Selection of Best Outlier Detection Method Using Regression Analysis

Fahad Sultan
Faculty of information technology
University of Central Punjab
Lahore, Pakistan

Mudassir Ahmed
Statistical Department
University of Punjab
Lahore, Pakistan

***Abstract*-Outliers are unusual data values that are inconsistent with most of the records. Such non-representative records can seriously affect the model to be produced, so detecting outlier is a significant job to achieve higher accuracy. Several outlier detection methods are used in literature for real as well as simulated data sets. The aim of this study is to compare the two outlier detection method i.e. Cook's Distance and Mahalnobis method with standard method for outlier detection. The 15 replicates for simple linear regression of total household expenditure and household size are used to find the most efficient outlier detection method. It is found that the standard method of outlier detection produce less residual to predict the actual values as compared to the other two methods. While among the other two methods Cook's method produce less prediction error as compared to the Mahalanobis methods.**

Keywords-Outlier; Outlier Detection; Regression; Analysis; Data; Applications

I. Introduction

Now day's companies are facing challenges of handling ever increasing data. So it is becoming more difficult to extract the hidden and desired information from the large data. As the quantity of data grows very fast and somewhat consequence is that the amount of the useful information decreases very rapidly. Data mining is a process of extracting hidden and useful information from the data. The knowledge discovered by data mining is previously unknown, potentially useful, valid and of high quality. Data mining help out in finding the hidden relationship and pattern that reside in the data, if the data is faithful mirror of the real world. Selection of data and its preprocessing is the major work of data mining in absence of data warehouses. These two factors take almost 75% of the total effort in KDD process. Knowledge discovery in database is an iterative process of various stages. Companies have large amount of the data from which useful information should be extracted. Today's large numbers of shopping malls, banks, insurance companies and hospitals etc are using KDD to flourish their business, performance and to reduce cost. Data mining also helps organization for policy planning and decision making.

The job of segmenting the future customer is becoming more difficult. Data mining can help in such situation. The seller needs to get the prospective customer through data mining with the offer people are interested in and then target their campaign on these customers [1]. Most of the companies keep their customer's information in the database for the interaction purpose. Today's competition is on the top. A customer which you have today could disappear in future. There should be some proper medium to interact with your customer to keep them associated with you. In this contest it is not simple to interact with your customers efficiently. Before launching any promotional item company must identify the behavior of relevant customers, so that relevant offer must be offered to the right person and at the right time. The entire customers should not be targeted for the campaign because this can be expensive and less effective. Incentive is to be offered to company potential customers. This can leads to customer satisfaction and make stronger relationship between customer and company [2].

Techniques of data mining are the result of huge research work and products growth. It starts when people starting to store their data on computers. Later on capabilities to improve their data, technologies that give platform to access the data in real time made huge advancement in data mining. Data mining is widely used in applications by the business community. Technologies are sufficiently matured which are to be required. Strong multiprocessor systems are available which have large capacity for storing huge amount of data in very low cost. Many data mining algorithms are easily available. Such techniques produce much faster and efficient results [3].

II. Data vs Information

Data are the recorded facts. There are various sources, which are producing huge amount of data e.g. Business, nature, sports, science and medicine etc. These raw data is useless until some data mining techniques should be applied to extract information from it. Information is the hidden patterns which reside in the data. This can be explained by an example of shopping mall. If we say number of sale in a week or number of particular item sold in certain time period is called data. Whereas the relationship within the data provides us information e.g. what are the characteristics people have, who are purchasing specific item. Which product will customer buy with any particular item? E.g. If he purchases a tooth paste, there might be a chance he will purchase tooth brush as well.

