Outlier Detection by using Statistical Approach

Nan Khaing Pwint Phyu, Thandar Aung
ComputerUniversity, Mandalay
phyupwint5@ gmail.com

Abstract

Outlier analysis is that the user do depends on the kinds data they have. An outlier is a data value, that has a very low probability of occurrence that is, it is unusual. There are three kinds of outlier detection in outlier analysis. This paper uses statistical approach in outlier analysis. The statistical approach to outlier detection assumes a distribution or probability model for the given data set. Therefore, this paper was distribution model for input statistical data in any fields and then identifies outliers by using Inter Quartile Range formula and hypothesis testing. Input data must be numerical data or statistical data. The case study in this paper is to find outlier detection for house selling price with two containing data set. The main objectives of this paper is to detect the fraud detection or inconsistent data, to trace whether outlier data are real or not and then to give decision markers in making better decisions.

Keywords: Statistical Data, Outlier Detection, Statistical Approach, Inter Quartile Range, Hypothesis Testing.

1. Introduction

Outliers can be caused by measurement error or execution error. For example, the display of a person's age as –999 could be caused by a program default setting of an unrecorded age.

Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. This, however, could result in the loss of important hidden information since one person's noise could be another person's signal. In other words, the outliers may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity.

Outlier mining has wide application. As mentioned above, it can be used in fraud detection, for example, by detecting unusual usage of credit cards or telecommunication services. In addition it is useful in customized marketing for identifying the spending behavior of customers with extremely low or extremely high incomes, or in medical analysis for finding unusual responses to various medical treatment [3].

2. Outlier mining

Outlier mining can be describe as follow: Given a set of n data points or objects, and k, the expected number of outliers, find the top k objects that are considerably dissimilar, exceptional, on inconsistent with respect to the remaining data.

The outlier mining problem can be views as two sub-problems: (1) define what data can be considered as inconsistent in a given data set, and (2) find an efficient method to mine the outliers so defined.

The problem of defining outliers is nontrivial. If a statistical model is used for data modeling, analysis of the residuals can give a good estimation for data "extremeness".

In general, a person must check that each outlier discovered by this approach is indeed a "real" outlier [3].

There are two basic procedures for detecting outliers:

Block procedures: In this case, either all of the suspect objects are treated as outliers, or all of them are accepted as consistent.

Consecutive (or sequential) procedure: An example of such a procedure is the inside-out procedure. Its main idea is that the object that is least "likely" to be an outlier is tested first. If it is found to be an outlier, then all of the more extreme values are also considered outliers [3].

3. Statistical-based Outlier Detection

The statistical approach to outlier detection assumes a distribution or probability model for the given data set (e.g., a normal distribution) and identifies outliers with respect to the model using a box-plotting method and then check by using hypothesis test to examine whether the identified outliers are really extremeness for the model we've assumed or not. Application of these test requires knowledge of the data set parameters (such as assumed data distribution), knowledge of distribution parameters (such as the mean and the variance), and the expected number of outliers [3].

3.1 Data Set

This paper can detect outliers from statistical data set. The statistical data are the results of experiments, observations, and research from which we can generalize a unique characteristic of our works (e.g., marketing survey data, market research data, etc). If the number of attribute statistical data exceed than one thousand statistical data sets, the user can use one of the sampling methods that are random sampling, systematic sampling, etc. The resulting outlier data can be different according to the user's choice sampling method. The fields of the data sets can be chosen by the user. The data sets fields depend on the user objective.

3.2 Distribution Parameters

3.2.1 Measure of Central Tendency

The solution of many statistical-based outlier detection problems in which large set of data is collected can be somewhat facilitated by the determination of single numbers that describe unique characteristics about the data. The most popular measure of this type is called the arithmetic mean or average [1].

The common expressions for the arithmetic mean of the population and of a sample are:

$$\mu=\sum x_i / n \text{ (population)}$$

$$\bar{x}=\sum x_i / n \text{ (sample)}$$

3.2.2 Measure of Variation

Measure of variation indicates the degree to which data are dispersed, spread out, or bunched together. It is reasonable to define this variation in terms of how much each number in the sample deviates from the mean value of the sample, that is $x_1-\bar{x}, x_2-\bar{x},\ldots,x_n-\bar{x}$. If the user wanted an average deviation from the mean, the user might try adding $x_1-\bar{x}$ through $x_n-\bar{x}$ and dividing by n. The resulting formula for the standard deviation of the entire population is

$$\sigma=\left[\sum(x_i-\mu)^2 / n\right]^{1/2}.$$

Another common measure of variation is called the variance; it is the square of the standard deviation [1].

