www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 7 July 2016, Page No. 17278-17281

A study on least square fit for outlier detection

T.Jagadeeswari[1], Dr.H.Venkateswara Reddy[2]
[1]Department of Computer Science and Engineering1,2
Vardhaman College of Engineering, JNTU Hyderabad
jagadeeswari2003@gmail.com
[2]Department of Computer Science and Engineering1,2
Vardhaman College of Engineering, JNTU Hyderabad
h.venkateswara reddy@vardhaman.org

Abstract: Data mining is the procedure of mining knowledge from data. This is extensively studied field for research area, where most of the work emphasized over knowledge discovery. Outlier detection is an important task in data mining and it has many real time applications . In most of the applications data contains unwanted and unrelated data. Finding and removing anomalous data is very important and there by improve the quality and accuracy. The outliers are pinpoint or group that depends on data and applications. This paper focus on outlier concept, taxonomy and outlier detection using least square fit

Keywords :outlier detection , outlier

1. Introduction

Data mining is a process of analyzing data from different perspective and summarizing it into useful information. Although they are many data mining techniques , they all have their origin based on science discipline like statistics or machine learning.
"An outlying observation , or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs"
A further outlier definition from Barnett & lewis (.Barnett and Lewis 1994)is:
An observation (or set of observations) which appears to be in inconsistent with the remainder of that set of data.[1]
Detecting these anomalies is an a significant issue in different applications to improve the accuracy of the data.

The most exhaustive list of applications that utilize outlier detection is: fraud detection, satellite image analysis, pharmaceutical research, intrusion detection, structural defect detection , network performance.

2. Literature Survey

Each data instance can be described with a set of attributes (also referred to as variables, characteristics , features) A key aspect of outlier detection techniques depends nature of the desired anomaly. So the outliers can be classified in to following categories.

2.1 point anomalies: If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point outlier. This is the simplest type of outlier and is the focus of majority of research outlier detection. As in the real life credit card fraud detection with data set corresponding to an individual's credit card transactions assuming data definition by only one feature: amount spent. A transaction for which the amount spent is very high compared to the normal range of expenditure for that person will be a point outlier.

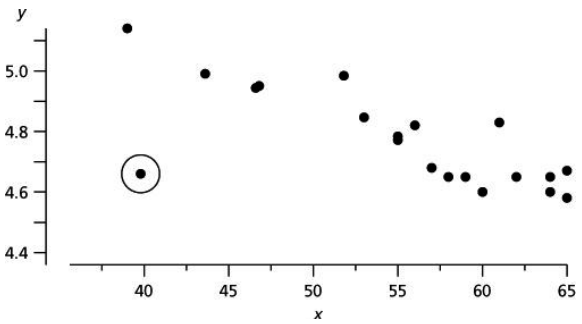


Figure 1: point outlier

2.2 contextual outliers: If the data is anomalous in specific context then it is termed as a contextual outlier[2]
The notion of a context is induced by the structure in the data set and has to be specified as a part of the problem formulation
Contextual attributes: The contextual attributes are used to determine the context(or neighborhood) for that instance .for example special data set and in time series data , time is a contextual attribute which determine the position of an instance on the sequence.
Behavioral attribute: The behavioral attribute define the non –contextual characteristic of an instance .For example in a spatial data set describing the average rainfall of the entire world , amount of rainfall at any location is behavioral attribute.

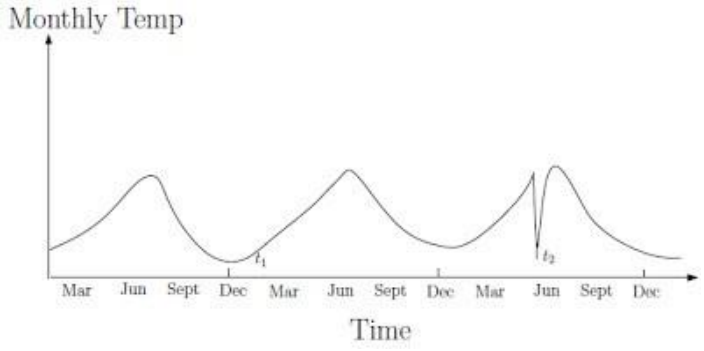


Figure 2: contextual outlier

2.3 collective outlier: If a collection of related data instance is inconsistent with respect to the entire data set is termed as collective outlier. The individual data instance in a collective outlier may not be outlier by themselves, but their occurrence together as a collection is anomalies.