III. DATA CLEANSING

Data has been collected from the various sources, the nature and format of data can be different. Data needs to cleaned and normalized after data gathering. Collation of various dataset can cause many problems because keeping of data in each dataset may be vary e.g. primary key issue, error, time periods etc. Duplicate records should be identified and eliminated from the dataset. Records with missing values should be discarded. There could be typographical errors in the nominal attribute, measurement error in numerical attributes and some deliberately made errors. Dataset can contain previously recorded information of a person, who might have changed his location and database still hold his/her old information. Most of the People purposely provide incorrect information about themselves e.g. A Person rejected by the insurance company will provide the incorrect name by slightly misspelling or by giving other false information on the second attempt. There is another example; some rigid kind of software required all fields to be filled up. A foreigner want to hire a car from car rental agency, which does not listed the postal code of other countries. Car rental representative have to enter their own office postal code in that field. If this happen on regular basis, the future data mining will produce result which shows that major customer of the rental company are living in neighborhood, which is totally bias. Let’s take another example, a petrol pump offer a scheme to its customers that after filling of 100 liter petrol, there will be a free car wash for one time. They provide a point’s card to their customer free of cost, which have to be scratched after filling of petrol. There are numerous customers which filled their car fuel tank but did not scratch the card, so at that time the salesman scratch his card. This way he can earn thousands of points. The future data mining will shows that salesman is the most valuable customer of the petrol pump, which is totally wrong information. Domain knowledge helps to identify such scenarios. You can build accurate model only if you have valid and accurate data, you cannot produce the hidden and useful information from the data, which is not faithful image of the real world. In data mining the answer of the respondent is very important, but in Pakistan people are annoyed to answer the question in proper way. This is the main hurdle in getting good and reliable information.

IV. DATA PRE-PROCESSING

Major part of data mining is the preprocessing of the data. Data sets need to be balanced; each class should receive equal proportion of data. There are two methods used for data balancing. First we can delete the record from the majority class to achieve the equal proportion. Second method is to duplicate the record by choosing randomly, this can lead to noisy model if many of the instances contain noise. Dataset may contain the abnormal values or unusual values which are inconsistent with most of the data are called outliers. Normally it is because of typing mistakes, measurement error or some time it is a real observation. Outliers can affect the accuracy of the model. There are many methods based on statistics to identify outliers. Outliers needs to be identified and should be removed. Outlier should be examined carefully, may contain important information. Three outlier detection methods have been compared in section 12, in term of their efficiency in detecting outlier by applying regression analysis.

Same feature could be store in different format in various datasets; all the data should be streamlined. In one database format of the currency is dollar and in other database the format is PKR. In such cases all the data should be transformed into a same currency format. There could be many instances with missing values, to cope up with such record we will predict a new value for these attribute or to identity the class which belongs to majority of the instances. The class which is most popular will be assigned to missing values instances. There are many techniques which produce quality data by analyzing the raw data. Data collection is the first step of data mining. The data should be collected from possible sources, after collection all the data should be integrated into one dataset. Data should have uniformity as the data are collected from various sources they might have different style of storing the same information. Data should be in clean format, duplicate record should be eliminated and null records should be discarded [4].

Data preparation is more time consuming as compare to data mining, this reveals the importance of data preparation. Initially the data is not mature, people have trend to giving the false data. There could be any typographical mistake in the nominal attribute and measurement error in numeric attributes. Data can have outliers; there could be noisy data. Incomplete data may consist of missing attribute values; there could be many attribute of interest which currently is not present in the data, such attribute should be added to the data. The data generated after the data preparation is more authentic and compact. This data reveal many quality patterns. This shows that data preparation is not a simple task. Thus data preparation is more challenging task in the data mining. 75% of the work is used on data preparation and attribute selection [5].

In some cases we add/remove some attributes to make the model simpler, efficient or due to our needs. There are many scenarios in which we required certain variable which are not already present in the database e.g. if we know the person income and their expenses, we can add new attribute saving by subtracting his expense from income. But this thing requires it expertise and this could make database more clumsy and difficult. Some time we remove attribute to make the model simpler with same level of accuracy. There are various techniques used for reducing data dimensionality. Sampling, feature subset selection and discretization are the techniques used for this purpose. Feature sub selection is most commonly and widely used technique to reduce the data dimensionality [6].

