Outlier Detection in High Dimensional Data Based on Elastic Net Regression

Ch. Anuradha, M. Ramesh, Patnala S.R. Chandra Murty

Abstract: ***Outlier detection in large datasets is the dynamic research area in computer science such as data mining, database systems, and distributed systems. Outlier detection faces many challenges due to the absence of data samples from the outlier class. Massive algorithms have been projected to conquer the challenges in this field to improve the efficiency of regression approach for large datasets. Currently, no particular efficient regression technique is designed for outlier detection. In this research, we proposed an ElasticNet regression model for detecting the outliers in high dimensional data. To validate the efficiency and competence of our projected algorithm, it is implemented in the open source software called Weka Explorer. The parameters such as Mean absolute error 0.0022, RMSE 0.0387, Relative absolute error (RAE) 0.4562 and Root relative squared error (RSE) 7.8722 are calculated using annthyroid dataset. ElasticNet model consumes less computational time, generates fast convergence results, provides high accuracy and correctly classified accuracy is 98.25%.***

Keywords**: Outlier detection, ElasticNet regression, High-dimensional data, Weka explorer, annthyroid.***

I. INTRODUCTION

Outlier detection is one of the most significant challenge in the field of machine learning. In today's modern society, availability and reliability of data have become crucial factors. One important task for any domain application is to detect abnormal data. Anomaly recognition approaches are utilized in several fields like fraud recognition in the banking system, intrusion detection in network security, unusual behavior in military surveillance, and also detection of tumors in MRI images. Outliers are defined as data points that happen very infrequently and/or lie far from the expected values. It may arise due to several reasons such as intrusion, human error, machine error, and changes in the behavior of the system. Because of all these various causes, outliers are difficult to detect, distinguish, and remove data consisting of noise. There have been many techniques proposed for outlier detection. Out of those techniques, some are specifically designed to suit some solicitation areas while others are more generic. The presence of outliers in data may carry important information. For example, outliers in Magnetic Resonance Imaging (MRI) may identify pixels, which are significantly different between two MRI scans and thereby indicate the presence of brain tumors. Similarly, an anomaly in digital photography may indicate that a terrorist is using steganography to hide messages in the low-order bits of a digital photograph in either plaintext or ciphertext form to disguise it from their enemies. Furthermore, an abnormal pattern in network traffic may signal an alarm of intrusion, which may indicate a compromised server is sending out unauthorized information. Other examples could be outliers in credit card transactions, which may raise attention to a credit card theft or interruptions in continuous signals from airplane to the ground due to inconsistent data acting as outliers, which may lead to accidents. In this work, we projected a heuristic outline for determining outliers in high dimensional data using ElasticNet Regression. The organization of this article is as follows: Segment 2 covers earlier research in outlier recognition. In Segment 3, we present regression based outlier detection methodology. Section 4 covers computational results and discussion, followed by summary of paper in Segment 5.

II. RELATED WORK

In current years, various methods are used to identify anomalies over large datasets. For managing large databases from high dimensional spaces, outlier recognition techniques are designed and that are non-parametric in nature. The quality of data is obtained by separating the ouliers from the databases and decreasing the effect of incorrect values in the process research. The significance of identifying anomalies can enhance the quality of saved data. Based on the annotations, an outlier recognition techniques are used by many machine learning algorithms and they are categorized into 3 methodologies: unsupervised, Semi-supervised and supervised anomaly detection methods. This section presents a brief summary of important anomaly detection methods and different types of anomalies detection techniques in data stream. Density-based probabilistic approach is projected by Charu C. Agrawal for anomaly detection using ambiguous datasets. In this projected research work, various probability functions are used in addition to estimation of density and sampling functions. For processing huge datasets effectively, micro clustering features are also used and compared the performance of projected method with existing methodologies. Finally, Projected Probability model is better over other. Distance-based anomaly detection approach was projected by Bin Wang on unclear data because previously projected approaches handling only certain data and the necessity of determining the unclear data is recognized.

