Detection of Discordant Observations and Visualization of Data

Loshma Gunisetti
CSE Department
Sri Vasavi Engineering College
Pedatadepalli,Tadepalligudem
W.G.District,A.P.,India

ABSTRACT

Discordant Observations are special values or extraordinary cases in the available data which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism. They can be used to identify special or extraordinary or fraudulent cases in day to day transactions. Preprocessing can be used to identify the noise in the data and removal of such noise improves data quality. Discordant Observations are also called Anomalies or Outliers. Anomaly Detection can be used for Traffic Analysis, Credit Card Fraud Detection. We applied Anomaly Detection to Traffic data set for identifying the anomaly traffic stations on the highway. Detected stations represent abnormalities in the traffic sensors data. This information is used by us to identify the faulty traffic sensors located at the highway stations. Two dimensional visualization of the outliers has been provided which can be used for analyzing the data in an efficient manner. Traffic Management becomes easier when the abnormal traffic sensors identified at the corresponding stations are identified. The method used here can be easily applied to very large datasets.

General Terms

Database Applications, Data Mining.

Keywords

Discordant Observations, Anomaly, Outlier.

1. INTRODUCTION

Discordant observations can be identified by Anomaly Detection. The goal is to find objects that are different from most other objects[6]. Anomalous objects are known as outliers since on a scatter plot of the data, they lie far away from other data points. Anomaly Detection is also known as deviation detection or exception mining. Outliers have been informally defined as observations which appear to be inconsistent with the remainder of that set of data [10], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [3]. Outlier detection [4] is a data mining technique like classification, clustering, and association rules.

Outliers [1] are frequently treated as noise that needs to be removed from a dataset in order for a specific model or algorithm to succeed (e.g. points not belonging in clusters in a clustering algorithm). Alternatively, outlier detection techniques can lead to the discovery of important information in the data ("one person's noise is another person's signal") [8]. Outlier detection strategies can also be used for data cleaning before any traditional mining algorithm is applied to the data. Anomaly Detection is often a part of data preprocessing. Anomaly Detection techniques [6] could be Model Based, Proximity – Based and Density Based. Applications for which anomalies are of considerable interest are Credit Card Fraud Detection [2], Public Health [7], Intrusion Detection, Ecosystem Disturbances, Medicine. This work uses the Traffic Dataset and can be effectively used to handle large datasets to detect outliers and visualize them.

Identifying the specific distribution of a dataset is the first task of our work.[6] If the wrong model or distribution is chosen then an object can be erroneously identified as an outlier. Computation of the test parameters for detection of anomalies is required in the next stage. This parameters are then used to identify the outliers. Identified outliers are the visualized. Visualization can be used as a convenient way to discover outliers; however, inspecting thousands of pictures is still a challenging task. For example, if there are 18 highways in our system. Each highway runs in 2 directions. If one picture is generated per day/highway/direction. This means 2*18=36 pictures are generated per day. To inspect one month's data, the traffic manager need to look at about 1000 pictures, this is tedious and error-prone. The growing volume of datasets makes it difficult for humans to browse the entire datasets. Thus we need data mining algorithms to discover suspected outliers so that they can be highlighted for further investigations. Fig 1 illustrates the main modules in our system.

