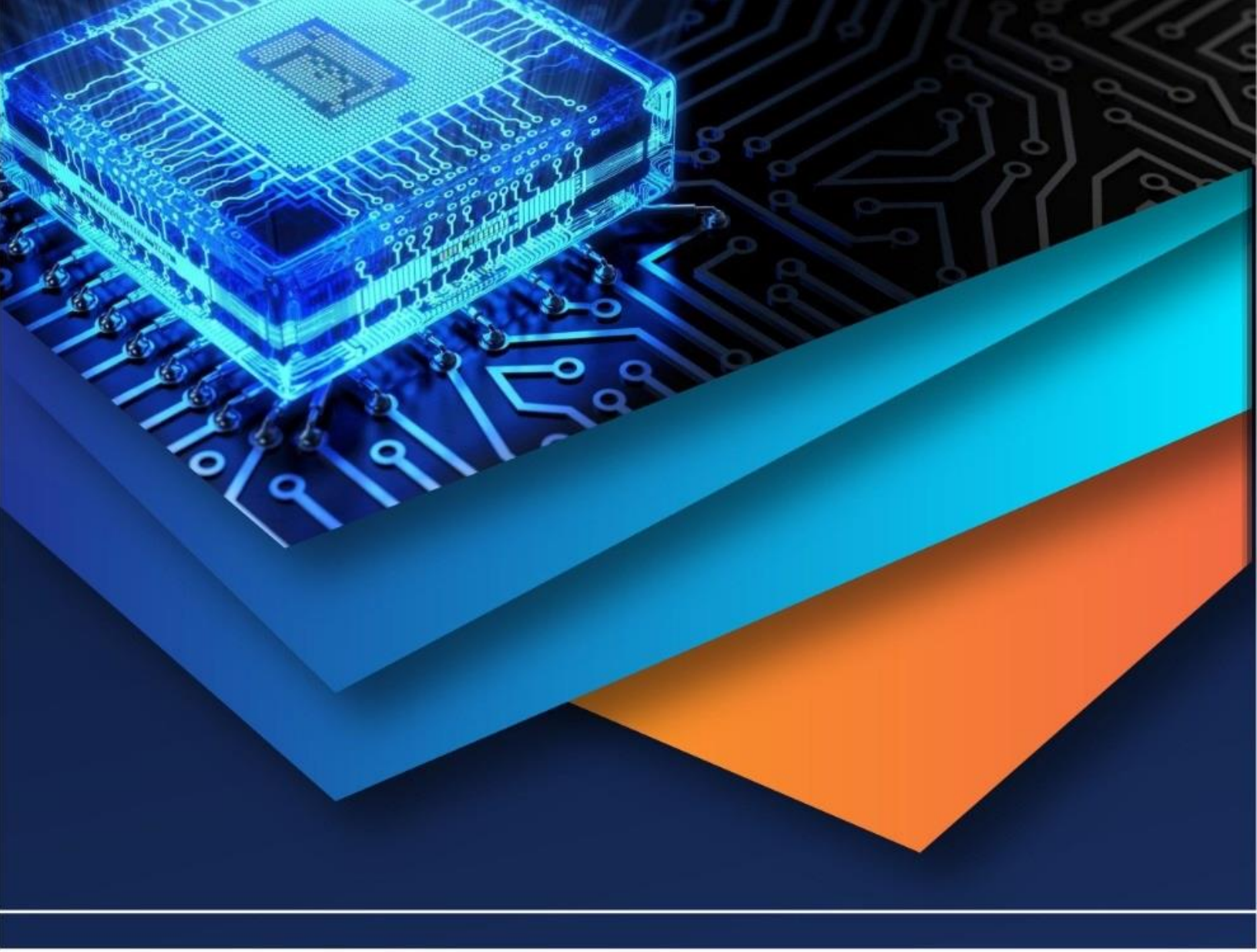
V. CONCUSION

Outlier detection find out the unmatching patterns from data set. Different techniques use that are Anit-hub, Logistic regression to detect the outliers. Outlier scores also plays an important role in outlier detection. The goal of this paper is locate the inconsistent objects in data which has high dimension through reduced calculation time, cost and increase the accuracy. We apply logistic regression rule on the results of Anti-hub dataset then obtained combination of data, prevention measures and Anti-hub calculation. It increase the efficiency of remove out irrelevant, redundant feature.

REFERENCES

[1] V. chandola, A.Banerjee and V.Kumar, "Anomaly detection: A survey" ACM comput sury, 2009
[2] S.Rmaswamy, R.Rastogi and K.Shim "Efficient algorithm for mining outliers from large datasets" SIGMOD Rec, 2000
[3] Milos Radovanovi, c. Alexandros and Nanopoulos "Reverse nearest neighbors in unsupervised distance based outlier detection", IEEE Transaction on knowledge and Data Engineering, 2014
[4] Miss.Gavale Swati S.[1] Prof. Kahate Sandip[2] "Outlier using Anti-hubs" IJSRD - International Journal for Scientific Research & Development, 2016
[5] H.P.Kriegel, M.Schubert, and A.Zimek "Angle based outlier detection in high dimensional data" in Proc 14th ACM, SIGKDDInt
[6] R.Lakshmi Devi and R. Amalraj "An Efficient unsupervised clustering adaptive Antihub technique for outlier detection in high dimensional data" Indian journal of science and technology, 2016
[7] Smita S. Patil, Prof. P. D. Chouksey "Outlier Detection Using Hub, Antihub & Semi supervised approach for Distance based Method International journal of research in advanced engineering technology, 2016.
[8] Unsupervised distance based outlier detection using nearest neighbor algorithm on distrusted approach: survey International journal of Innovative Research in computer and communication engineering, 2014
[9] "Unsupervised distance based outlier detection in reverse nearest neighbour" south Asian journal of engineering and technology, 2016.

©IJRASET (UGC Approved Journal): All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)