3.3 Data Set Parameter

3.3.1 Normal Distribution

Among the many continuous distributions used in statistics, the normal distribution is by far the most useful.

The normal distribution is a theoretical frequency distribution for a specific type of data set. Its graphical representation is a bell-shaped curve that extends indefinitely in both directions. As can be seen in Figure 1.

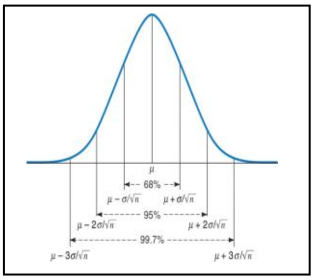


Figure 1 A normal Distribution Curve and it's Parameters

The location and shape of the normal curve can be specified by two parameters: (1) the population mean (μ), which locates the center of the distribution, and (2) the population standard deviation (σ), which describes the amount of variability or dispersion of the data [1].

3.4 Identifying Outliers by using Box-Plots

A method for displaying a set of data uses not the individual values, but rather a set of summary statistics taken from that data. The plot is called a box-plot. A box-plot is a graphical display that uses summary statistics to display the distribution of a set of data. A box-plot summarizes a sample or a population by using the quartiles and the median. The median is a measure of center. The quartiles are the medians of two sub data sets which drive from dividing a data set by using it's median. In a box-plot diagram, we describe the median of a data set as median quartile (q_M), the quartile of upper part of the same data set as upper quartile (q_U), and the quartile of lower part the same data set as lower quartile (q_L).

Inter quartile range (use to define the expected outliers):

IQR = $q_U - q_L$
IQR = Inter quartile range
q_U = upper quartile
q_L = lower quartile

q_M = middle quartile
Upper Inner Fence = q U + 1.5 IQR
Upper Outer Fence = q U + 3 IQR
Lower Inner Fence = q L – 1.5 IQR
Lower Outer Fence = q L – 3 IQR

By empirical rule, we can assume as possible outliers if the data values in a data set which fall between inner fences and outer fences of a box-plot and as probable outliers if the data values in a data set which fall beyond the outer fences of a box-plot [4].

3.5 Checking Outliers by using Hypothesis Test

3.5.1 Hypothesis and Hypothesis Test

In statistic, a hypothesis is an idea, an assumption, or a theory about the characteristics of one or more variables in one or more populations.

A hypothesis test is a statistical procedure that involves formulating a hypothesis and using sample data to decide on the validity of the hypothesis.

3.5.2 Procedure for Hypothesis testing using Traditional Methods

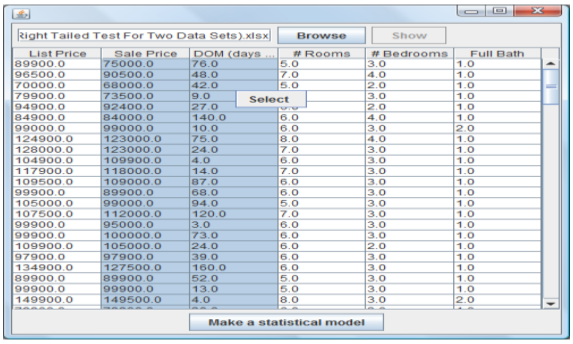
Step 1: State the hypothesis and identify the claim.
Step 2: Find the critical value(s) from the significance bounds table.
Step 3: Compute the Test Value.
Step 4: Make the decision to reject or not reject the null hypothesis.
Step 5: Summarize the results.

3.6 System Implementation

In this paper, the system implementation is to find outlier data based on the user statistical input data sets and then to help decision makers in making better decisions according to the obtained outlier data.

3.6.1 Data Set of the Case Study

The survey of the original house selling price data contain six fields. The case study of this paper uses two fields that are sale price and day on market to detect outlier data in the system. The data set use in case study is as follow in Figure 2:



