ISSN:2320-0790

An International Journal of Advanced Computer Technology

Outlier free Real Estate Predictive Model

Geetali Banerji[1], Kanak Saxena[2]

[1]Professor & Head – Computer Science, Institute of Innovation in Technology & Management, New Delhi
[2]Professor & Head – Department of Computer Applications, Samrat Ashok Technological Institute, Vidisha, M.P., India

Abstract: Studying human behaviour is a difficult task for many reasons:-The reasons are ranging from an intentional task to a cognitive sign, which are processes other than those of interest occurring at the same time. The processes can be psychological, social constraint and all cognitive. The processes operated in the background and have either some time no effect on the collective data or they may reflect the results occasionally. All those undesired behaviours produce measurable responses sometimes happens to be correct by chance. Some responses however may attract attention due to their unusual aspects, denoted as outliers. If not properly handled the outliers in the design phase the resultants may affect the resultant inferences or the experimental outcome at initial stage. Thus, it is required to treat the outliers before it is too late in the design phase itself. The influence of outliers is more importance if the sample size is small with the examined statistics, which is less robust. In this paper, we have detected the outlier patterns in the Real Estate era and removed it up to a great degree, which not only reflects the drastic change in the results but also improves the rules formation. Thus, we provide a structure and comprehensive overview of the research on outlier detection.

Keywords: AODE (Average one dependence estimators), Classification based outlier detection, clustering based outlier detection, statistical based outlier detection component

I. INTRODUCTION

In Real Estate domain, a real estate broker who sends a direct mail catalog to its current customer database promoting various schemes. Customer Interest, recorded in a transactional database recoding his demographic, educational and professional data. The need or requirements of customers keep on changing that is of more dynamic in nature. By grouping its existing customers by income range for instance, into groups of "willing to invest", "may be interested" and "not interested" may not give us the accurate image of the investing behavior of a customer. This may be due to the changing environment of a customer as only the income can't responsible for the decision, where as the other factors also play key role in the decision, i.e., the factors that are treated to be outliers.

An outlier is an observation that deviates so much from other observations as to arouse suspicious. The problem to detect outlier is to finding patterns in data that do not conform to expected behavior. The following algorithm detects the outlier in the data set.

Step 1: Analyze the data set.

Step 2: Compare the data as per the defined model rules.

Step 3: If it matches as per the norms then keep it as data set 1 else data set 2. The data set 2 contains outliers.

Thus, outliers are patterns in data that do not conform to a well-defined notion of normal behavior. Fig 1 illustrates outliers in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., point o1 and o2, and points in region o3, are outliers.

In order to treat the outliers, the major problem occurs that no unanimously accepted theoretical framework exists in the literature. In spite of various approaches formulated, their impact varies as the change in scenario and the number of fields used with weights increase in dimensionality results in difficulty in identification of outlier.

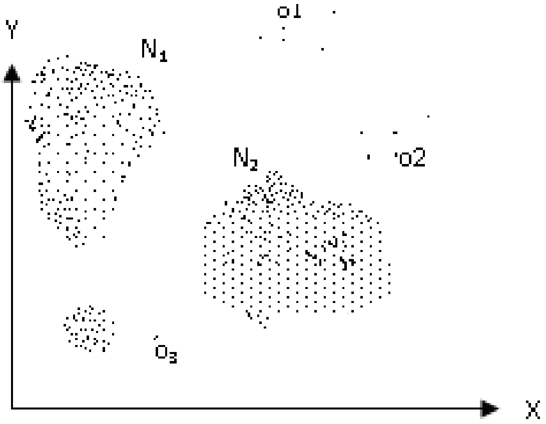


Fig 1 A simple example of outliers in a 2-dimensional data set

In this paper, we have detected the outlier patterns in the Real Estate era and removed it up to a great degree, which not only reflects the change in the results but also improves the rules formation. The organization of this paper is as follows: section II illustrate the concepts used in the paper, section III elaborate the various tests and lastly conclusion.

II. CONCEPTS USED

A. *Outlier Detection Techniques*

The approaches used in this paper are: Statistical based outlier detection, Classification based outlier detection and Clustering based outlier detection, which comes under the categories of statistical, unsupervised and supervised respectively as in the fig 2.

