www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 2 Issue 6 June , 2013 Page No. 2045-2049*

# IDENTIFICATION OF OUTLIERS BY COOK'S DISTANCE IN AGRICULTURE DATASETS

***T.Jagadeeswari*** [1] **,** ***N.Harini*** [2]
C.Satya Kumar [3]M.Tech(CSE) MISTE.
Department of Computer Science and Engineering[1,2,3]
Vardhaman College of Engineering, JNTU Hyderabad

## Abstract

*Data mining play a vital role in computer field . A huge and valuable knowledge is extracted from the large collection of data. outlier detection is currently an important and active research problem in many fields and is involved in numerous applications. This paper applies minimum volume ellipsoid (MVE) with principle component analysis (PCA) extension, a powerful algorithm for detecting multivariate outliers. If the data points exceed the cut-off value, the cook's distance is used for the outliers. The paper also compares the performance of the suggested frame work with statistical methods to demonstrate its validity through simulation and experimental applications for incident detection in the field of agriculture.*

*Keywords: outlier detection, PCA, MVE, cook's distance.*

## 1. Introduction

Outliers are data objects in a data base that do not comply with the general behaviour or model of the data. Most data mining methods discard outliers or noise or exceptions, in some applications such as fraud detection the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining. Outliers may be detected using statistical tests that assume distribution or probability model for the data, or distance measure where objects that are at a substantial distance from any other cluster are considered outliers.

Outlier is defined as "an observation whose value is not in the pattern of values produced by the rest of the data". A more comprehensive definition is due to
Beckman and Cook (1983).
- "An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism " (Hawkins, 1980).
- "An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of the dataset" (Barnet & Lewis, 1994).
- "An outlier is an observation that lies outside the overall pattern of a distribution" (Moore and McCabe, 1999).

Many data mining algorithms try to minimise the influence of outlier in data sets. They are extensively used in a wide variety of applications such as fraud detection in credit card transactions, intrusion detection in cyber security, identifying novel molecular structures in the field of bioinformatics as part of pharmaceutical research, loan application processing of problematic customers, etc. Their importance in data is due to the fact that they can translate into actionable information in a wide variety of applications.

### 1.1. Defining Outliers

Outliers are considered as noise lying outside a set of defined cluster. The outliers lie outside the cluster but are separated from the noise.(incorrect data). Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

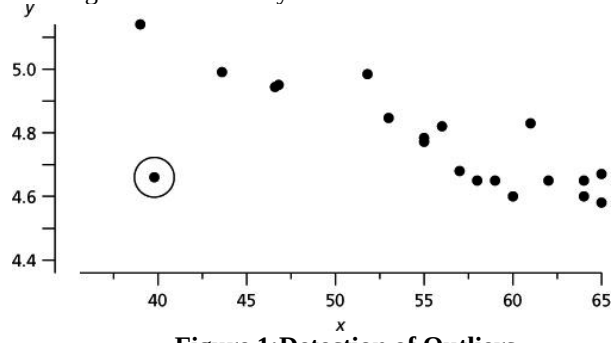


**Figure 1:Detection of Outliers**

Identification of potential outliers is important for the following reason- An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

## 2. Problem identification

The outlier detection methods are derived from three fields of computing: statistics (proximity-based, parametric, non-parametric and semi-parametric), neural networks (supervised and unsupervised) and machine learning.

Statistical approaches were the earliest algorithms used for outlier detection. Some of the earliest are applicable only for single dimensional data sets. One such single dimensional method is Grubbs' method. In this method the first step is to quantify how far the outlier is from the others. Calculate the ratio Z as the difference between the outlier and the mean divided by the SD. If Z is large, the value is far from the others. Note that you calculate the mean and SD from all values, including the outlier.

**Mean:**

Suppose we have sample space $\{x_1,x_2...x_n\}$. Then the arithmetic mean *A* is defined as the equation

$$A := \frac{1}{n}\sum_{i=1}^{n} x_i$$

**The Standard Deviation (SD)**

The standard deviation of a data set is a measure of how spread out the data is. "the average distance from the mean of the data set to a point". The way to calculate it is to compute the squares of the distance from each data point to the mean of the set, add them all up,divided by n-1 , and take the positive square root.

Standard deviation is a statistical measure of spread or variability. The standard deviation is the root mean square (RMS) deviation of the values from their arithmetic mean.

**Formula for Standard Deviation**

$$S=\sqrt{\frac{\sum(X-M)^2}{n-1}}$$

where Σ = Sum of

X = Individual score

M = Mean of all scores

N = Sample size (Number of scores)

**Variance:**

The square of the standard deviation. A measure of the degree of spread among a set of values; a measure of the tendency of individual values to vary from the mean value

**Principal Component Analysis**

Principal Component Analysis (PCA), developed by Karl Pearson in 1901, is a simple, non parametric method of extracting relevant information from confusing data. The aim of this method is to reduce the dimensionality of multivariate data and is a linear transformation that transforms the data to a new coordinate system. This method calculates the covariance matrix, because covariance is always measured between two dimensions. It measures how much the dimensions vary from the mean with respect to one another. If we calculate the covariance between one dimension and itself, we will get the variance of that dimension. The covariance matrix describes all relationships between pairs of measurements in the considered data set.

