International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Finding Multiple Outliers from Multidimensional Data using Multiple Regression

L. Sunitha[1*] , Dr M. Bal Raju[2]

[1*]*CSE Research Scholar, J N T U Hyderabad , Telangana (India)*

[2]Professor and principal, *K MI T& Engineering, Hyderabad, Telangana(India)*

ABSTRACT

The knowledge of weather is useful for finding climate change over a period. In this present frame work uses 15 years of weather of Hyderabad city , data a real time the datasets collected from weather station. Weather data is a time series and multidimensional data. Outliers are the objects whose behavior is different from the rest. Outliers in weather data represent the cyclone, drought, seasonal change or heavy rains .In this paper multiple regression model is used on weather data. All the parameters are strongly related so regression model is well suited for weather data. We are used an Excel statistical tool is used for visual and models generated.

Key terms: *Outliers, multidimensional, multiple regression, climate change*

I .INTRODUCTION

Outliers are the exceptional or critical objects which are abnormal from normal characteristics, significant behavioral difference from whole database. Outliers from weather data are classified into three categories weekly , monthly and yearly. Due to heavy destruction of environment, pollution in environment leads, the future is danger. The environment scientist and research agencies are warning the world save our earth, serious causes may occur in near future temperature increasing year by year some times heavy rains or drought may occurred. Research is going on to get facts of climate change over the historical data.

II. RELATED WORK

Form statistical data analysis to find the outliers early of 19th century [1] many of the researchers have developed. The objects which are so far from remaining objects the objects are called abnormal objects generated by different models and mechanisms [2].An anomalous patterns in computer network the hacked computer is sending data to unauthorized destination [3].In credit card transactional [4] data Outliers gives the fraud detection or identify the theft. Bayesian weighted regression algorithm [5] that is able to automatically detect and eliminate outliers in real-time, without requiring any interference from the user, parameter tuning, sampling or model assumptions about the underlying data structure. We compared this algorithm to standard

approaches for outlier detection, such as thresh holding using Mahalanobis distance, mixture models. A clustering based framework for outlier detection [6] in evolving data streams that assigns weights to attributes depending upon their respective relevance. Incremental and adaptive to concept evolution. Experimental results on synthetic and real world data sets show that other existing approaches in terms of outlier detection. Weather forecasting data model utilizes the k-means unsupervised learning technique [7] for performing the clustering on the entire training dataset. This clustering is performed for finding the pattern level pattern similarity among two instance data. Using the extracted observations and available class labels the data is re-organized in terms of observation matrix and the transition matrix. The trained data model is used for prediction or the pattern recognition work.

III. PROPOSED WORK

The framework is two phase outlier detection , first building a model , next step is classification of objects either normal or outlier.

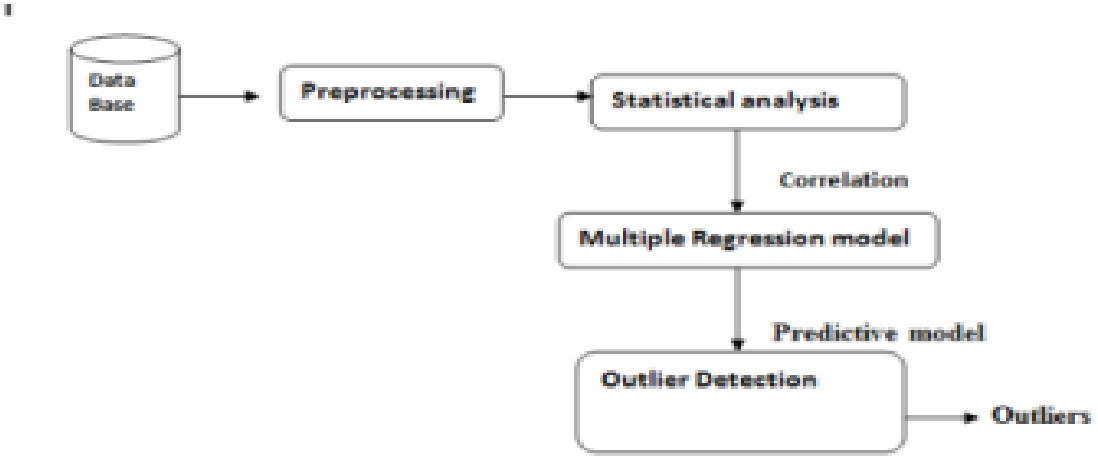

Fig1. Proposed work

A. Building a model using multiple regression

The linear regression for multi value attributes , system was developed and trained a model the equation for linear regression Y=a+ X b , predict for supervised classification. Where Y is dependent variable and X is a independent variable. In multiple regressions Y is temperature and the X is set of variables atmospheric pressure, humidity , wind speed and dew point .How the temperature is depend on other parameters. All the parameters are interrelated and these parameters influence on rain fall. Here we are collected and study of 15 years data from 2001 to 2015.In this period climate change which day is abnormal and which year is abnormal.