Right Tailed Test For Two Data Sets).xlsx

List Price	Sale Price	DOM (days ...	# Rooms	# Bedrooms	Full Bath
89900.0	75000.0	76.0	5.0	3.0	1.0
96500.0	90500.0	48.0	7.0	4.0	1.0
70000.0	68000.0	42.0	5.0	2.0	1.0
79900.0	73500.0	9.0	[illegible]	3.0	1.0
94900.0	92400.0	27.0	[illegible]	2.0	1.0
84900.0	84000.0	140.0	6.0	4.0	1.0
99000.0	99000.0	10.0	6.0	3.0	2.0
124900.0	123000.0	75.0	8.0	4.0	1.0
128000.0	123000.0	24.0	7.0	3.0	1.0
104900.0	109900.0	4.0	6.0	3.0	1.0
117900.0	118000.0	14.0	7.0	3.0	1.0
109500.0	109000.0	87.0	6.0	3.0	1.0
99900.0	89900.0	68.0	6.0	3.0	1.0
105000.0	99000.0	94.0	5.0	3.0	1.0
107500.0	112000.0	120.0	7.0	3.0	1.0
99900.0	95000.0	3.0	6.0	3.0	1.0
99900.0	100000.0	73.0	6.0	3.0	1.0
109900.0	105000.0	24.0	6.0	2.0	1.0
97900.0	97900.0	39.0	6.0	3.0	1.0
134900.0	127500.0	160.0	6.0	3.0	1.0
89900.0	89900.0	52.0	5.0	3.0	1.0
99900.0	99900.0	13.0	5.0	3.0	1.0
149900.0	149500.0	4.0	8.0	3.0	2.0

Figure 2 Select Two Data Set from Houses Modify Excel Data Form

3.6.2 Modeling of the System

In modeling of the system contains two steps, this steps are used to find mean and standard deviation of the selected two data sets of the case study.

3.6.2.1 Measure of Central Tendency

$$\mu = \sum x_i \, / \, n$$

$$\text{mean for sale price } = \frac{53000+ 65000+\ldots+159900}{150}$$

$$= 104883.9867$$

$$\text{mean for day on market } = \frac{1+ 2+ 2+\ldots+ 350}{150}$$

$$= 52.5$$

3.6.2.2 Measure of Variation

$$\sigma = \left[\sum (x_i - \mu)^2 \div n\right]^{1/2} .$$

Standard deviation for

$$\text{sale price } = \frac{[\Sigma\,(53000- 104883.9867)2+ \ldots +(159900- 104883.9867)2]½}{150}$$

$$= 21366.2993$$

Standard deviation for

$$\text{day on market } = \frac{\left[\Sigma(1- 52.5)2+ (2- 52.5)2+\ldots+ (350- 52.5)2\right]^{½}}{150}$$

$$= 59.4235$$

3.6.3 IQR Formula of the Case Study

The result of the IQR for sale prices and Day on Market data set formula in section 3.4 is as follows:

Table 1 Outlier for Sale Price

Upper Outer Fence	207300
Upper Inner Fence	163275

Upper quartile	119250
Median	101500
Lower quartile	89900
Lower Inner Fence	45875
Lower Outer Fence	1850
Inter quartile range	29350
Possible outlier	0
Probable outlier	0

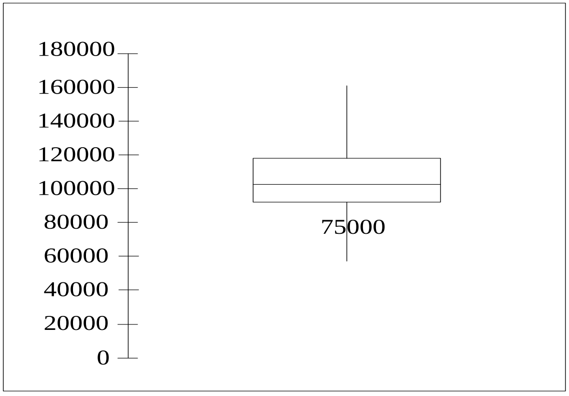


Figure 3 Box Plot Diagram for Data Set with no Outlier

Table 2 Outlier for day on market

Upper Outer Fence	276
Upper Inner Fence	177
Upper quartile	78
Median	27.5
Lower quartile	12
Lower Inner Fence	–54
Lower Outer Fence	–186
Inter quartile range	66
Possible outlier	6
Probable outlier	1

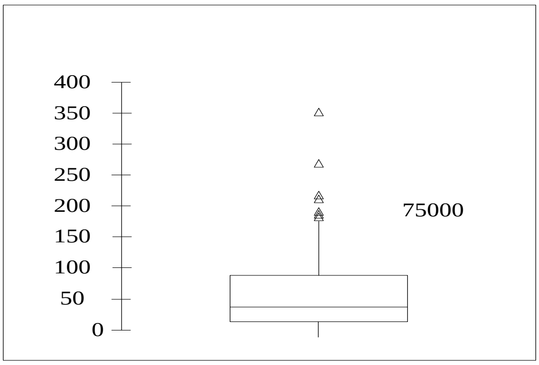


Figure 4 Box Plot Diagram for Data Set with seven outliers (179,181,196,202,215,263,350 are the most extreme data in the data set)

3.6.4 Hypothesis testing of the Case Study Hypothesis

There are five steps in hypothesis testing.