Statistical outlier detection techniques: A traditional approach to solve the outlier detection problem, based on the construction of a probabilistic data model and the use of mathematical methods of applied statistics and probability theory. A probabilistic model can be either a priori given or automatically constructed by given data. Having constructed the probabilistic model, one sets the problem of determining whether a particular object of the data belongs to the probabilistic model or generated in accordance with some other distribution law. The object which does not suit the probabilistic model, considered as an outlier.

1) *Tests:* The probabilistic model is not sufficient for detecting outliers. A procedure that determines whether a particular object is an outlier is required referred as a test. A standard test consists in the verification of the basic hypothesis (null hypothesis) [2]. In the Smart- Sifter algorithm, the objects processed successively, and the model uses supervised learning while processing each data object in order to detect the outlier by the introduction of a special measure is metric.

2) Regression Analysis: There are two approaches in the regression methods for the outlier analysis reverse search and direct search. In this paper, we have used both the methods by considering the whole data for the model preparation and considering a part data for the model preparation and then perform the various regression techniques for the outlier detection as they are mathematically justified.

Classification based outlier detection: Classification [3], [4] learns a model (classifier) from a set of labeled data instances (training) and then, classify a test instance into one of the classes using the learnt model (testing). Classification based outlier detection techniques operate in two-phase fashion. The training phase learns a classifier using the available labeled training data. The testing phase classifies a test instance as normal or outliers using the classifier. Based on the labels available for training phase, classification based outlier detection techniques grouped into two broad categories: multi-class and one-class outlier detection techniques. In this paper, initially one class outlier followed by the multi class outlier detection technique (multilayer perceptron) is used.

Semi-Supervised outlier detection: Techniques that operate in a semi-supervised mode, assume that the training data has labeled instances for only the normal class. Since they do not require labels for the outlier class, they are more widely applicable than supervised techniques. Such techniques are not commonly used, primarily because it is difficult to obtain a training data set which covers every possible behavior that can occur in the data. Kmeans clustering can be used in semi supervised mode also.

B. *WEKA*

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. It is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools. WEKA is open source software issued under the GNU General Public License [6]-[7]. The main interface in WEKA is the Explorer. It has a set of panels, to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer performs further analysis. Fig 3 is a snapshot of Real Estate data set in WEKA.

C. *Data Sets*

We have used two variations of Real estate datasets for testing purpose. The first one is, Real Estate complete data set (complete data set) and the, second one is Real Estate selected data set (selected data set).

Real Estate Complete data set: Real estate complete data set contains the social status, income and educational background of the customer. There are total 43 attributes and 5821 records.

Real Estate Selected data set: We have removed the demographic details from the real estate data set to determine their impact. In context to Indian scenario these factors plays a significant role. In this data set demographic detail such as caste marital status, number of children, religions, nationality and type of customer are removed. The resultant is in the data with 26 attributes and 5821 records.

III. MODELING & TESTING

Average one dependence estimators (AODE), applied for detecting the outliers. The outliers, removed from both the variations of the real estate data sets and the empirical results are compared.

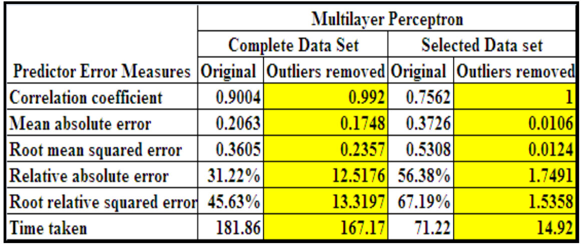
AODE is a probabilistic classification learning technique. It achieves the highly accurate classification by averaging overall of a small space of alternative naïve bayes like models and hence less detrimental, independence assumptions than naïve bayes. The resulting algorithm is computationally efficient while delivering highly accurate classification on many learning tasks.
Multilayer Perceptron, used for testing, Kmeans and Pace Regression and association rules, generated using Apriori.

A. *Multilayer Perceptron*

We have applied Multilayer perceptron on both the variations of data before and after removal of outliers. Table I depicts the results. The result of Multilayer perceptron on complete data set and selected data sets with outliers in [8].

Table I show the result of the classifier detection technique on original data, after preprocessed (complete data set and selected data set) and then we have applied the outlier detection technique and removed the outlier data.