1 Hypotheses

Effect of each variable can be estimated separately by using multiple regressions. Let S denote weather data of 15 years.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

We always start a regression analysis by formulating a model for our data. One possible linear regression model with four quantitative predictors for temperature.

$Y_i = (B_0 + B_1X_{i1} + B_2X_{i2} + B_3X_{i3}) + C_i$ where Y_i is temperature for the month and year. The independent error term C_i follows a normal distribution with mean 0 and equal variance σ^2.

2. Key Points about Model

1. Because we have more than one predictor(X) variables like X_{i1} , $X_{i2,}$ X_{i3} are subscripted with a 1, 2 , 3 as a way of keeping track of the three different quantitative variables. We also subscript the slope parameter with corresponding numbers B_1 , $B_{2,}$ $B_{3.}$

2. The LINE conditions must still hold for the multiple linear regression model .The linear portions comes from the established regression equation.

3. We use the term linear in the parameters .This simply means that every B coefficient multiplies a predictor variable or transformation of one or more predictor variables.

3. $Y=B_0 +B_1X + B_1X^2 + C$ is a multiple linear regression model even though it represent a curved relationship between Y and X.

B. Correlation Analysis

1. Correlation Coefficient: A single summary that tells that whether relationship exists between two variables, strong that relationship is and whether the relationship is positive or negative.

2. The Coefficient of Determination: Which tells us much variation in one variable is directly related to variation in another variable.

3. Linear Regression: A process that to make predictions about variable "Y" based on knowledge of variable "X".

4. The Standard Error of Estimate: It shows how accurate the predictions are likely to be when to perform Linear Regression.

5. Using correlation analysis to find out if there is a statistically significant relationship between TWO variables. We use linear regression to make predictions based on the relationship between two variables.

C Outlier Detection

There are several ways to identify outliers, including residual plots and three stored statistics: leverages, Cook's distance, and DFITS. It is important to detect outliers because they can significantly affect the model, providing potentially misleading or incorrect results. If you identify an outlier in your data, we should examine the observation to understand why it is not normal and identify an certain method.

Leverage

Leverage (Hi) measures the distance from an observation's x-value to the average of the x-values for all observations in a data set. Use to identify observations that have unusual predictor values compared to the remaining data. Observations with large leverage can have a large effect on the fitted value, and thus the regression model. Investigate observations with leverage values greater than 3p/n, where p is the number of

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

model and n is the number of observations .Observations with leverage values greater than 3p/n or .99, whichever is smaller, with an X in the table of unusual observations.

Cook's distance (D)

Geometrically, Cook's distance is a measure of the distance between the fitted values calculated with and without the i^{th} observation. Use to identify observations that have unusual predictor values compared to the remaining data and observations that the model does not fit well. Observations with large Cook's Distances can have a large effect on the fitted value, and thus the regression model. Investigate observations where D is greater than F (0.5, p, n-p), the median of an F-distribution, where p is the number of model terms and n is the number of observations. A different way to examine distance values is to compare distance values to each other graphically, using a line plot. Observations with large distance values relative to other observations can be influential.

DFITS: DFITS represents approximately the number of standard deviations that the fitted value changes when each observation is removed from the data set and the model is refit. Use to identify observations that have unusual predictor values compared to the remaining data and observations that the model does not fit well. Observations with large DFITS values can have a large effect on the fitted value, and thus the regression model. Investigate observations with DFITS values greater than 2*sqrt (p / n), where p is the number of model and n is the number of observations. A different way to examine DFITS values is to compare DFITS values to each other graphically, using a time series plot or a line plot. Observations with large DFITS values relative to other observations can be influential. To determine how much effect the unusual observation has, fit the model with and without the observation and compare the coefficients, p-values, R^2, and other model information