V. Data Visualization

Visualization plays an important role in data mining. The purpose of visualization is very simple. Data mining involve with the extracting hidden and useful information from the database. The understanding for a normal user can be complex. Visualization techniques should be used to present the data in such a way that a normal person can also understand the result of data mining from which they are previously unaware. The output should be precise and in some graphical format which gives high level of understanding of the problem [7].

The result of data mining or the data on which we are mining can be very large. So it is very necessary to present the data in some graphical form. Role of graphics is very significant in data visualizing. Large number of data can be shown in the form of a graph. It plays a significant job in identifying pattern, or understanding the structure of the data. Many research shows that graphics are not suited for large and exploratory work. Outlier is the main reason, which can make some uncertainty in the graphical form of data. Nominal attributes can be shown using histogram, for numeric attributes we can use graph [8].

There are many methods for data visualization. Most of them only work well with discrete variables. Graphical tools for continuous variables are considered to be less effective. So the continuous variable should be discretised. One of the methods is known as mosaic plots, which may be used to explore the association rule using linking. Association rule are good approach to use when you have many variables in large data sets. The result produced by this method is too large, thus required filtering the records. Association rules are only works with the subset variable categories, so all the other categories are neglected. Mosaic plots are capable of examining all the background rules, so that power of other rules is to be judged [9].

VI. Data Modeling

The word model in data mining can have several meaning. Mostly it is referred to as data modeling process. Suppose if we are applying neural network on a particular problem, in such scenario the model refers to general type of model. Another use of model is referred with the output of the modeling process. In many other cases the model is referred to as mapping between the input and output associated to it [10].

Modeling is the way of building model on one situation where you know the output and then to apply on another situation where you are unaware of the output. Modeling is a technique of data mining to get the unknown information from the dataset. To find a high sale ratio of a superstore, we must identified some attributes e.g. age group of customer, location of superstore, competitor around that store, peak sales time, least sales time, item sold most, item sold less, items sold together etc. According to these information a model will be generated, which includes characteristics that are common to that superstore. If you are able to make a good model you can find the reason of high sale of particular item in that superstore [11].

VII. Decision Trees

Decision tree is the predictive model that makes the prediction based on the series of decisions. The first branch is called the root of the tree and rest intermediate nodes are the branching nodes with each leads to some leaf node. The tree is built by selecting one attribute at a time that best separate the classes. The set of examples are portioned according to the attribute value. This is repeated at each branch node until the segmentation is completed. For example if you want to see buying trend of male and female, you can build a model, which segmented the whole tree by gender attribute [12]. Decision tree are widely used for data exploration, preprocessing and prediction. Data is segmented according to the series of decision based on attributes. It gives the high level understanding of the data. Decision tree is also used for preprocessing of the data. It is used to find the subset of useful predicators which can be later used for the data mining. Initially the decision tree was used only for data exploration and data preprocessing. But later on it is widely used for the prediction e.g. if you want to predict whether a customer will buy a product and given person attributes gender, country and age. Decision tree will help you out in exploring and making prediction of such information. Decision tree needs to be simple with higher accuracy rate. Complex tree have a higher computational cost and higher complexity [13]. Neural network and decision tree are commonly used in data mining. Neural network have many significant advantages. One of them it is very accurate predictive models that can be applied on large number of distinct attributes. There is also a term artificial neural network. Artificial neural network are based on computer programmed based. True neural network can make prediction, find out pattern and can learn [14].

VIII. Data Mining Tasks

Data mining is the search of valuable from large stream of data. DM is the extraction of hidden knowledge and useful information from the databases. Data mining generates a model from the features in the database drawn from machine learning and other fields. This model represents the patterns which exist in the dataset. There are many model function used today in data mining. Which are classifications, regression, clustering, summarization, dependency, link analysis and sequence modeling. Classification maps the data into predicting target classes. Target class is known and it is also called supervised learning. Regression maps the data on numeric attributes e.g. you can predict the weight of a person by its age and height. Clustering maps the data into similar groups according to similar characteristics and also called unsupervised learning. Summarization provides the compact view of the large data. Dependency shows the association between different attributes e.g. age attribute is dependent on data of birth attribute. Links analysis shows the relation between attributes, it is very much similar to dependency. Sequence analysis is related with time factor. Detailed explanations of these tasks are given below [15].

