ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 6, June 2014*

# Automatic Outlier Identification in Data Mining Using IQR in Real-Time Data

**L.Sunitha[1] , Dr M.BalRaju[2], Dr J.Sasikiran[3] , E.Venkat Ramana[4]**

Department of computer Science and Engineering, Vidya Vikas Institute of Technology, Hyderabad, India [1, 3,4]

Malla Reddy College of Engineering, Hyderabad, India[2]

**Abstract:** Some of the Real-Time databases containing exceptional values or extreme values called outliers. The separation of outliers is very much useful for improving the data quality and also removing which is not required and unrelated for knowledge discovery. The outlier will influence the results of data mining techniques like clustering, classification and association. Outlier Identification and removing is a part in the in preprocessing. In this paper we are using various measures central tendency methods like mean ,mode, median and Inter-Quartile – Range(IQR) on real time databases and the experimental results are generated by using Weka a data mining tool.

**Key Terms:** Outlier, Preprocessing, IQR

## 1. INTRODUCTION

Outliers are comes into data mining area outlier detection and analysis is a dada mining task hence it is referred as outlier mining. Discrete attributes Frequency of each value,Mode = value with highest frequency, Continuous attributes Range of values, i.e. min and max, Mean (average) Sensitive to outlier Median, Better indication of the ”middle” of a set of values in a skewed distribution. Data preprocessing useful and it is necessary to get overall picture of data. Descriptive data summarization techniques can be used to identify the properties of data and focus which data values should be treated as noise or outliers before working with data pre-processing. In many data pre-processing tasks users interested to learn the characteristics regarding both central tendency and dispersion of data. The central tendency (normal behavior of data) measures are mean, median, mode and midrange, when finding data dispersion include quartiles Inter quartile(IQR) and variance. These descriptive statistics are very much useful in understanding the distribution of the data. In data mining we need to examine how we can compute efficiently in large data bases.

## 2. Measuring the central Tendency

(i) **Mean**: The most common measure is “center” of a set of data is mean. Let x1,x2,x3,….,xN be a
Set of N values or observations for some attribute like marks of the student average marks of the student will be calculated by

$\overline{X}$ = (x1+x2+……+xN) / N

The mean is often used as a summary statistic. However, it is affected by extreme values (outliers): either an unusually high or low number. When we have extreme values (outliers) at one end of a data set, the mean is not a very good summary statistic.
If values are

(ii)**Median**: Median is the number that falls in the middle position once the data has
Been organized data means the numbers are arranged from small to large or from large to small. The median for an odd number of data values is the value that divides into two equal parts. If ‘n ‘ is odd n/2 and if n is even (n+1)/2 position is median. For asymmetric data, median is the measure suitable for center of the data.

(iii)**Mode:** The mode of a set of data is simply the value that appears most frequently in the set. If two or more values appear with the same frequency, each is a mode. The downside to using the mode as a measure of central tendency is that a set of data may have no mode, or it may have more than one mode. However, the same set of data will have only one mean and only one median. If a data set has only one value that occurs the set is called **unimodal**.A data set that has two values that occur **bimodal**. When a set of data has more than two values that occur with the same greatest frequency, the set is called **multimodal**. When determining the mode of a data set, calculations are not required, but keen observation is a must. The mode is a measure of central tendency that is simple to locate, but it is not used much in practical applications.

### 2.1 Measuring the Dispersion of Data

The degree to which numerical data tend to spread is called the dispersion. The most common measures of data dispersion are range, the five-number summary min, max, median or mean ,Q1,Q3
**Example: 49, 53, 62, 41, 67, 58, 68, 65**
The five-number summary consists of the numbers we need to draw the box-and-whisker plot: the minimum value, $Q_1$ (the bottom of the box), $Q_2$ (the median of the set), $Q_3$ (the top of the box), and the maximum value (which

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 6, June 2014*

is also $Q_4$). So we need to order the set, find the median and the sub-medians, and then list the required values in order.

Ordering the list: 41, 49,53,58,62,65,67,68 so the minimum is 41 and the maximum is 68

finding the median: (58 +62 ) ÷ 2 = 60 =$Q_2$

lower half of the list: 41,49,53 so $Q_1$ = 49

upper half of the list: 65,67 and 68 so $Q_3$ = 67

**five-number summary: 41, 49, 60, 67, 68**

Now we plot a box-and-whisker plot is to show how spread out your values are. But another of our values is way out of line , we need to consider "outliers".

### 2.3 OUTLIER CALCULATION USING INTERQUARTILE RANGE

1. Arrange data in order.
2. Calculate first quartile (Q1)
3. Calculate third quartile (Q3)
4. Calculate inter quartile range (IQR)=Q3-Q1
5. Calculate lower boundary= Q1-(1.5*IQR)
6. Calculate upper boundary= Q3+(1.5*IQR)
7. Anything outside the lower and upper boundary is an outlier.