Revised Manuscript Received on September 29, 2019.
***** Correspondence Author
Ch. Anuradha*, Dept. of CSE, ANU, Guntur, Andhra Pradesh, India
Dr. M. Ramesh, Dept. of CSE, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India
Dr. Patnala S.R. Chandra Murty, Dept. of CSE, ANU, Guntur, Andhra Pradesh, India

For this purpose, Wang projected grid-based pruning and dynamic programming method for recognition of anomalies and works efficiently in unclear data. Graph-based anomaly recognition approach is projected by Ville Hautamaki by detecting drawbacks in existing graph-based approach using kNN. Performance of projected method is tested using synthesis and real databases. Projected system generates efficient outcomes with synthesis database as compared to teal databases and also identified reasons for poor performance.

An efficient spatial anomaly recognition algorithm was projected by Chang-Tien Lu for overcoming the drawbacks in existing spatial anomaly recognition approach. The performance of the projected approach is test with real world census databases. The projected system works efficiently and generates accurate results and also determines what are the false anomalies and true anomalies in the given database. Clustering-based anomaly recognition approach was projected by Elahi et al. for dynamic databases. In this research work stream of data is divided into groups and portions and stable no of cluster groups are generated using k-mean algorithm. The anomalies in the given database is identified by exploiting average value of present portion with the average value of preceding one. Experimental values demonstrates the effectives of projected method. It is clear from the literature survey, detection rate and detection accuracy is needed to be improved.

III. ELASTIC-BASED OUTLIER DETECTION

In this section, we projected ElaticNet-based regression method for detection of outliers. Here, we use simple and shorter statistic expressions for defining the robust ElasticNet model. Explicitly, shorter inner products are replaced with original inner products in the ElasticNet regression model. The simple ideal behind this shorter inner products is the points that are very large in the inner products are corrupted. That means the outliers having too large magnitude have positive impact on the correlation and they can be treated as true anomalies. Else, they have negative impact on correlation and can be conspired as outliers. Consider the following simple mathematical model of ElaticNet regression approach:

$$\tilde{\beta} \hat{} \in \arg \min_{\beta 1 \le R} \{ \frac{1}{2} \beta T \Gamma \hat{} \beta - < \tilde{\gamma} , \beta > \quad \ldots\ldots (1)$$

Here $\Gamma\hat{}$, $\gamma\hat{}$ indicates the unbiased estimations of Σx and $\Sigma x\beta*$ respectively. $\beta*$ is the single resolution to the given program $R \geq k\beta *k1$.

Lasso regression is different circumstance with $\Gamma\hat{}$ LAS = XTX and $\gamma\hat{}$LAS = XT y in equivalence 1. Under a appropriate λ value and a optimistic semi-positive $\Gamma\hat{}$, Equivalence (1) has a comparable standardised formula as

$$\beta\hat{}\tilde{\beta}\hat{}\in \arg \min_{\beta 1 \le R} p\{\frac{1}{2} \beta T \Gamma \hat{} \beta - <\tilde{\gamma},\beta> + \lambda \|\beta 1\|\} \ldots (2)$$

Here $\|\beta 1\|$ would be greater constrained by b $\sqrt{}$ k for a appropriate fixed value b if $\Gamma\hat{}$ has undesirable eigenvalues. If $\Gamma\hat{}$ EN $=\alpha XTX + (1-\alpha) \cdot I$, $\alpha \in [0, 1]$ and $\gamma\hat{}EN = XTy$, Equivalence (1) develops elastic net prototype. But, both the elastic net and the Lasso are delicate once exploitation happens, particularly below our clamour situations. Therefore we usage the subsequent shorter statistics as strong substitutes for $\Gamma\hat{}$ and $\gamma\hat{}$:

$$\{\Gamma\hat{} REN\}ij = \alpha \sum_{k=1}^{n0} [Xi , Xj](k) + (1 - \alpha) \cdot Iij \{\gamma\hat{}REN\}j = \sum_{k=1}^{n0} [Xj,y] (k) \quad \ldots\ldots (3)$$

Wherever $\alpha \in [0, 1]$, and $[u, v](k)$ $(u, v \in R\ h\ )$ designates the kth minutest mutable in $\{qi = |ui \cdot vi|, \forall i\}$ such that $[u, v](1) \leq [u, v](2) \leq \cdots \leq [u, v](h)$. I is an unit matrix. Xi designates the ith column of the matrix X.