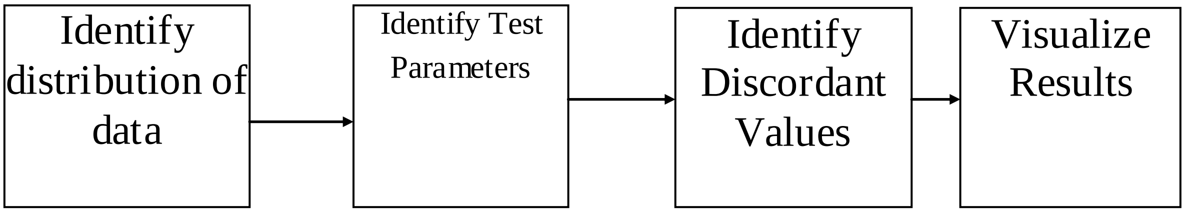


Fig 1: Main Modules of our system

2. DETECTION OF DISCORDANT OBSERVATIONS

Most studies in KDD (Knowledge Discovery in Databases) focus on finding the common patterns. However, finding the outliers (rare events or exceptional cases) may be more interesting and useful than finding the common patterns. Outlier detection is a data mining technique like classification, clustering, and association rules. A Spatial Outlier[5] is an object whose non-spatial attribute value is significantly different from the values of its spatial neighbors. A Temporal Outlier is an object whose non-spatial attribute value is significantly different from those of other objects in its temporal neighborhood. Comparison between spatially referenced objects are based on non spatial attributes.

The input data set used is the Traffic Dataset. The data set consists of 288 rows of the 5-min intervals, starting from 00:00AM; each row contains traffic volume for 150 stations. Data file format consists of number of stations, and timeslots and data, for example,

```
@relation 20101015
@stations 150
@timeslots 288
 1 2 3 4 5…
 2 4 2 6 2 …
```

Choice of Confidence Interval (68%, 95%, 99.7%) and Number of Outliers to be identified depends on the user choice (from 1 to 50).

In this application, the outlier would be the one station which detects a very high volume compare to the neighboring station. For instance, at 1:00 AM, station A detects a volume of 250, which the two neighbor stations B and C only collect single digits volume, then in this case station A would be considered as an local outlier.

The distribution of the input dataset is identified as Normal Distribution as in Fig 2 or Chi Square Distribution as in Fig 3.