Example: Consider
25,26,32,27,11.6,28.5,24.6,33.2,41.2,36.1,18.9,48
Q1=24.6, Q3=32.9, IQR=Q3-Q1=32.9-24.6=8.3
Lower limit= Q1-(1.5*IQR) =12.5, Upper limit= Q3+(1.5*IQR)=45.35
The values which are beyond these are extreme less than 12.5 is 11.6 outlier and greater than 45.35 is 48 outlier.

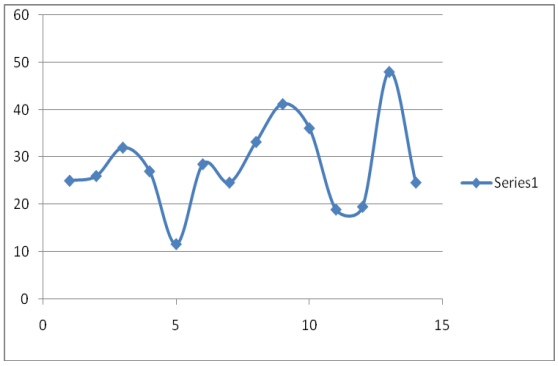


Fig 1. Line Graph

### 3. Experimental Results with weka

Outlier and extreme values in a data set performance of algorithms will effect, hence we need to identify and remove outliers from data set before applying the data mining techniques. In weka it is possible to identify outliers By IQR.

Following steps require to detect outliers

1. Open a file
2. Go to Filters
3. Select Unsupervised
4. Select attributes
5. Click on IQR

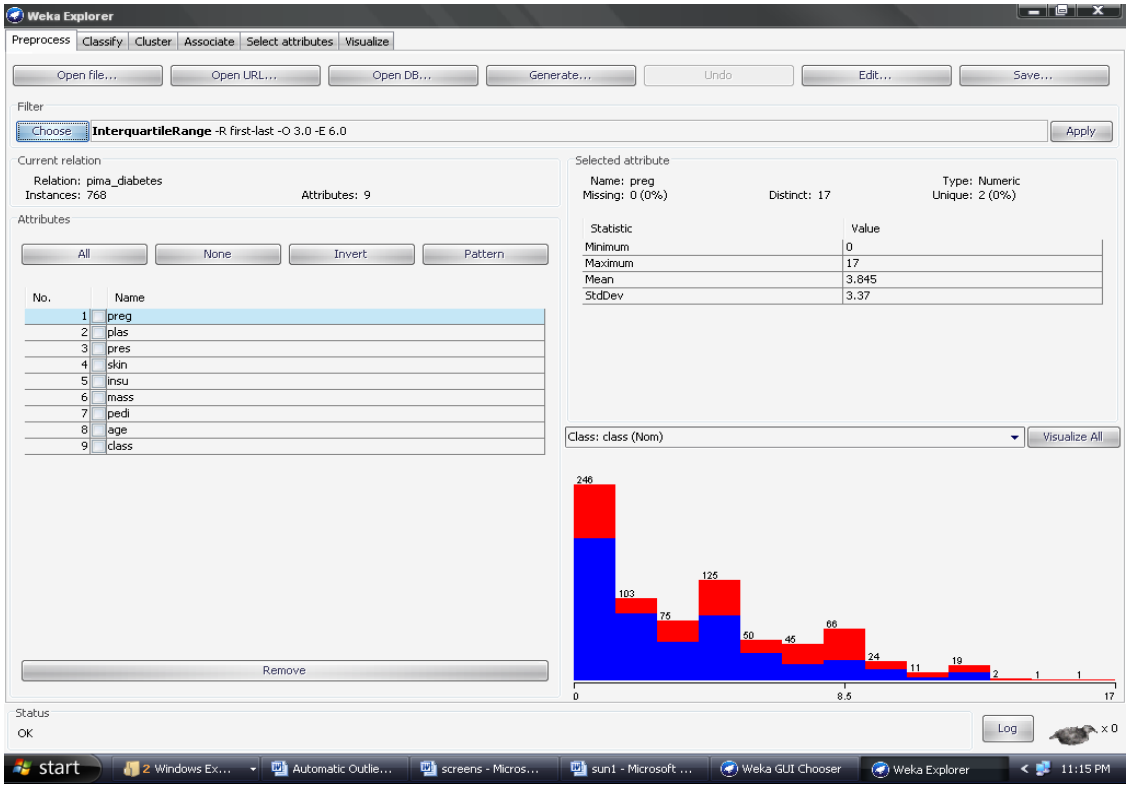


Fig 2.output screen of before applying IQR

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 6, June 2014*

3.1 Output Screen

On diabetes data set before applying filter IQR ,data set consists of 768 instances and 9 attributes. By observing the results automatically two attributes are added one is Outlier and other is Extreme value ,so now 11 attributes present.

