International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

Noise Reduction In Data Using Polynomial Regression

Penta Venkata Sai Dinesh Kumar[1], Kapileswarapu Girish Kumar[2], Nallamada Gyanadeep[3]

[1] [2] [3] School of Computing Science and Engineering, VIT University

Abstract:-Noise is common in data which hinders the data analysis. We consider noise as low-level data errors or objects that are irrelevant to data analysis. Data cleaning technique reduces the low-level data errors but not irrelevant objects. To reduce both types of noise there are three traditional outlier detection techniques distance-based, clustering-based, and an approach based on the Local Outlier Factor (LOF) of an object. In this paper we introduce a new method for noise reduction using polynomial regression and spearman's rank correlation coefficient ρ. This approach allows a high recognition of noise with low false rate.

1. Introduction

Database may contain data objects that do not adhere with the general behavior or model of the data. Those data objects can be considered as noise or outliers. Analysis of noise or outlier data is called as outlier mining.

In this paper we explain four noise removal techniques. In which three of them are based on outlier analysis techniques:

1) Distance-based outlier detection
2) Density-based local outlier detection
3) Deviation-based outlier detection

The fourth technique which is a new method that we are proposing is PRCLEANER which is based on creating polynomial regression function for the noiseless data set and using the obtained models equation for testing the new data set where they adhere with the trained data set or not.

1.1 Distance-based outlier detection

Data objects which does not have enough neighbors are considered as distance-based outliers, where neighbors are defined based on distance from the given object. An object O, in a data set *D*, is a distance-based (DB) outlier with parameters pct and dmin, that is a DB(pct, dmin)-outlier, if at least a fraction pct, of the objects in *D* lie at a distance greater than dmin from O.

There are several algorithms for mining distance-based outlier, they are

1) Indexed-based algorithm
2) Nested-loop algorithm
3) Cell-based algorithm

1.2 Density-based local outlier detection

This outlier detection method is designed to identify outlier in data object based on varying density. Based on local density of an object neighborhood, local outlier factor is determined for an object, where an object's neighborhood is defined by the minpts nearest neighbors of the object. minpts is a parameter that specifies the minimum number of objects (points) in a

neighborhood. Each data object is assigned a local outlier factor(LOF) and objects which are closer to dense cluster will have high LOF The data objects with high local outlier factor are considered as outliers.

1.3 Deviation-based outlier detection

In this method it identifies outliers by observing the main characteristics of objects in a set. Objects that deviate from these characteristics are considered as outliers. For example simulate a process familiar to humans, after seeing a series of similar data, the data object disturbing the series is considered an exception

There are two methods in this technique
1) Sequential exception technique
2) OLAP data cube technique

2. PRCLEANER: Polynomial regression cleaner

In this section we propose a PRCLEANER method. The idea behind this method is to generate n models for n dimensional data using polynomial regression. In each model, one dimension will be taken as response variable and other n-1 dimensions as predictor variables. Let us consider 3 dimensional data set (x, y, z), so we produce 3 models using polynomial regression. For model x as a polynomial function of y and z is expressed as,

$x = k_0+k_1y^{n1} +k_2z^{n2}$

Similarly for model y and model z can be expressed as

$y = k_3+k_4x^{n3} +k_5z^{n4}$

$z = k_6+k_7y^{n5} +k_8x^{n6}$

Now let us take a data set which has no noise in it and by applying the polynomial regression for each dimension in the data set, we obtain polynomial regression for each model. After the equations are obtained we take a data set to test. Using the equations we get values for x and y. Now we apply spearman's rank correlation coefficient ρ for the obtained results. If $\rho \in (-1,1)$ then the data is not noise, if the ρ value is not in that range then we consider it as noise.

3. Numerical Evaluation

Let us take data set as follows

X	Y
5	8
6	9
7	10
8	11
9	12
10	13

The equations obtained from the above data set are

Y=0.990 * x ^ 1.000+ 3.931

X=0.989 * y ^ 1.000- 3.249

Now we take another data set to test using above equations and if the results are not approximately similar to the obtained results for all the models then we consider the data to be noise or outlier. For example we take the test data set to be as follows,

X	Y
112	125
167	171

For test case 1 (112,125)

Taking x(112), then y=114.811

Taking y(125), then y=120.376

For test case 2(167,170)

Taking x(167) ,Y=169.261

Taking y(170) ,X=166.87

For the obtained results we apply spearman's rank correlation coefficient ρ,

$\rho = 1-(6\sum d_i^2/n^3-n)$

if the obtained result lies between (-1,1) then the data belongs to that set and if not in that range then it is considered as noise or outlier.

By applying for the above results we get,

For test case1 (112,125)

$\rho = 1- ((6*((114.811-125)^2+(120.376-112)^2))/(2*(4-1)))$

$\rho = -172.973$

For test case2 (167,170)

$\rho = 1- (((6*(167-166.87)^2+(170-169.261)^2))/(2*(4-1)))$

$\rho =0.436979$

so from the obtained results test case1 is considered as noise or outlier and test case2 belong to the data set.

Conclusion

The goal of work presented in this paper is to improve the quality of data analysis techniques to remove very high level of noise. Three outlier detection techniques were described in this work. We proposed a new technique PRCLEANER. The above experimental results show high detection of noise for given data set with low false rate.

References

[1] Jaiwei Han and Micheline Kamber "Data Mining: Concepts and Techniques".

[2] Hui Xiong, Michael Steinbach "Enhancing Data Analysis with Noise Removal". IEEE transactions on knowledge and data engineering, vol. 18, no. 3, march 2006.

[3]http://en.wikipedia.org/wiki/Rank_correlation.