A. *Classification*

Classification is also called supervised learning. At a time it can have one target class. In classification target class is known. Outcome is called the class of the example, we will measure success on fresh data for which class labels are known. We will arrange the data into pre defined groups e.g. Instead of writing number of each student in a class, we will make a groups of range like if number fall in 50% to 60% that student will get A. Classification tree focus on learning in nominal attributes, if we want to apply it on continuous data, it should be discretized first.

B. *Clustering*

Clustering is also called unsupervised learning. Target class is not known in it. It is very much same like classification but it will not have predefined groups. We will try to make groups that are very much similar according to their characteristics. Success will be measured subjectively.

Clustering is the way to divide the similar records in to groups according to their characteristics. Clustering gave us high level understanding of the database. Clustering can also be defined as segmentation or partition. There are two clustering system by Claritas corporation and Micro vision. These companies have segmented the human population broken down by age, income, education etc. They believed it will be very helpful for marketing and sales promotions. This clustering information is then used by the data mining experts. They can analyze which clustering information is related to their business and which is not. They can find out the customer behavior so that they can target to potential buyer of the system for their promotional campaign [16].

Nearest neighbor is the prediction technique which is very similar to clustering. If you want to predict attribute values of one instance you must find out the similar instances based on their common attributes. You can pick the value from instances which are nearest to the missing attribute instance. For example if you want to predict the monthly income of a person which is missing in an instance, you must find out the instances which have common attribute value like house location, educational history, age, current employer, previous job history, total experience. After having these information checks in the data which instances are closer to these information's. Then you can make a prediction of person income by viewing closest neighbor income. This technique works in the same way as people normally thinks by examining the closely matching examples [17].

C. *Regression*

Regression analysis is a method that is used to find the relation between one or more independent variables with dependent variable, in a view to estimate or predict the average value of the dependent variable. Regression analysis has been performed on 15 dataset to check the efficiency of outlier detection method in sec 12.

D. *Dependency Modeling*

Many attributes have a relational dependency on each other. Attributes are associated to each other by changing one attribute can effect the other attribute. Association techniques find groups of items that are sold in a single transaction, called market basket analysis. Association provides valuable information in term of marketing point.

There can be three type of association.

- Positive Association (i.e. Paste and brush)
- Negative Association (i.e. Diesel car and petrol)
- Zero Association

E. *Time Series and Event Stream Analysis*

These are associated with certain time periods. If we talk about the valentine day, the sale of flowers dramatically increases on that day as compare to normal days. In spring season people often visit more to parks and garden as compare to other seasons. Usage of water in summer seasons is comparatively high.

Some part of the dataset in data warehouse is related to time base events. Taking all the data as whole is qualitatively different than taking each data superlatively e.g. taking data of all the year is qualitatively different than each month of a year. There can be different trend of data in each month. Data mining expert should understand this fact to create better model. Background knowledge is required for such task. If data miner doesn't have background knowledge the model produce by the data set will be bias [18].

IX. Ethics in Data Mining

Ethical issue can arise in data mining. Some of the ethical issue depends on the nature of application e.g. religion and race information in loan application is unethical. Whereas these attributes are acceptable in medical applications. The purpose of data collection should be cleared before the time of collection. Suppose a data is collected for issuing the credit cards, later on this data used for any other purpose is the violation of ethics. The primary purpose should be identified before collection of data. It should be clear that who is permitted to access the data and for what purpose data will be used and what kind of conclusion will be drawn from the data [19].