Hence it is essential to build a procedure which can estimated the optimal, else the enactment assured may be unusable. We usage the subsequent polynomial-interval expected gradient descent update:

$$\beta\ t+1 = \Pi l1(R)\ (\ \beta\ t\ -\ 1\ /\ \eta\ (\Gamma\hat{}\ REN\beta\ t\ -\ \gamma\hat{}REN) \quad \ldots\ldots (4)$$

Wherever $\Pi l1(R)(v) = \arg \min z\{v - z\}2 \mid \{z1\ \} \leq R\}$ designates Euclidean prediction onto a l1 ball of area R. The optimal fault is also constrained so that the projected gradient descent can estimated the optimal with adequate precision.

We present a novel projected approach for the REN prototype. In command to authenticate the enactment of ElasticNet model, it necessary that to consider an efficient methodology. The gradient loss function is defined as follows:

$$\nabla L(\beta)=\Gamma\hat{}REN\beta-\gamma\hat{}REN \quad \ldots\ldots (5)$$

We use the projected gradient descent that produces a series of repeats $\{\beta t, t = 0, 1, 2, \cdots\}$ by calling the expression itself:

$$\beta\ t+1 = \arg \min_{\beta \le R} \{L(\beta\ t\ ) + <\nabla L(\beta\ t\ ), \beta - \beta\ t\} + \eta\ /2\ \{\beta - \beta\ t\ \}\ 22 \quad \ldots\ldots (6)$$

Where $\eta > 0$ is a step-size factor. His expression transform from l2 to l1 expression that is equal to Equivalence (4). The equations 1 and 2 are convex, or consistently, then the above equivalence is guaranteed to converge to the global optimal. The procedure is shortened as follows.

Algorithm for detection of outliers based on ElasticNet Regression:

ElasticNet-based Regression

Response: X, y, R, no, α

Production: $\beta\hat{}$

Calculate $\Gamma\hat{}$REN and $\gamma\hat{}$REN via Equivalence (4)

while not influence highest repetition or fulfilled accurateness do

{

Execute the estimated gradient descent procedure by repeating Equivalence (7) to resolve Equivalence (2).

}

Output the final β^.

IV. RESULTS AND DISCUSSION

To authenticate the efficiency and competence of our projected algorithm, proposed method is implemented in the open source software called Weka Explore. To evaluate the performance of projected method, the annthyroid dataset was obtained from the University of Alberta. This dataset has 6633 observations and 21 attributes and this data set contains medical data on hypothyroidism. Outputs of given dataset are categorized into 3 classes normal, hyper function, and subnormal functioning. Data items with distinct behaviour are recognized as anomalies here. Sample annthyroid dataset is shown in table 1:

Table-I: Sample Page blocks dataset

att1	att17	att20	id	outlier
0.75	0.001132	0.300926	1	'no'
0.239583	4.72E-04	0.537037	2	'no'
0.479167	0.003585	0.527778	3	'no'
0.65625	0.001698	0.337963	4	'no'
0.229167	4.72E-04	0.337963	5	'no'
0.708333	4.72E-04	0.24537	6	'no'
0.875	4.72E-04	0.402778	7	'no'
0.489583	0.003925	0.282407	8	'no'
0.6875	0.002453	0.425926	9	'no'
0.78125	1.89E-04	0.513889	10	'no'
0.635417	0.020755	0.263889	11	'yes'
0.177083	1.89E-04	0.314815	12	'no'
0.604167	0.001509	0.37963	13	'no'
0.5	0.001132	0.393519	14	'no'
0.541667	0.00434	0.361111	15	'no'
0.395833	1.89E-04	0.25	16	'no'
0.395833	0.001132	0.384259	17	'no'
0.666667	0.003019	0.347222	18	'no'
0.65625	0.060377	0.458333	19	'no'
0.510417	0.115094	0.458333	20	'yes'