IV. EXPERIMENTAL RESULTS

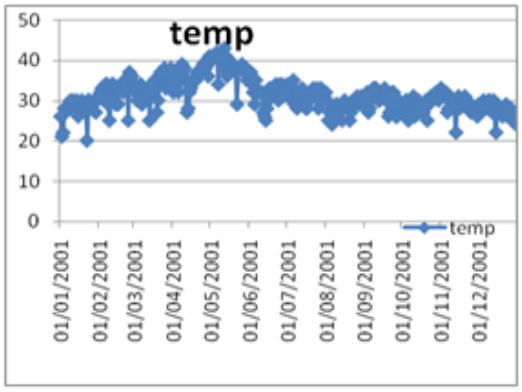

Fig2. Temperature

	Temp	humidity	Dew point	pressure	Wind speed
Temp	1				
humidity	-0.1403	1			
Dew point	-0.1391	0.75013	1		
Pressure	-0.1581	0.31663	0.388	1	
Wind speed	0.30270	0.27352	0.3017	0.12141	1

Table1. Correlation matrix

Negative sign indication of inverse relation as temperature increases humidity, dew point And Atmospheric pressure are decreases ,where as temperature and wind speed are in positive relation.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

	df	SS	MS	F	Significance F
Regression	4	2679.103157	669.7757892	74.62769	5.12969E-46
Residual	360	3230.962597	8.974896102		
Total	364	5910.065753			

Regression Statistics	
Multiple R	0.673284415
R Square	0.453311904
Adjusted R Square	0.447237591
Standard Error	2.995813095
Observations	365

Table 5 : Residuals

Temp=constant +humidity*x1+Dew point*x2+pressure*x3 + wind speed*x3 –1

Predictive model temp=178.6-0.176*humidity+0.321*dewpoint-0.138*pressure+0.002*windspeed --- (2)

	Coefficients	Standard Error	t Stat	P-value
Intercept	178.6035	49.25638	3.62599	0.000329
Humidity	-0.176087	0.012337	-14.272	6.13E-37
Dew point	0.321179	0.058095	5.52845	6.22E-08
pressure	-0.138132	0.048304	-2.8595	0.004489
Winds peed	0.0021623	0.014012	0.15431	0.877452

Table2: Regression Analysis

V. CONCLUSION

Outliers Detection is an important and essential task in data mining. Outlier detection as a branch of data mining, outlier detection has important applications in different domains and it need more attention from data mining. In this paper we are started from the review of existing outlier detection schemes and clustering and other methods multiple regression is statistical model and it well suited for multidimensional that is more than two variables. Three measures are used for outlier detection leverages, Cook's distance, and DFITS.

REFERENCES

[1]. Edgeworth F.Y 1887 , On discordant observations, XLI. On discordant **o**bservations: Philosophical Magazine Series

[2] Karanjit Singh and Dr. Shuchita Upadhyaya, Outlier Detection: Applications and Techniques, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814.

[3] Kumar v 2005 Parallel and distributed computing for cyber security , IEEE Distributed Systems Online (Volume: 6, Issue: 10, 2005).

[4] Aleskerov E, A neural network based database mining system for credit card fraud detection.

[5] Automatic Outlier Detection : A Bayesian Approach , IEEE International Conference on Robotics and Automation Roma, Italy, volume-10 , Issue -14, Apil-207, pg : 2489-2494.

[6] A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering. Science Direct, volume -6, Issue -1, April -2012, Pg-214 – 222.

[7] A Weather Forecasting Model using the Data Mining Technique, International Journal of Computer Applications ,volume :139,Issue-19 ,pg:1-9.

[8]Peters J., Suraj Z., Shan S., Ramanna S., Pedrycz W., Pizzi N., "Classification of meteorological volumetric radar data using rough set methods," Pattern Recognition Letters, pp.911–920. 2003.

[9]. Data Mining: Estimation of Missing Values Using Lagrange Interpolation Technique, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013,

[10]L. Sunitha, A Comparative Study between Noisy Data and Outlier Data in Data Mining, International Journal of Current Engineering and Technology ISSN 2277 - 4106 © 2013 INPRESSCO.

[11]Automatic Outlier Identification in Data Mining Using IQR in Real-Time Data,(IJA RCCE) Vol. 3, Issue 6, June 2014.

Author Profile

Lingam sunitha received her MCA from Kakatiya University in 1999, and M.Tech (CSE) from JNTU, Hyderabad in 2009. She is now working as Assistant Professor in GITAM University, Hyderabad and also pursuing PhD in Computer Science and Engineering from JNTU Hyderabad, Telangana, India. Her area of specialization is Data Mining.

Dr M. Bal Raju He received both Graduation B.Tech (ECE) and Post Graduation M.Tech (CSE) from Osmania University and PhD from JNTU Hyderabad in 2010. Now he is working as Professor and Principal in Krishna Murthy Institute of Technology & Engineering, Hyderabad, India. His area of interest includes Data Base, Data Mining and image processing; He was published 30 research papers in various National and International Journals. He was attended and presented 10 research papers in National and International conferences.