Table I: Multilayer Perceptron Results on complete and selected data set. An assessment of original with outlier removed data



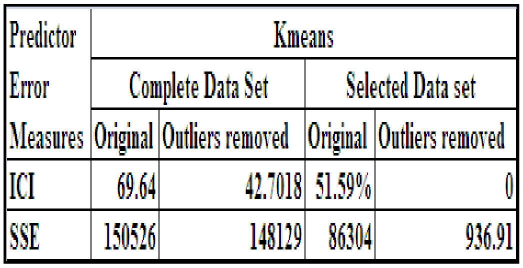
	Multilayer Perceptron			
	Complete Data Set		Selected Data set	
Predictor Error Measures	Original	Outliers removed	Original	Outliers removed
Correlation coefficient	0.9004	0.992	0.7562	1
Mean absolute error	0.2063	0.1748	0.3726	0.0106
Root mean squared error	0.3605	0.2357	0.5308	0.0124
Relative absolute error	31.22%	12.5176	56.38%	1.7491
Root relative squared error	45.63%	13.3197	67.19%	1.5358
Time taken	181.86	167.17	71.22	14.92

We found that there is a great difference in all the predictor measures. Initially, on the original data (complete or selected) correlation coefficient increased to .992 and 1 respectively. All the errors decreased and in selected data set, they are almost zero. This shows that the demographic data plays key role in Indian scenario. This also reflects that even after doing preprocessing of data, certain data treated to be normal where as they are hidden in the normal data set and only detected when applied the technique of outlier detection in the real estate domain. The coefficients found are very near to realistic structure of the real estate domain as the refinement in the weights at every step does not affected by the outlier data, which causes problem in the preprocessed data but removed after detecting the data, which are out of interest.

B. *Kmeans*

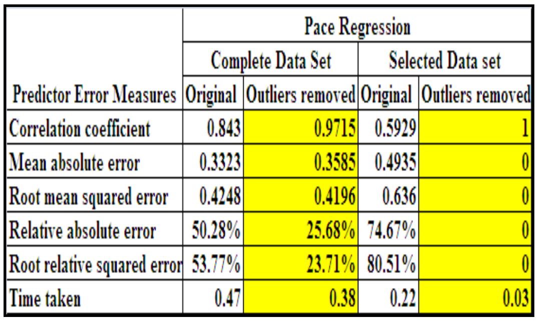
Table II depicts the results of kmeans, ie, the semi supervised outlier detection technique. The results of Kmeans on complete data set and selected data sets with outliers published in [5]. It is found, that there is a great difference in all the predictor measures. Initially, on the original data (complete or selected) incorrectly classified instances (ICI) reduced to 42.71% and 0% respectively. The sum of squared error (SSE) also reduced to 148129 and 936.91 respectively. This shows that the demographic data plays key role in Indian scenario, which also reflects that due to the presence of the outliers which are not the part of the data but are to be treated as part of the data affects the overall impact. It also shows that even after doing preprocessing of data, certain data are treated to be normal where as they are hidden in the normal data set and only detected when applied the technique of outlier detection in the real estate domain.

Table II Kmeans Results on Complete and selected data set



Predictor	Kmeans			
Error	Complete Data Set		Selected Data set	
Measures	Original	Outliers removed	Original	Outliers removed
ICI	69.64	42.7018	51.59%	0
SSE	150526	148129	86304	936.91

Table III Pace regression on complete and selected data set



	Pace Regression			
	Complete Data Set		Selected Data set	
Predictor Error Measures	Original	Outliers removed	Original	Outliers removed
Correlation coefficient	0.843	0.9715	0.5929	1
Mean absolute error	0.3323	0.3585	0.4935	0
Root mean squared error	0.4248	0.4196	0.636	0
Relative absolute error	50.28%	25.68%	74.67%	0
Root relative squared error	53.77%	23.71%	80.51%	0
Time taken	0.47	0.38	0.22	0.03

C. *Pace Regression*

As per the paper [1] we worked on four variations of regression and found pace regression is best among isotonic, least median square and linear regression. After considering the full data as a complete training set, we found that 210 tuples are under the category of outlier which occurs due to non acceptance of the hypothesis. After removal of outlier data when we applied the same regression techniques we got the better results which are shown as per table III.

The above table III shows the result on complete data set and selected data set and found that there is a great difference in all the predictor measures. Initially, on the original data (complete or selected) correlation coefficient increased to .9715 and 1 respectively. All the errors are improved and in selected data set they are zero. This shows that the demographic data plays key role in Indian scenario.