X. Importance of Data Mining

The main objective of data mining is getting hidden information from the collected information, and then explaining the hidden information in precise way so that policies can be made in business according to data mining results, when competition is on top. The importance of data mining is very high in today’s challenging environment. Many companies are using data mining for increasing their revenue and sales of their products and hence to beat their competitor in the market. If we apply data mining in our business we can save lot of money e.g. sending a company new products newsletter to all the customer of a shopping mall can be in-affective and costly. By apply data mining we can find potential customers who can respond to company new products. Newsletter should be send to potential group of people extracted through data mining. Data mining is highly important in medical diagnosis and treatment. By analyzing the previous history of a patient we can forecast many things which can be beneficial for the patient. Data mining is emerging area of research. The dimension of data mining is not only limited to business only, but it covers other fields of life. Data mining are very commonly used in health industry, marketing, ecommerce etc. Use of mathematical and statistical tools has great importance for data mining because variable of interest cannot be segregated without use of mathematical algorithms and statistical tools.

Database marketing software has a great impact on business growth. As the data mining is a here today, developer need to combine the data mining technology with currently exist software’s so that better result can be obtained. For this developer need to understand the business problem and required to show the result in understandable format. After getting the knowledge contained in a database, people can turn what should be done [20].

XI. Applications

Data mining is widely used in large scale of organizations. The dimension of data mining is not only limited to business only, but it also covers other fields of life. Some of the areas of real world where data mining applied are the following.

A. Marketing/Sales

Large number of data is produced by the shopping malls. Such raw data can be used to find the hidden pattern underlying that data. Identify the potential distinct customers in customer database and find their buying behavior and characteristics. Email or SMS marketing can be done on such target people to increase the sale and revenue.

B. Fraud-Detection

Data mining helps us in fraud detection. While withdrawing amount from the ATM machine previously history is to be checked. Suppose a person ATM card has been lost or theft. The thief will try to get withdrawal the entire amount from the ATM machine. While at the time of withdrawal, the history of already withdrawal amount of that person will be checked. If the current transaction matches to previous one then payment will be released. Otherwise ATM will hold the card and will not issue the amount.

C. Risk of Loan Payment

Data mining is very helpful in processing loan queries. Companies have made a questioner in which number of questions has been asked regarding person monthly income, age, already defaulter, saving, employment history, qualification, property and dependents etc. 90% of the cases can be processed by using statistical methods. Rest of the cases is called borderline cases. These cases are referred to loan officer. After evaluating the questionnaire, expert judgment is to be made to analyze the risk. If it has minimal risk then loan is to be granted to that person otherwise his/her loan application will be rejected. Boarder lines cases are the most active customers. Normally 50% of the boarder line cases defaulted.

D. Medical Diagnosis and Treatment

The use of data mining is very common in medical applications. Data mining helps us in finding the diseases by critically analyzing the symptoms. Through data mining we can forecast the future problems. Many doctor using data mining negatively. They apply data mining on patient before an operation. They want to check whether patient is going to recover from disease in future or not. If there is high risk involve that he will not be survive then doctors withdrawal from that case, because if he dies this affect the doctor repute.

E. Judiciary

Data mining can play a vital role in judiciary. Large amount of data can be obtained, which contains the previous history of the cases. Data mining helps in judiciary by analyzing the historical data on judgment of similar cases.

F. Car Insurance

Data mining can help car insurance companies in detecting whether a particular customer can made any accident in near future. They have made a questionnaire which asked different questions related to their personality. Data mining has been applied on customer information if the outcome shows the risk of accident then company can charge more car insurance fee from the customer as compare to other who has very less risk of accident.

G. Real Estate

Data mining can also be beneficial in real estate. Price of a specific house can be calculated by analyzing the different houses in their surrounding. For this number of attributes are required e.g. area of the house, stories, location, number of rooms, availability of resources, house material etc. By analyzing these data, data miner can make an expert judgment about the cost of the house.

H. Biomedical and DNA data Analysis

Data mining was widely used in biomedical and DNA data analysis. But now it is banned to apply data mining on DNA data analysis. In America females who are expecting babies made DNA test of their babies for applying data mining on it. If their babies DNA slightly matches to world famous and noble people, they keep the babies alive otherwise they made abortion.

I. Telecommunication

Telecommunication companies are using data mining before the launch of their newly offers to check the consequences of it. If company wants to launch any package data mining will be applied first, whether this package/offer is going to click in their customers or not. Company will analyze the customer information like their age, their current package, their peak time of usage of phone, monthly cards consumption etc for data mining.