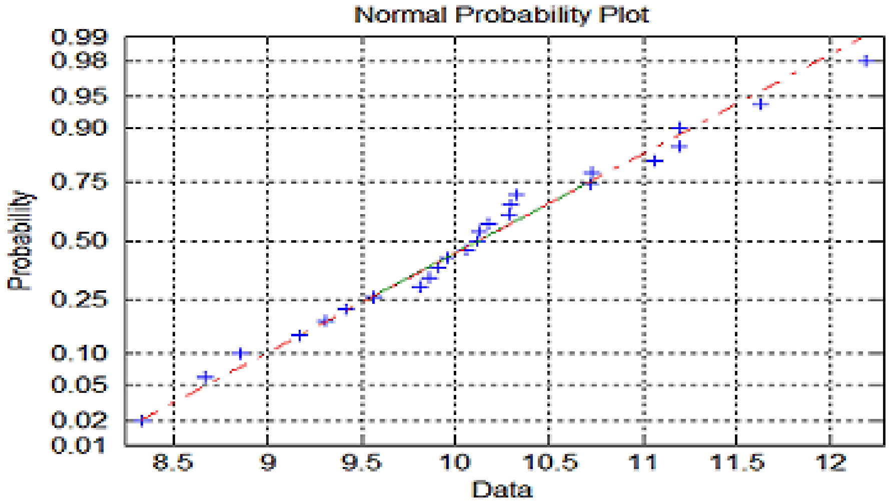


Fig 2: Normal Distribution Model of Data

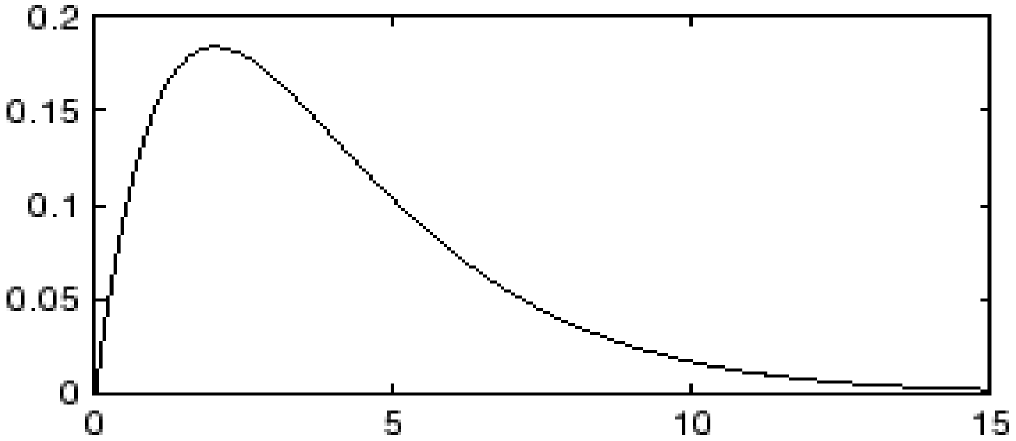


Fig 3: Chi Square Distribution Model of Data

3. CHOICE OF STATISTIC

First, the neighborhood can be selected based on a fixed cardinality or a fixed graph distance or a fixed Euclidean distance[6]. Second, the choice of neighborhood aggregate function can be mean, variance, or auto-correlation. Third, the choice for comparing a location with its neighbors can be either just a number or a vector of attribute values. Finally, the statistic for the base distribution can be selected from various choices such as normal distribution and chi-square distribution. The technique used here is the Spatial Statistic Approach [9]. Every location is compared to its neighborhood using the function

$$S(x) = [\, f(x) - E_{y\in N(x)}(f(y))], \qquad (1)$$

where
f(x) - attribute value for a location x
N(x) - set of neighbors of x
$E_{y\in N(x)}(f(y))$ - average attribute value for the neighbors of x
S(x)- difference of the attribute value of a sensor located at x and the average attribute of x's neighbors.

The Spatial Statistic [11] is used for detecting spatial outliers for f(x) is shown in equation (2).

$$Zs(x) = \left| \frac{s(x) - \mu s}{\sigma s} \right| > \theta \qquad (2)$$

μ_s - Mean value of S(x)
σ_s - Standard Deviation of S(x)
θ - Specified Confidence Level

The test for detecting an outlier can be described as $Zs(x) > \theta$. For each data object x with an attribute value f(x), the S(x) is the difference of the attribute value of data object x and the average attribute value of its neighbors; μ_s is the mean value of all S(x) and σ_s is the standard deviation of all S(x). The choice of θ depends on the specified confidence interval. For example, a confidence interval of 95 percent will lead to $\theta \approx 2$.Our objective is to determine stations that are "outliers" based on the volumes of the traffic measurements from each station.

4. ALGORITHMS

We now present algorithms to calculate the Parameters, e.g., mean and standard deviation for the statistics, as shown in Algorithm 4.1. The computed mean and standard deviation can then be used to detect the outlier in the incoming data set. Almost all data structures used in this application are matrix (i.e., 2D array).

Reasons that matrix data structure is chosen:

- Easy to access and modify specific cell
- Easy to convert from data to image.

The following algorithm is the pseudo code of the proposed system.

4.1 Parameters Computation Algorithm

This Algorithm is used to compute the Parameters.

Input: Dataset f(x)
Output: Parameters
Method:

```
1. Select an object in f(x)
2. Find its neighbors
3. Compute its Mean
4. Compute its Standard Deviation
5. Return Parameters
```

4.2 Discordant Observations Detection Algorithm

This Algorithm is used to detect the Outliers using the Parameters.

Input: Parameters
Output: Outlier station
Method:

1. Compute s(x) for f(x) using an object and its neighbors.
2. Compute Zs(x)= $\left|\frac{s(x)-\mu s}{\sigma s}\right|$
3. Compare Zs(x) with values in Normal Distribution for corresponding Confidence Interval(ɵ)
4. Return detected Outlier Station