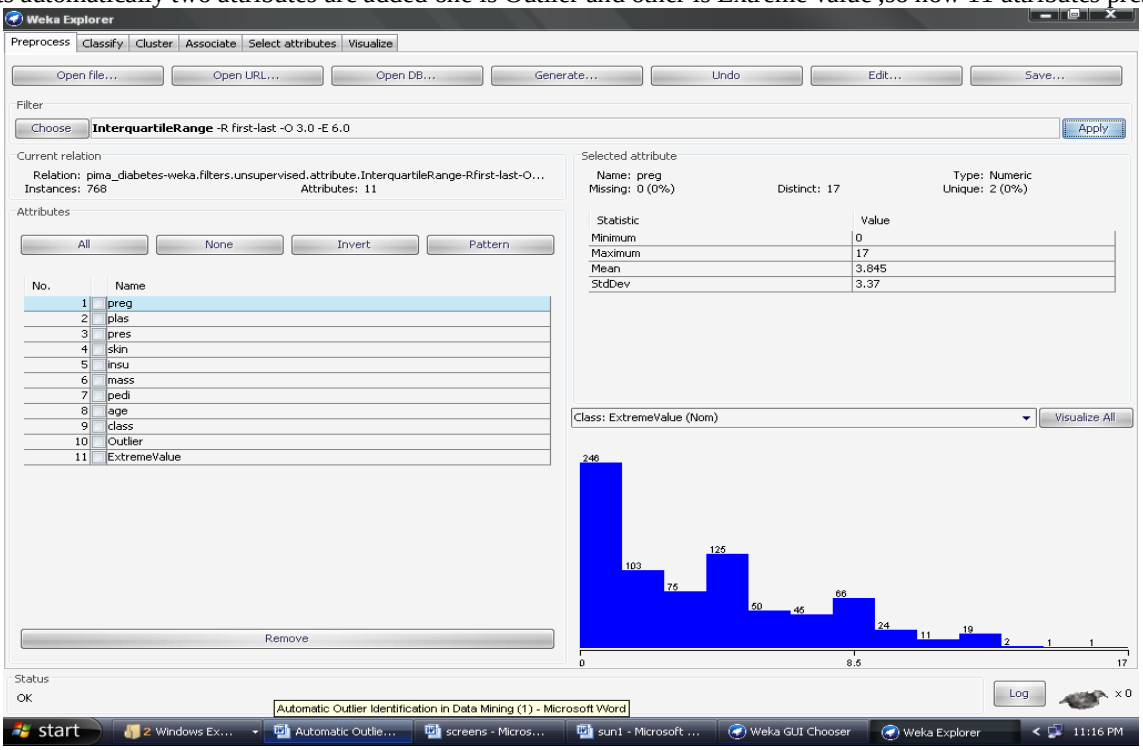


Fig 3.Output screen after applying IQR

| SNO | Name of Data set | Instances | Attributes Before IQR | Number of Outliers | Number of Extreme values |
|---|---|---|---|---|---|
| 1 | Diabetes | 768 | 9 | 49 | 0 |
| 2 | Supermarket | 4627 | 217 | 0 | 0 |
| 3 | Vote | 435 | 17 | 0 | 0 |
| 4 | Soybean | 683 | 36 | 0 | 0 |
| 5 | German Credit | 1000 | 21 | 25 | 155 |
| 6 | Iris | 150 | 5 | 0 | 0 |
| 7 | Labor | 57 | 17 | 1 | 0 |
| 8 | Breast cancer | 286 | 10 | 0 | 0 |
| 9 | CPU | 209 | 7 | 36 | 9 |
| 10 | Glass | 214 | 610 | 16 | 42 |

Table 1.Expertimental results on 10 sample data sets

## 4. CONCLUSION

Outlier detection is interesting and important task in data pre processing, Outliers will affect on results of data mining techniques. By using central tendency of the data set which will provide normal behavior of the data. In this current research work automatically detecting outlier by weka data mining tool. Filters option on unsupervised data and Inter Quartile Range (IQR) we apply a data set. after applying IQR two attributes are added ,outlier and extreme value .In this paper 10 experiments are performed ,in this some data sets does not have outliers means all instances are normal.

## REFERENCES

[1]. OUTLIER DETECTION Irad Ben-Gal Department of Industrial Engineering Tel-Aviv University Ramat-Aviv, Tel-Aviv 69978, Israel.
[2]. V. Barnett and T. Lewis, Outliers in Statistical Data (John Wiley & Sons, 2d ed., New York, NY, 1985).
[3]. Text book *Data Mining*: *Concepts and Techniques*. Second Edition.Jiawei Han and. Micheline Kamber. University of Illinois at Urbana-Champaign
[4]. http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf
[5].weka data mining tutorial youtube
[6]. http://www.cs.waikato.ac.nz/ml/weka/documentation.html
[7]. A Survey of Outlier Detection Methodologies. Victoria J. Hodge (vicky@cs.york.ac.uk)∗ and Jim Austin(austin@cs.york.ac.uk)
[8]. n.wikipedia.org/wiki/Data_mining
[9]. D. Hawkins, Identification of outliers. Chapman and Hall London,1980
[10]. ].Outlier Detection Techniques, Hans-Peter KriegelPeer Kröger, Arthur Zimek