XII. Data Analysis

The 15 replicates of total household expenditure and household size of 30 instances are used in the analysis. Relationship between two features is displayed using Scatter plot in "Fig. 1".

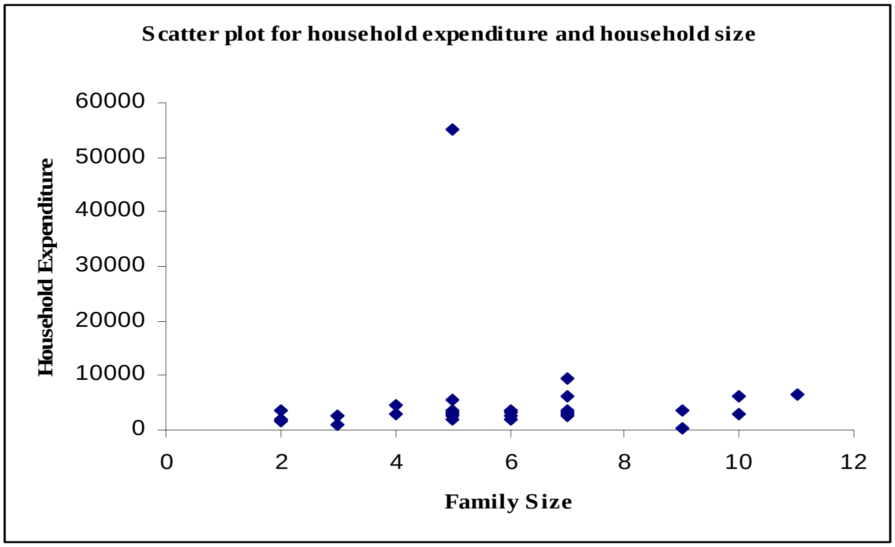

Figure 1. Date presentation using Scatter plot

Scatter plot is an informal method of data presentation. Scatter plot shows the relationship between the variables and trend of the data, it also helps us to find the minimum, maximum values and outliers present in the data. In figure1 all the data values are showing the same tendency except 1. By initial observation we can consider it as outlier. But to confirm it we have to apply some outlier detection methods.

After analyzing the data, regression analysis has been applied on all the dataset without eliminating the outliers. The residual value of each dataset has been calculated which is displayed in the table 1.

TABLE 1. RESIDUAL WITH OUTLIERS

Residuals of 15 Data Set	
Dataset	***Residuals***
1	9.427935995
2	10.80948704
3	10.82356637
4	8.765587803
5	10.7962479
6	10.5551665
7	10.53678968
8	10.46824725
9	10.13042292
10	9.883151843
11	10.11691448
12	10.25431436
13	10.13242244
14	10.02357961
15	9.944007971
Total Residual	162.8456983

After that three method of outlier detection has been applied on all dataset. These methods are:

- Standard Method (Mean ± 2 S.D)
- Cook's Method
- Mahalanobis Distance Method

These methods are based on statistics. Standard method is to find the mean and standard deviation of attributes value and use these parameters to find the threshold value as shown in the "Fig. 2".