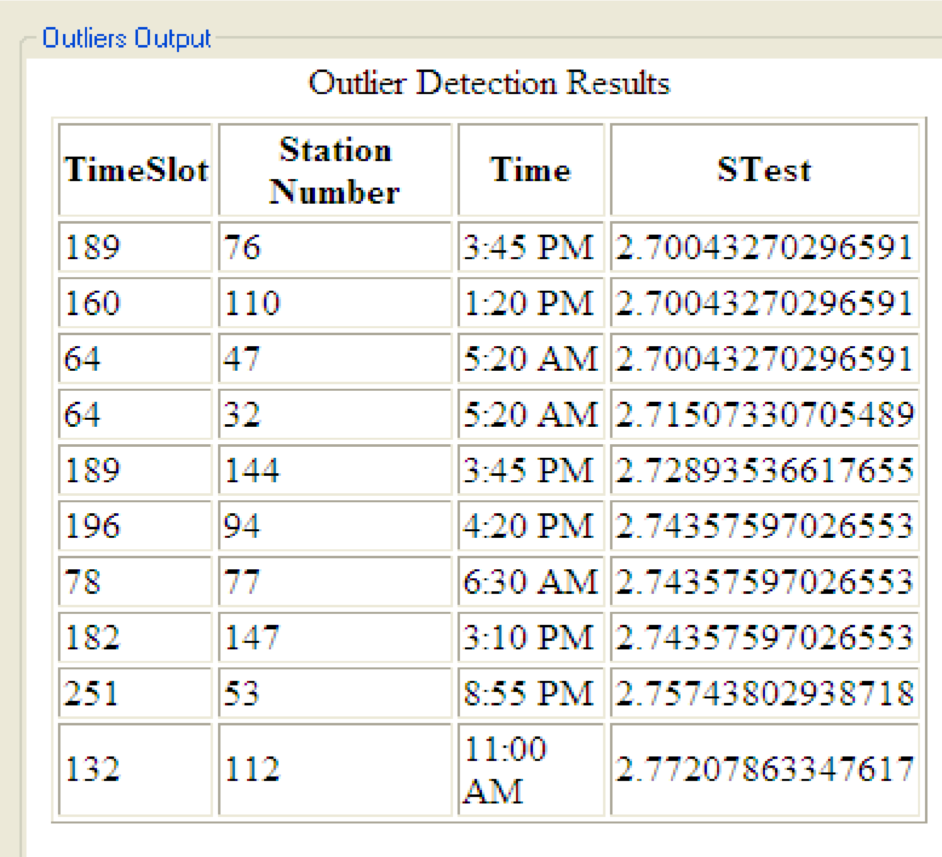
5. EXPERIMENTAL RESULTS

After computing the test parameters which are mean and standard deviation for the entire dataset, the outliers are detected and displayed in a tabular format. User can provide the desired confidence interval.

Detected outlier indicates the presence of a faulty sensor at the corresponding traffic station for the corresponding time duration. This information can be utilized by the traffic department to replace the faulty sensors or to identify the location of abnormality.

Table 1 illustrates the detection of 10 outliers found with respect to the stations and time and STest represents the S(x) value.

Table 1. Detected Outliers



Outlier Detection Results

TimeSlot	Station Number	Time	STest
189	76	3:45 PM	2.70043270296591
160	110	1:20 PM	2.70043270296591
64	47	5:20 AM	2.70043270296591
64	32	5:20 AM	2.71507330705489
189	144	3:45 PM	2.72893536617655
196	94	4:20 PM	2.74357597026553
78	77	6:30 AM	2.74357597026553
182	147	3:10 PM	2.74357597026553
251	53	8:55 PM	2.75743802938718
132	112	11:00 AM	2.77207863347617

Every outlier station and its corresponding timeslot and the total volume of the traffic at that station is indicated using a graph.