The following screen (as shown in figure 1 and 2) short shows detection of anomalies are present in the given annthyroid dataset and also shows the pre-processing stage:

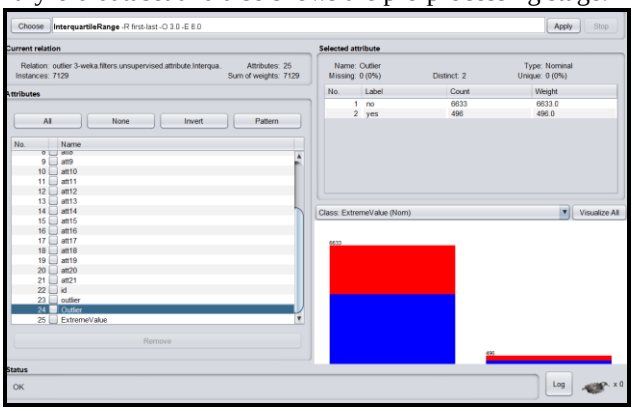

Fig. 1. Detection of Outliers of annthyroid dataset

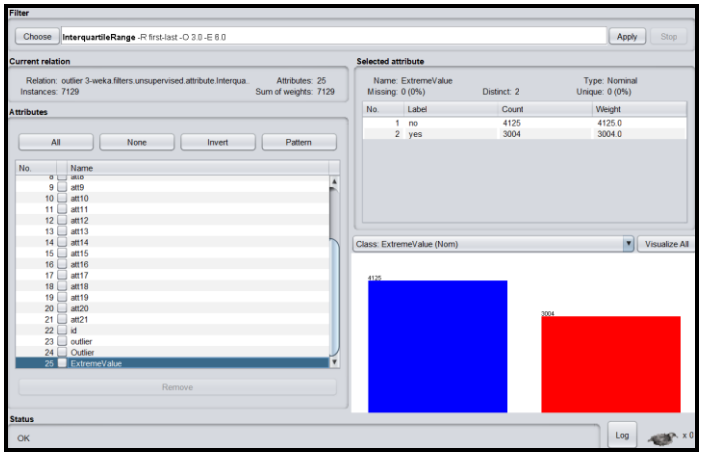

Figure 2: Pre-processing in annthyroid dataset

The following screen (as shown in figure 3) short shows removing anomalies from the given annthyroid dataset using projected ElasticNet regression methodology:

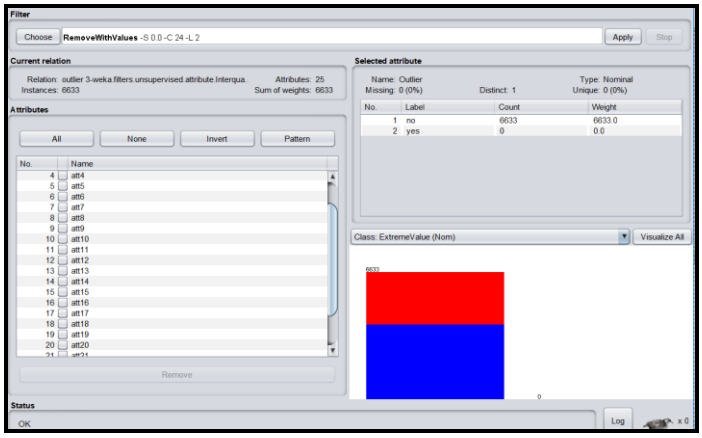

Fig. 3. Removing the outliers in annthyroid dataset using Elastic Regression

Our proposed algorithm is implemented in the open source software called Weka Explorer as shown in figure 1-3. The parameters such as Mean absolute error 0.0022, RMSE 0.0387, RAE 0.4562 and RSE 7.8722 are calculated using annthyroid dataset. Projected method generates the outputs accurately, quickly, consumes less time and correctly classified accuracy is 98.25%. From result simulation it has been found out that best algorithm is ElasticNet regression with optimal time complexity. Several methods are used for detection of outliers from a given data set. Each method aims to detect the outliers and gives the best result to other. We compare these methods according their advantages and disadvantages to find best method. The table 2 illustrates that the comparison of distinct existing methods with our proposed method to find an outlier detection.