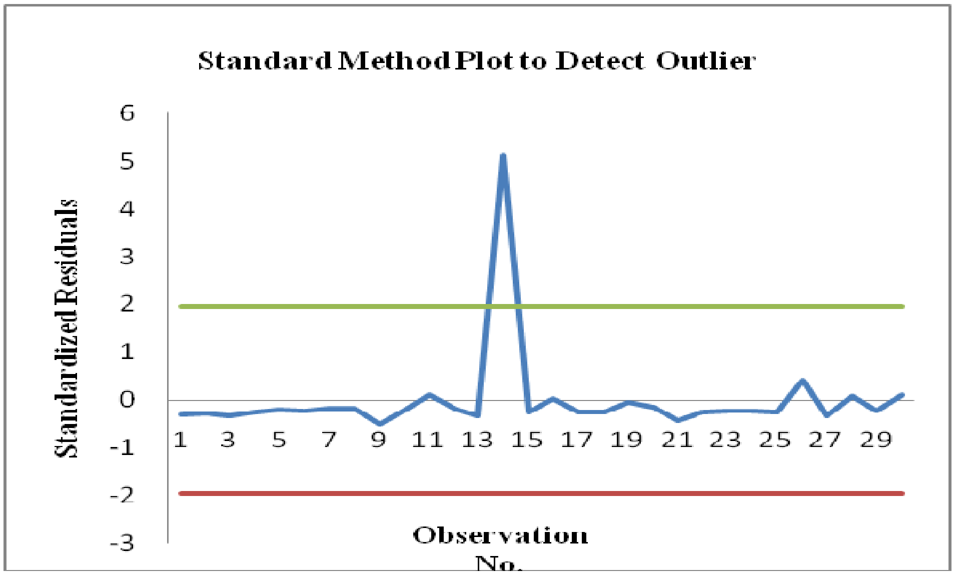

Figure 2. Outlier Detection using Standard method

Any observation outside this range is considered as outlier.

TABLE 2. Residual Comparison

Comparison of Different Outlier Detection Methods			
Dataset	***Standard Method***	***Cook's Method***	***Mahalanobis***
1	7.902977	7.902977	9.427677
2	9.426933	9.426933	9.426607
3	8.001877	8.001877	10.82333
4	7.404063	8.619537	8.749243
5	10.30009	10.30009	10.79492
6	8.790136	8.790136	10.55515
7	8.025477	8.025477	10.52786
8	7.557991	7.557991	9.450656
9	8.786246	8.786246	10.13017
10	8.910762	8.910762	8.910762
11	8.960141	8.960141	10.11417
12	8.829424	10.01985	10.01985
13	9.370301	9.370301	10.1208
14	9.115246	9.115246	10.02048
15	8.256322	8.256322	9.935637
Total Residual	129.638	132.0439	149.0073

Cooks distance is used for detecting whether any observation can affect the regression estimates. Distance is calculated from the mean of dependent and independent variable. Observations which have cook distance of 1 or more are considered to be as outlier. Mahalanobis distance method is based on correlation between variables, through which dissimilar patterns can be identified and analyzed.

Standard method and Cook's distance method shows almost same tendency in predicting outlier. In most of the dataset same observation are pointed out by these two methods. Whereas Mahalanobis result are totally different from these two methods.

Regression analysis has been applied on each dataset after eliminating the outliers. Residual value of each dataset is calculated. Residual values of each datasets are presented in the table 2.

Table 2 shows the residual values of each dataset. Most of the residual values are identical in Standard method and Cook's method because same outliers were pointed out by both methods. In last row sum of all the residual is mentioned. Standard method has the less residual values and hence the most effective method in outlier detection compared to Cook's and Mahalanobis distance method. There is an immense difference in residual values while applying regression analysis without deleting outlier as shown in table 1 and after deleting outlier's observations shown in table 2 from each dataset.

XIII. Conclusion

Finally, 15 different data sets have been applied on Standard method for outlier detection, Cook's method and Mahalanobis distance method by using the regression analysis. The decision drawn after analyzing the dataset that standard method has the less overall residual value and hence the most effective method for outlier detection. It is also noted that Standard method and Cook's method detect the same outlier observation in most of the datasets, whereas Mahalanobis has totally different observation in outlier detection.

Acknowledgment

We would like to thank the Dr. Abdul Aziz for their careful reading and constructive comments, which helped to improve the research skills. With out his guidance and sincere effort we are not able to write this paper.

REFERENCES

[1] T. M. Chan. Low-dimensional linear programming with violations. In Proc. 43rd Annual. IEEE, Symposia. Found. Computer. Sci, 2002.

[2] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer- Verlag, Berlin, Germany, 2001

[3] T. Cheng and Z. Li. A hybrid approach to detect spatial temporal outliers. In Proc. of the 12th International Conference on Geo informatics, pp 173–178, 2004.