Now, even when we compare the results after outlier detection, we found that classification outlier detection technique is the best when applied on complete data set where as statistical outlier technique is best when applied on selected data set.

D. Apriori Rules

The rules induced after removal of outliers are found to be very near to realistic structure of the real estate domain as compared with the rules generated in paper [5]. The following rules are induced from the outlier free data set.

i. IPC='(2.6-3]' 690 ==> AI='(2.6-3]' 690 confidence (1)
ii. AI='(2.6-3]' 690 ==> IPC='(2.6-3]' 690 confidence (1)
iii. HS='(-inf-0.5]' 960 ==> INH='(-inf-0.9]' 919 confidence (0.96)
iv. NOH='(-inf-1.2]' HS='(-inf-0.5]' 840 ==> INH='(-inf-0.9]' 799 confidence (0.95)
v. INMH='(-inf-0.5]' 754 ==> INH='(-inf-0.9]' 712 confidence (0.94)

From the above rules it is very much clear that the investment power depends on Annual Income with confidence 100% higher status implies a customer belongs to higher income group. Number of houses also plays a significant role in determining the investing power of a class

IV. CONCLUSIONS

In this paper, we have discussed numerous ways to detect the outlier. We have found that even the results provided by the multilayer perceptron is better prior to outlier detection where as when same method is applied we found that the statistical detection techniques plays an important role in the outlier detection and gives the best results specially when applied on the selected data set. For each category of outlier detection technique, an efficient rule is induced. These rules not only improve the result but also enhance the knowledge in the form of efficient rule generation which will directly helpful in our research area i.e. Real Estate.

REFERENCES

[1] Banerji Geetali Saxena Kanak, Predictive Model- A Boon for Real Estate, in International Journal for Wisdom Based Computing Volume 2(1), April 2012, 6-11

[2] Yamanishi, K, Takeichi, J., and Williams, G., On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms, Proc. of the Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, (Boston, 2000), 320–324.

[3] Tan, P.-N., Steinbach, M., and Kumar, V., Introduction to Data Mining. Addison-Wesley., 2005

[4] Duda R. O., Hart P. E., and Stork, D. G. 2000. Pattern Classification, Wiley, University of Michigan, 2007

[5] Banerji Geetali, Saxena Kanak, 2012, Developing Rule Dynamics on Bank transactions and Design Specifications in Real Estate (Communicated).

[6] R. Bouckaert Remco, Eibe Frank et al, WEKA Manual for Version 3-6-2, January 11, 2010

[7] Witten I. Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Elsevier, San Francisco, 2005.

[8] Banerji Geetali Saxena Kanak, Analysis of Data Mining techniques on Real Estate, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, July 2012 Volume-2 (3), 223-230