Table -II: Comparison of various outliers detection approaches

Methods	Accuracy	Detection Rate	Sensitivity	False alaram rate
Distance-Based Approach	84%	76.14%	95%	65.01%
Deviation- based Approach	80%	67.27%	93%	71.34%
Depth-based Approach	85%	75.54%	96%	59.74%
Statistical- based Approach	78%	74.16%	95%	69.38%
Regression-based Approach	98.25%	78.18%	98%	80.12%

The performance comparison of distinct anomaly recognition methods is done with the help of distinct databases and factors. It is difficult for deciding the performance of projected method when it implemented individually without any comparison. In this research, the assessment is made with different parameters such accuracy, detection rate, sensitivity, false alaram rate etc. Calculations of these parameters designate that an outlier recognition method based on ElasticNet Regression was the most capable out of the studied approaches. In this comparisons, fiver algorithms i.e Density-based, distance-based, deviation-based, depth-based and regression-based approaches; has been compared based on proposed dataset. The comparative results demonstrates that the projected method detects the anomalies in high dimensional data with potentially higher accuracy.

V. CONCLUSION

The outlier detection plays an interesting and significant role because the elimination of false outliers may impact the extracted outcomes to a larger level if it is a significant information required for investigation. Outlier detection faces many challenges due to the absence of data samples from the outlier class. Massive algorithms have been projected to conquer the challenges in this field to improve the efficiency of regression approach for large datasets. Currently, no particular efficient regression technique is designed for outlier detection. In this research, we proposed an ElasticNet regression model for detecting the outliers in high dimensional data. To validate the efficiency and competence of our projected algorithm, it is implemented in the open source software called Weka Explorer. The parameters such as Mean absolute error 0.0022, RMSE 0.0387, Relative absolute error (RAE) 0.4562 and Root relative squared error (RSE) 7.8722 are calculated using annthyroid dataset. ElasticNet model consumes less computational time, generates fast convergence results, provides high accuracy and correctly classified accuracy is 98.25%.

ACKNOWLEDGMENT

The authors would like to thank the publisher and reviewers for many helpful remarks and recommendations, which resulted in a substantial enhancement in the presentation of this article.

REFERENCES

1. Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi "An Outlier Detection Method based on Clustering", Second International Conference on Emerging Applications of Information Technology, 2011.
2. V. Barnett and T. Lewis. Outliers in Statistical Data. John Wiley&Sons, 3rd edition, 1994.
3. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, CURE: an efficient clustering algorithm for large databases, ACM LIBRARY, 1999.
4. R. Ng. E and Knorr "Algorithms for Mining Distance-based Outliers in Large Data Sets", VLDB Conference Proceedings, September 1998.
5. Chang-Tien Lu, Dechang Chen, Yufeng Kou, "Algorithms for Spatial Outlier Detection," Third IEEE International Conference on Data Mining, 2003.
6. T. Ptacek, T. Newsham, Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection, Secure Networks Inc, 1998.
7. K. Bache and M. Lichman. UCI machine learning repository, 2013.
8. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. IEEE Transactions on Knowledge and Data Engineering, 24(5):823– 839, 2012.
9. Pachgade, Ms SD, and Ms SS Dhande. "Outlier detection over data set using cluster-based and distance-based approach." International Journal of Advanced Research in Computer Science and Software Engineering 2, no. 6 (2012): 12-16.
10. Ghosh S, Reilly D (1994) Credit card fraud detection with a neural network. Proceedings of 27th Hawaii International Conference on Systems Science 3:621–630.