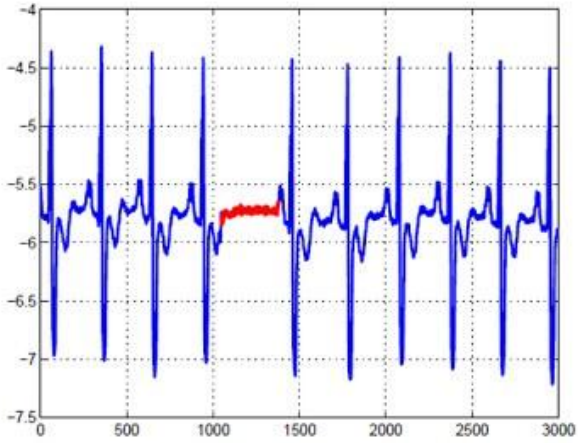


Figure 3: collective outliers

3. Taxonomy Of Outlier Detection Method in Statistics

The Outlier Detection Method is classified into various methods which are charted in the following figure

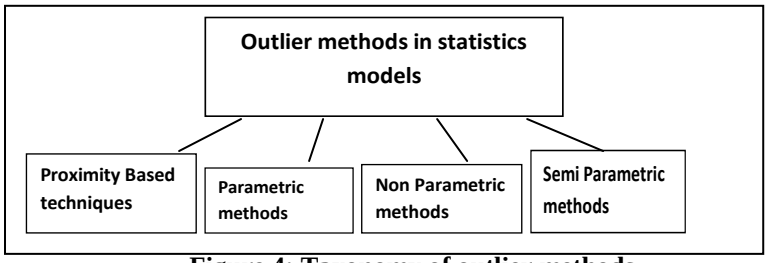


Figure 4: Taxonomy of outlier methods

3.1 proximity based techniques: objects far from the other are outliers. The proximity of an outlier deviates significantly from that of most of the other in the data set. For example density based outlier detection , an object is an outlier if its density is relatively much lower than that of its neighbor. Distance-based outlier detection: An object is an outlier if its neighborhood does not have enough other points .

3.2 parametric methods: parametric methods involve assumption of some underlying distribution such as normal distribution .Assumes that the normal data is generated by a parametric distribution with parameter θ .so outliers could be identified by calculating the probability of the occurrence of an observation on calculating how far the observation is from the mean.

3.3 Non parametric method: A statistical method is called non-parametric if it makes no assumption on the population distribution or sample size.

This is in contrast with most parametric methods in elementary statistics that assume the data is quantitative, the population has a normal distribution and the sample size is sufficiently large. In general, the non-parametric methods are not as powerful as the parametric ones. However, as non-parametric methods make fewer assumptions, they are more flexible, more robust, and applicable to non-quantitative data.

3.4 semi parametric method: A semi parametric model for observational data combines a parametric form for some components of a data generating process(usually the observational relation between the dependent and explanatory variables) with weak non parametric restrictions on the remainder of the model..

4. Least square regression model

The basic idea of the method of least square is easy to understand . It may seem unusual that when several people measure the same quantity , they usually do not obtain the same results .In fact , if the same person measures the same quantity several times, the result will vary. The method of least squares gives a way to find the best estimate , assuming that the errors (i.e the difference from the true values) are random and unbiased .

The regression analysis or least square estimation is a statistical technique to estimate a linear relationship between two variables. This is highly sensitive to the outlier and influential observation. A graph is used to estimate the best fit line as opposed to calculating the best fit line from the data points. The basic is to draw a line through the data with a many data points above it as below it with in errors

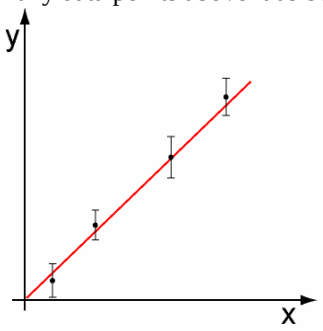


Figure 5(a)

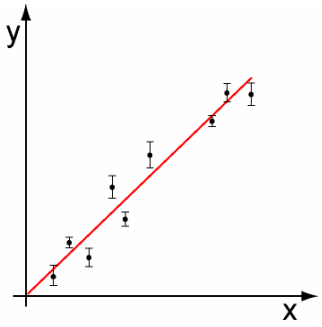


Figure 5(b)

Figure 5(a) data where it's easy to estimate the best fit of the line
Figure 5(b) data where its best to use a least square fit

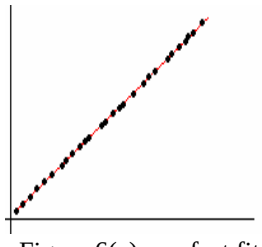


Figure 6(a) ; perfect fit

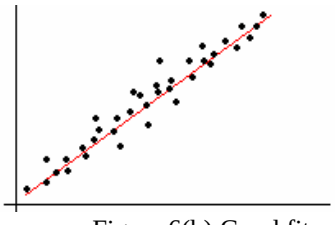


Figure 6(b):Good fit

DOI: 10.18535/ijecs/v5i7.17

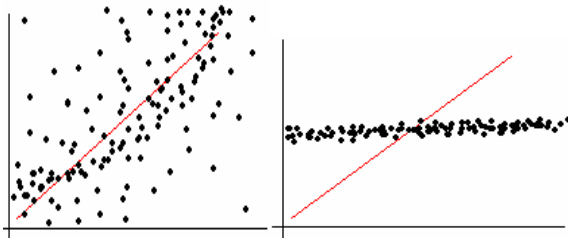


Figure 6(c): Bad fit Not linearly related

Figure (d) bad fit , zero relation ship between two

Figure 6 : Graph with different χ^2 values,

Figure 6(A) is almost a perfect fit, for the points fall very close to the best fit line. Figure 6(b) is a good fit , for the points fall fairly close to the best fit line . Figure 6(c) is bad fit for the points do not fall close to the best fit line(and if you look closely you will note that they seem to fall in a parabola and not a line. Figure 6(d)is a bad fit for the points do not fall along the chosen best fit line. The data in figure 6(d) fall on a horizontal line , which indicates that the variables are unrelated for changes in the independent variable produce no change in the independent variable.