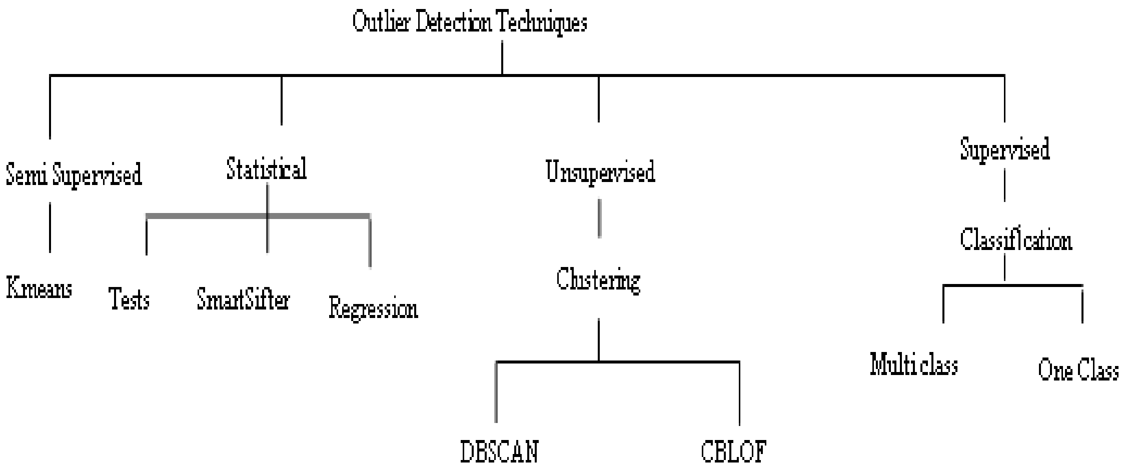
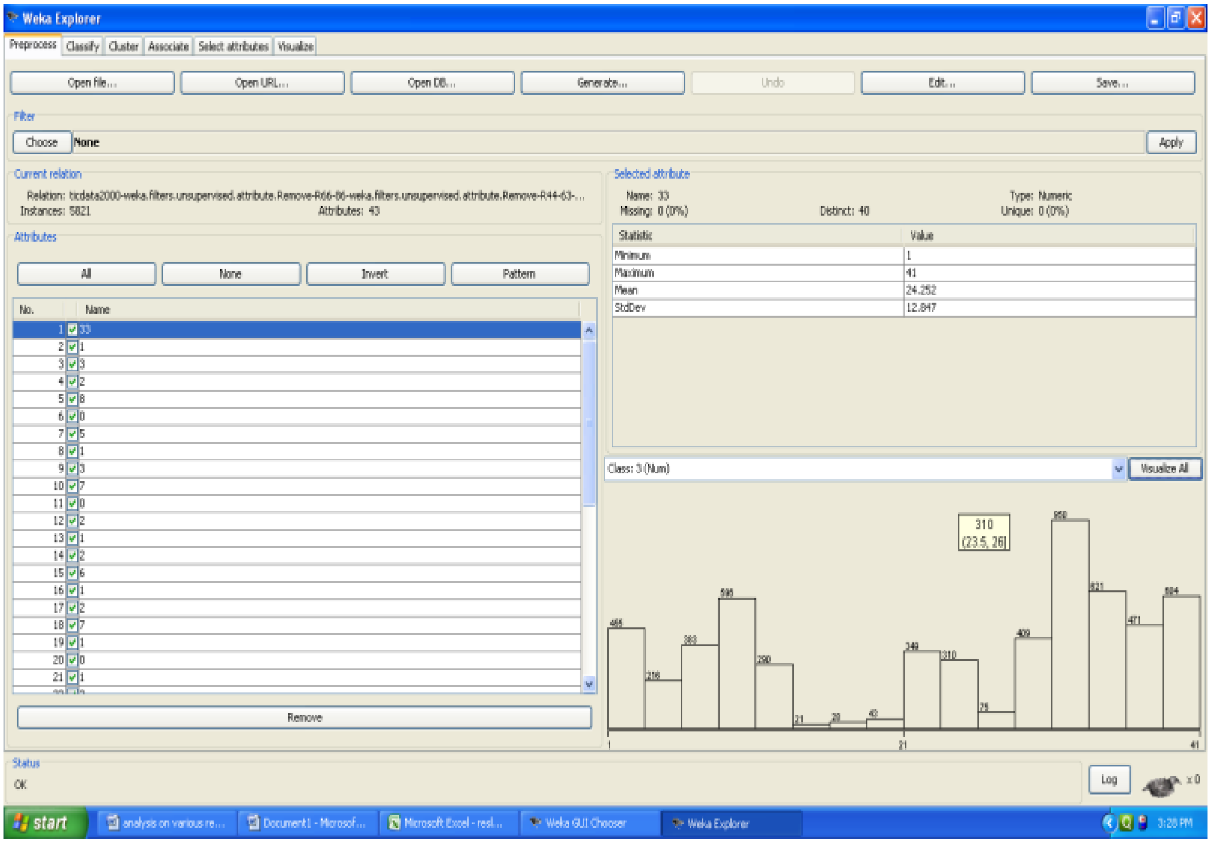


Fig 2 Outlier Detection Techniques



Filter: None

Current relation

Relation: ticdata2000-weka.filters.unsupervised.attribute.Remove-R66-86-weka.filters.unsupervised.attribute.Remove-R44-63-...
Instances: 5821
Attributes: 43

Attributes

No.		Name
1	☑	33
2	☑	1
3	☑	3
4	☑	2
5	☑	8
6	☑	0
7	☑	5
8	☑	1
9	☑	3
10	☑	7
11	☑	0
12	☑	2
13	☑	1
14	☑	2
15	☑	6
16	☑	1
17	☑	2
18	☑	7
19	☑	1
20	☑	0
21	☑	1

Selected attribute

Name: 33
Missing: 0 (0%)
Distinct: 40
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	41
Mean	24.252
StdDev	12.847

Class: 3 (Num)

Status

OK

Fig 3 Snapshot of Real estate data set in WEKA