Fig 4 indicates the volume of traffic at timeslot 189

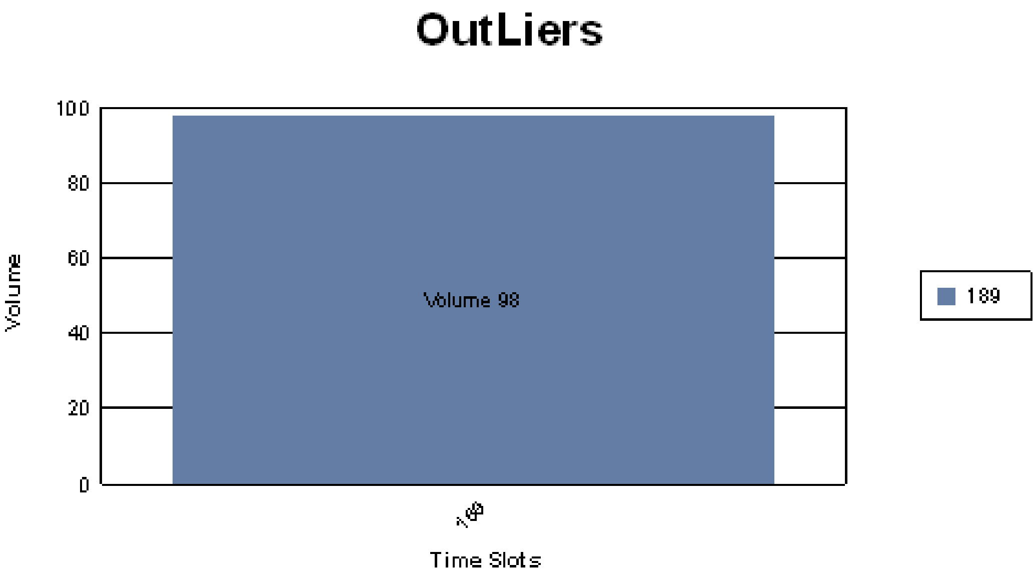


Fig 4: Volume of Traffic at Outlier Station

For one complete day the frequency of the traffic at one traffic station can be visualized which can be used to analyze the flow of traffic through that station. Figure 5 Visualizes the Traffic Volume at one station for one day

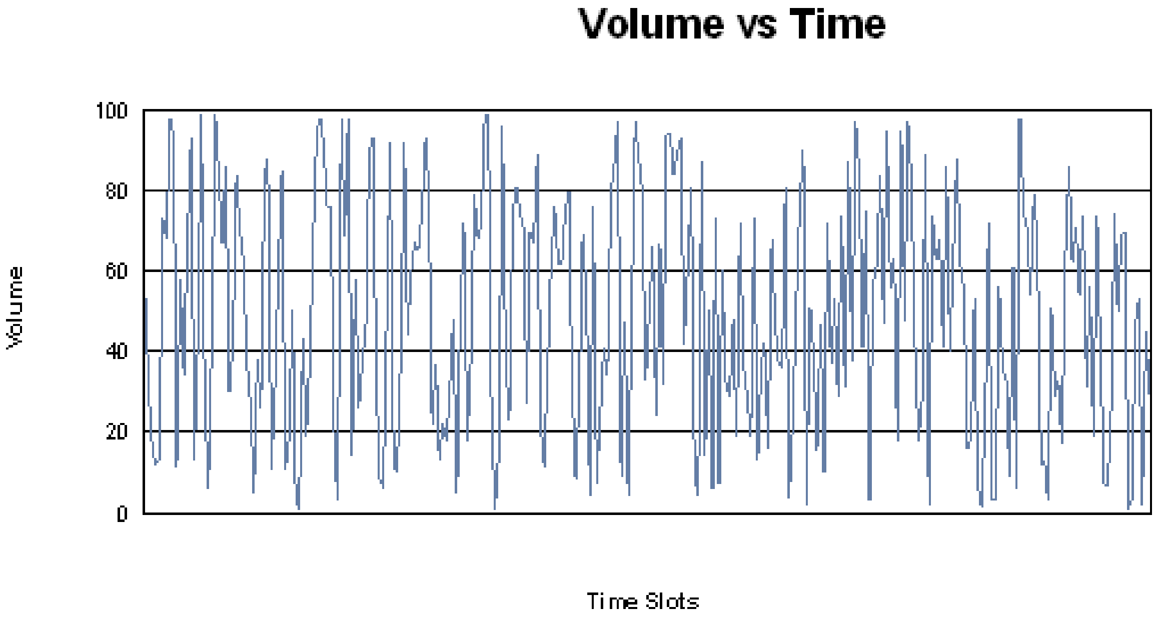


Fig 5: Traffic Volume of one day

Comparison between neighboring stations could be done by visualizing the traffic volume of a station and its neighbors. Fig 6 indicates the Traffic at station 76 in comparison with its neighboring stations 75 and 77 respectively.