3.6.4.1 Stating the Hypotheses and identifying Claim

H_0 : O_i is in distribution F, Where i = 1,2,3, …,n (Claim)

(the entire data set of n objects come from an initial distribution model or there is no outlier in given data set)

H_A : O_i is in distribution G, Where i = 1,2,3,…n

(the objects come from another distribution model G or there is outlier in given data set for distribution F)

H_0 = Working hypothesis or Null hypothesis

H_A = Alternative hypothesis.

3.6.4.2 Finding critical value from Significance Bounds Table.

To find critical value, we use number of data in the data set (n) and significance level that we define for the hypothesis testing (α).

Then we find the appropriate value in the Significance Bounds Table.

n = 150,

α = 0.05,

Critical Value (C.V) = 3.283.

3.6.4.3 Computing the Test Value

Here, number of data in the data set is greater than 25. So, we use formula for large sample.

T (Test value) = (Value of suspected outlier – Sample mean) ÷ Sample Standard Deviation

The suspected outliers are,

Sample mean : $\bar{x}$ = 52.5

Sample Standard Deviation : s = 59.2435

$$T(\text{ Test value}) = (x_i - \bar{x}) \div s = (179 - 52.5) / 59.2435$$

$$= 2.1288., \text{etc...}$$

3.6.4.4 Making Decision for Hypothesis testing

By using the Formulae for finding Test value (test statistic) for suspected outliers and consecutive procedure for outlier detection, we can make the decisions for suspected values.

Table 3 Hypothesis Testing Decision Table

Suspected outliers	Critical value N = 150, α = 0.05	Comparison	Test Value or Test Statistic	Decision
179	3.283	Is greater than	2.1288	There is no enough evidence to reject Ho.
181	3.283	Is greater than	2.1625	There is no enough evidence to reject Ho.
196	3.283	Is greater than	2.4159	There is no enough evidence to reject Ho.
202	3.283	Is greater than	2.5159	There is no enough evidence to reject Ho.
215	3.28	Is greater than	2.7346	There is no enough evidence to reject Ho.
263	3.283	Is less than	3.5424	There is no enough evidence to reject Ho. The outliers are from this value on.
350	No need to claclate	No need to claclate	No need to claclate	No need to claclate

3.6.4.5 Summarize the Results

In table 3 hypothesis testing decision table for day on market to find two real outliers from expected seven outliers as shown in Figure 5.

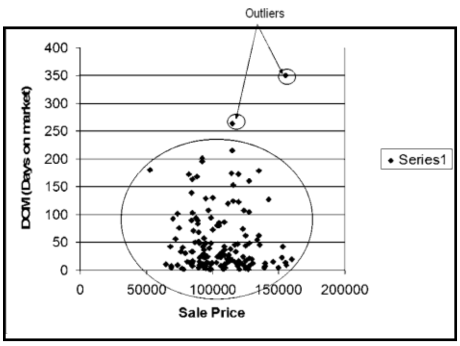


Figure 5 Scatter Diagram

4. Conclusion

In this paper, the effectiveness, of Outlier Detection system in selling house price application has now been demonstrated. This system helps statistical data detect outliers by utilizing statistical method in outlier analysis. The hierarchical approach to constructing the system uses the use of over 150 records of datasets. This statistical data sets are used a branch of applied mathematics. This system can test the possible outlier depending on any statistical data set. This paper present an approach to discover outliers from a statistical data set by using statistical methods for outliers detection and statistical hypothesis test using statistical significance testing method [2].

REFERENCES

[1] Arvid R. Eide, Roland D. Jenison, Lane H. Mashaw, Larry L. Northup, "Engineering Fundamentals & Problem Solving", ISBN 0-07-113022-5, McGraw-Hill Publish.

[2] Jacobs, Roberts, " Outliers in Statistical Analysis: Basic Methods of Detection and Accommodation", Paper presented at the Annual Meeting of the Southwest Education Research Association, New Orleans, LA, February 1-3, 2001.

[3] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", ISBN 1-55860-489-8, Morgan Kaufmann Publish.

[4] Marilyn K. Pelosi, Theresa M. Sandifer, "Elementary Statistic from Discovery to Decision", ISBN 0-471-42903-1, John Wiley & Sons, Inc.