The basic formula for the covariance is

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{(n-1)}$$

Where X and Y are two separate dimensions of data. By getting the covariance matrix, the Eigen vector and Eigen values are calculated. Before outlier detection the observations on the eigenvectors are scored with positive Eigen values for co-variance matrix.

For a square matrix *A* of order n, the number **λ** is an Eigen value if and only if there exists a non-zero vector *C* such that

$$AC=\lambda C$$

Using the matrix multiplication properties, we obtain

$$(A - \lambda I_n)C=0$$

Where $I_n$ is the unit vector

Covariance measures the degree to which two variables change or vary together (i.e. co-vary). On the one hand, the covariance of two variables is positive if they vary together in the same direction relative to their expected values (i.e. if one variable moves above its expected value, then the other variable also moves above its expected value). On the other hand, if one variable tends to be above its expected value when the other is below its expected value, then the covariance between the two variables is negative. If there is no linear dependency between the two variables, then the covariance is 0.

Correlation is a measure of the relation between two or more variables. The correlation coefficient ρX, Y between two random variables X and Y values

$$\rho xy = \frac{cov(X,Y)}{\sigma_x\,\sigma_y}$$

To counter this masking problem, Rousseeuw (1985) introduced the robust Minimum Volume Ellipsoid (MVE) method for detection of outliers in multidimensional data. By the term change point, we mean a time point at which the data properties suddenly change.

John (1978) studied the problems that arise in detecting the presence of outliers in the results from factorial experiment. He actually applied the technique of Gentleman and wilk(1975) and John and Draper(1978), who investigated The problem of detecting outliers in two-way table and provided a statistic $Q_k$ which is difference between

the sum of square of residuals from original data and sum of squares revised residuals resulting from fitting the basic model after deleting K-influential observations .The cook statistics also detect the outlying observations in experimental data.

Cook's distance measures the effect of deleting a given observation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a large Cook's distance are considered to merit closer examination in the analysis.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ \mathrm{MSE}}.$$

The following is an algebraically equivalent expression

$$D_i = \frac{e_i^2}{p\ \mathrm{MSE}} \left[ \frac{h_{ii}}{(1-h_{ii})^2} \right].$$

In the above equations:

$\hat{Y}_j$ is the prediction from the full regression model for observation *j*.

$\hat{Y}_{j(i)}$ is the prediction for observation *j* from a refitted regression model in which observation *i* has been omitted.

$h_{ii}$ is the i-th diagonal element of the hat matrix $\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$

$e_i$ is the crude residual (i.e., the difference between the observed value and the value fitted by the proposed model).

MSE is the mean square error of the regression model;

P is the number of fitted parameters in the model.

## 3. Implementation and design methodology

The suggested PCA-MVE method is for finding Outliers in datasets. The sample data is generated and taken in form of numeric values then PCA method is applied. In PCA, the Covariance and Correlation matrix for the data is found. Then by applying MVE algorithm, whether it is an Outlier or a Cluster is known. By giving the weight as 0 for Outlier and 1 for Cluster, the Outliers are found and then projected in the form of a Graph.

### 3.1. Procedure

Input: Take aricultural samples such as boro rice, mustard, brinjal, tomato, potato
Step 1: generate or create numeric data set.
Step 2: Apply PCA for sample dataset.
Step 3: Calculate covariance matrix and then find Eigen vector, Eigen values and correlation matrix for the data.
Step 4: Then apply MVE to know whether it is an outlier or cluster by giving the weights as 0 for outlier and 1 for cluster.
Step 5: Apply cooks distance, if data point exceeds the cut-off value.
Step 6: Generate reports to visualize the outliers in the data set by using statistical charts.

## 4. Experimental results

The experimental results taken from cropping system research such Boro rice,Mustard, Brinjal, Tomoto, potato generate above procedure
The sample input data is as follows:

| Treatment | Kharif | Rabi | Summer |
|---|---|---|---|
| T1 | Rice | | |
| T2 | Rice | Boro Rice | - |
| T3 | Rice | Mustard | Rice |
| T4 | Rice | Brinjal | Rice |
| T5 | Rice | Tomoto | Rice |
| T6 | Rice | French Bean | Rice |
| T7 | Rice | Potato | Rice |

**Table 1**. The sample data

The

| No of observations | cooks distance |
|---|---|
| 1 | 0.0026807 |
| 2 | 0.0636333 |
| 3 | 0.1049018 |
| 4 | 0.0003942 |
| 5 | 0.0255928 |
| 6 | 0.0029281 |
| 7 | 0.0149541 |
| 8 | 0.1131487 |
| 9 | **0.6901545** |
| 10 | 0.01862 |
| 11 | 0.1295856 |
| 12 | 0.1117699 |
| 13 | 0.0303065 |

| | |
|---|---|
| 14 | 0.0693472 |
| 15 | 0.0032514 |
| 16 | 0.0010388 |
| 17 | 0.0375675 |
| 18 | 0.0008375 |
| 19 | 0.0003222 |
| 20 | 0.0334282 |
| 21 | 0.0016305 |
| 22 | 0.0116759 |
| 23 | 0.0095747 |
| 24 | 0.0009105 |
| 25 | 0.2274856 |
| 26 | 0.0018246 |
| 27 | 0.0013758 |
| 28 | 0.1577261 |

Table 2:The sample data by cook's statistics

### 4.1. Statistical charts for representing the outliers

The statistical graphs are useful for representing the data in a meaningful way. A good graph conveys information quickly and easily to the user. Graphs highlight the hidden features of the data.