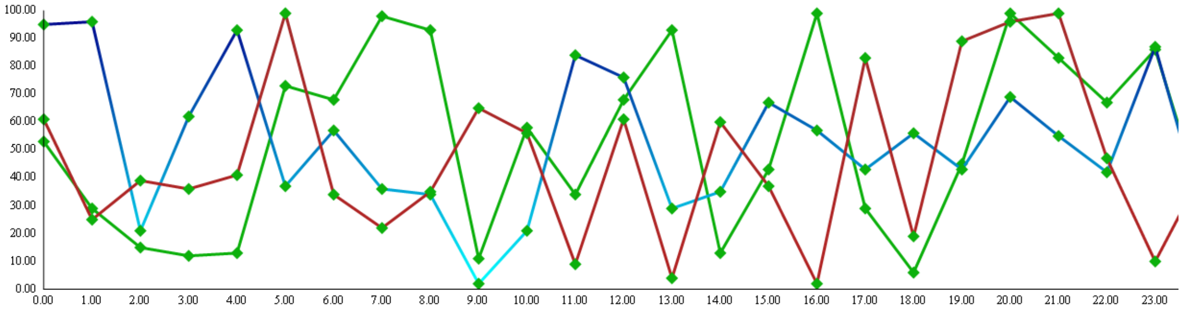


Fig 6: Neighboring Stations Volume Comparison

Visualization of the results provides an insight to the volume of traffic at different traffic stations and at different timeslots which can be utilized by the traffic department for effective planning and timely action.

6. CONCLUSION

This work was designed and tested for accuracy and quality. During this work, all the objectives have been accomplished and this work meets the needs of the Data Analyst. Users can choose various numbers of outliers from same dataset and also detect top k outliers by iterations to find one outlier. All results are shown through three different mechanisms: plain text, traffic volume image, and neighbor relationship graph between stations.

- Up to 50 Outliers can be found
- Confidence Interval can be selected by user
- Top k Outliers Query Processing can be done

File format and data type used in the algorithm to detect outliers are fixed in this project but various formats and types might be allowed

7. ACKNOWLEDGMENTS

The author would like to thank Dr. A V Bhaskara Rao, Principal, and Mr.Ch.V.V.Subbarao, Secretary and Correspondent of Sri Vasavi Engineering College, Pedatadepalli for their encouragement and support.

8. REFERENCES

[1] A. Koufakou, E.G. Ortiz, M. Georgiopoulos, G.C. Anagnostopoulos, K.M. Reynolds . A Scalable and Efficient Outlier Detection Strategy for Categorical Data, 19th IEEE International Conference on Tools with Artificial Intelligence

[2] Bolton, R.J., Hand, D.J. Statistical fraud detection: A Review, Statistical Science, 17, pp. 235–255, 2002.

[3] D.Hawkins. Identification of Outliers. Chapman and Hall, 1980

[4] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, Morgan Kaufman Publishers.

[5] Jingke Xi. Outlier Detection Algorithms in Data Mining, Second International Symposium on Intelligent Information Technology Application

[6] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2005

[7] Penny, K.I., Jolliffe, I.T. A comparison of multivariate outlier detection methods for clinical laboratory safety data, The Statistician, Journal of the Royal Statistical Society, 50, pp. 295–308, 2001

[8] Knorr, E., Ng, R., and Tucakov, V. Distance-based outliers: Algorithms and applications, VLDB Journal, 2000.

[9] S. Shekhar, C. T. Lu, and P. Zhang, A Unified Approach to Detecting Spatial Outliers, GeoInformatica, pp. 139-166. 2003

[10] V. Barnett and T. Lewis. Outliers in Statistical Data, John Wiley, New York, 3rd Edition, 1994.

[11] Shashi Shekhar, Sanjay Chawla. Spatial Databases A Tour, First Edition,2003