The method of least square calculates the line of the best fit by minimizing the sum of squares of the vertical distance of the points to the line. The least square model for a set of data (x_1, y_1), $(x_2,y_2)\ldots(x_n,y_n)$ passes through the points (x_a, y_a) where x_a is the average of the $x_{i's}$ and y_a is the average of the $y_{i's.}$

The least square fit or regression equation (1)

$$Y= a+bX \qquad (1)$$

$$\text{Slope (b)} = \frac{(n\,\Sigma xy - (\Sigma x)(\Sigma y))}{n\Sigma\chi^2 - (\Sigma x)^2}$$

$$\text{Intercept (a)} = \frac{(\Sigma y - b(\Sigma x))}{n}$$

Where x and y are variables
b =slope of the regression line
a= The intercept point of the regression line and the y-axis
n= Number of values or elements
x = First score
y = Second score
Σxy= sum of the product of first and second scores
Σx= sum of the first score
Σy = sum of the second score
Σx^2= sum of the square of first score

The correlation coefficient r is a statistical measurement of a linear relationship(correlation) between dependent variable and independent variable and its values between -1 and +1.

The calculation of the least square fit is the distance from each point to the best fit line is given in figure 5(b) called deviation..If the line is really good fit those deviations will be a small as possible. The least square fit method attempt to minimize the square of deviations.
The sum of all of the square of the deviations is called the residual χ^2 given by the equation (2)

$$\sum_{i=1}^{n}(y_i - y) \qquad (2)$$

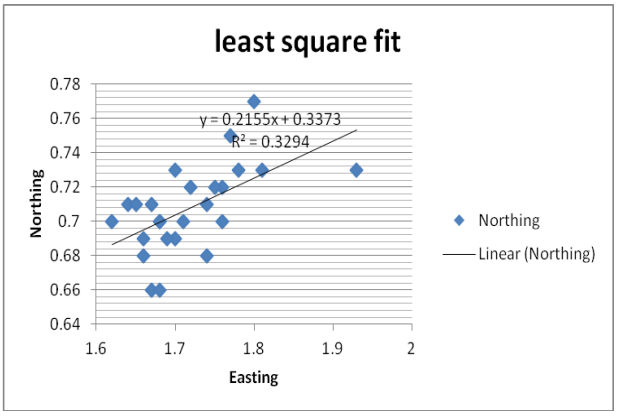
Where y_i are the data points and y is y-value from the best fit line .In the data follows a linear relation then y can expressed in the general form y = a+bx.

5. Experiment results

Least square fit using sample data(climate data).Excel uses the least square fit method to calculate the best fit line.

Table 1 Instance of sunshine database

Easting	Northing
1.74	0.68
1.93	0.73
1.67	0.66
1.68	0.7
1.68	0.7
1.68	0.66
1.76	0.72
1.81	0.73
1.72	0.72
1.67	0.71
1.64	0.71
1.7	0.69
1.76	0.7
1.78	0.73
1.62	0.7
1.71	0.7
1.75	0.72
1.66	0.69
1.69	0.69
1.74	0.71
1.77	0.75
1.8	0.77
1.66	0.68
1.65	0.71
1.7	0.73



6. Conclusion

Calculating the best fit line by using least square fit method is good for data sets and give better results for more data points. The analytic method is good for when error is small. This method will give slope, intercept and correlation coefficient (an indication of how good the data fit the line).This article gives deviation and method of using least square fit and at end expressed the experiment results through excel spread sheet.

Jagadeeswari Tanukonda received the Master of Computer Applications degree from Andhra University Vizag in 2002 and the Master of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad in 2010. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering ,Vardhaman College of Engineering, Hyderabad. Her research interests include Data Mining, Database Management Systems.

References

[1] Hodge, V. and Austin, J. 2004. A survey of outlier detection methodologies. Artificial Intelligence Review 22, 2, 85-126

[2] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. ACM Computing Surveys, (3), 1–58. Doi:10.1145/1541880.1541882

[3]Barnett, V, & Lewis, T. (1994). Outliers in statistical data (3rd ed.). New York: Wiley.

[4]Rorabacher, David B. (1991) "Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's' Q' Parameter and Related Subrange Ratios at the 95% Confidence Level." Analytical Chemistry 63, no. 2 (1991): 139–46

[5]P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection* (John Wiley & Sons, New York,

[6]. *USP 29–NF 24* (US Pharmacopeial Convention, Rockville, MD, 2006).

[7] V. Barnett and T. Lewis, *Outliers in Statistical Data* (John Wiley & Sons, 2d ed., New York, NY, 1985).