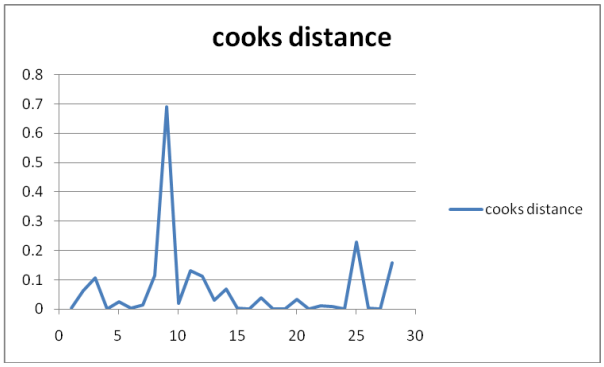


Figure 2: line chart

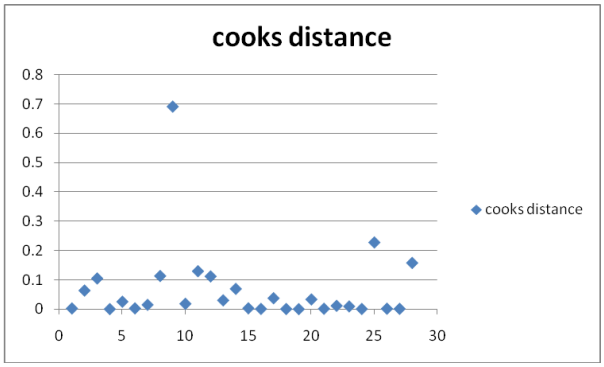


Figure 3: Scatter chart

## 4. Conclusion

The important feature of the statistics developed for identifying outlier in linear regression models .This paper suggests the use of PCA-MVE, which is one of the data mining techniques, along with **Cook's distance** to detect the outliers in two dimensional or multivariate data sets. Such detection of the outliers helps in identifying errors in agriculture data set.

## 5. Bibilography


[1] Jun-ichi Takeuchi and Kenji Yamanishi (2006) Unifying Framework for Detecting Outliers and Change Points from Time Series

[2] Hodge, V.J. and Austin, J. (2004) A survey of outlier detection methodologies.

[3] Romy Shioda and Levent Tuncel (2005) Clustering via Minimum Volume Ellipsoid

[4] Hongxia Pang, Jiaowei Tang, Su-Shing Chen, and Shiheng TaLac Statistical distributions of optimal global alignment scores of random protein sequences

[5] V. Barnett and T. Lewis 1994 "Outliers in Statistical Data,"JohnWiley&Sons

[6] Wold, S., Esbensen, K., Geladi, P. (1987).Principal Components Analysis. Chemometrics and Intelligent Laboratory Systems. 2, 37-55.

[7] T.Cover and J.A. Thomas, Elements of Information Theory. Wiley-International, 1991

[8] M. Huskova, "Nonparametric Procedures for Detecting a Change in Simple Linear Regression Models," Applied Change Point Problems in Statistics, 1993.

[9] K. Yamanishi and J. Takeuchi 2001 "Discovering Outlier Filtering Rules from Unlabeled Data,"Proc. Fourth Workshop Knowledge Discovery and Data Mining, pp. 389- 394.

[10] K. Yamanishi and J. Takeuchi, "A Unifying Approach to Detecting Outliers and Change- Points from Nonstationary Data," Proc of the EighthACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.

[11] T. Fawcett and F. Provost, "Activity Monitoring: Noticing Interesting Changes in Behavior," Proc. ACM-SIGKDD Int'l conf. Knowledge Discovery and Data Mining, pp. 53-62, 1999.

[12] V. Guralnik and J. Srivastava, "Event Detection from Time Series Data," Proc. ACM-SIGKDD Int'l Conf. Knowledge Discovery and DataMining, pp. 33-42, 1999

[13] D.M. Hawkins, "Point Estimation of Parameters of Piecewise Regression Models," J Royal Statistical Soc. Series C, vol. 25, no. 1, pp. 51-57, 1976.

[14] Kenji Yamanishi and Jun-ichi Takeuchi, "A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data" in IEEE transactions on knowledge and data engineering vol17. No.6, June 2006

[15] Barnett, V. and Lewis, T.: 1994, Outliers in statistical Data. John Wiley & Sons.,3 edition.

[16] Aggarwal, C. C. and Yu, P. S.: 2001,Outlier detection for High Dimensional Data'.In: Proceedings of the ACM SIGMOD Conference 2001

[17] V. Guralnik and J. Srivastava, "Event Detection from Time Series Data," Proc. ACM-SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 33-42, 1999