[4] Shekhars and Duen-Ren, CCAM: a connectivity- clustered access method for networks and network computations by Liu Knowledge and Data Engineering, IEEE Transactions on Volume 9, Issue 1, Jan/Feb 1997

[5] C.-T. Lu, D. Chen and Y.Kou. Detecting spatial Outliers with multiple attributes. In Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003), Sacramento, California, USA, pp 122–128. IEEE Computer society, 2003

[6] G. S. Brodal and R. Jacob. Dynamic planar convex hull. In Proc. 43rd Annu. IEEE Sympos. Found. Comput. Sci.

[7] L. Breiman, Bagging Predictors, Machine Learning, vol. 24, 2, pp. 123-140, August 1996.

[8] V.Barnett and T.Lewis, Outliers in Statistical Data. New York, NY, John Wiley and Sons, 1994

[9] Kurt Thearling, Some Thoughts on the Current State of Data Mining Software Applications, Published in the January 13, 1998 edition of DS.

[10] Kurt Thearling, Data Mining and Privacy: A conflict in the making, Published in the March 17, 1998 edition of DS.

[11] M. Joshi, R. Agarwal and V. Kumar, Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?, In Proceedings of the Eight ACM Conference ACM Sig kdd International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, July 2002

[12] Sariel Har-Peled and Vladlen Koltuny, Separability with Outliers, September 12, 2005

[13] E. Arkin, F. Hurtado, J. Mitchell, C. Seara, and S. Skiena. Some lower bounds on geometric separability problems. In 11th Fall Workshop on Computational Geometry, 2001

[14] Pei Sun and Sanjay Chawla, On Local Spatial Outliers, School of Information Technologies, University of Sydney, June 2004

[15] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, New Algorithms for Fast Discovery of Association Rules, Proc. Of the 3rd Int'l Conf. On Knowledge Discovery and data Mining (KDD-97), AAAI Press, 1997.

[16] N. Serbedzija, The Web Supercomputing Environment, Seventh International World Wide Web Conference, Brisbone, Australia, April 1998.

[17] U. Fayyed, G. Shapiro, P. Smith, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

[18] M. Chen, J. Han, and P. Yu, Data Mining: An Overview from a Database Prospective, IEEE Trans. Knowledge and Data Engineering, 8, 1996.

[19] Unwin, A. R., Hawkins, G., Hofmann, H., and Siegl, B. Interactive Graphics for Data Sets with Missing Values - MANET. Journal of Computational and Graphical Statistics, 5(2), pp 113-122, 1996

[20] Langley, P., and Simon, H.A. Applications of machine learning and rule induction. Commun. ACM 38, 11, pp 55–64, Nov. 1995

[21] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, I. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Cambridge, Mass., 1996.

[22] Garrett R.G. The chi-square plot: A tool for multivariate outlier recognition. Journal of Geochemical Exploration. Vol. 32, pp. 319-341, 1989

[23] Gervini D. A robust and efficient adaptive re-weighted estimator of multivariate location and scatter. Journal of Multivariate Analysis. Vol. 84, pp. 116-144, 2003

[24] Rousseeuw P.J., Van Zomeren B.C. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association. Vol. 85(411), pp. 633-651, 1990

[25] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA, pages pp 93–104. ACM, 2000.

[26] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers: algorithms and applications. In Proc. of the 7th International Conference on KDD, 2001.

[27] M. Joshi, R. Agarwal and V. Kumar, Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?, In Proceedings of the Eight ACM Conference ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, July 2002.

[28] N. Billor, A. Hadi and P. Velleman BACON: Blocked Adaptive Computationally-Efficient Outlier Nominators, Computational Statistic & Data Analysis, vol. 34, pp. 279-298, 2000.

[29] P. Sun, S. Chawla, On Local Spatial Outliers, In Proceedings of Fourth IEEE International Conference on Data Mining (ICDM'04), Brighton, United Kingdom, November 2004.

[30] E. M. Knorr and R. T. Ng. Algorithms for mining distance based outliers in large datasets. In Proceedings of 24th International Conference on Very Large Data Bases, New York City, New York, USA, pp 392–403. Morgan Kaufmann, August 24-27, 1998.