SIMPLE TECHNIQUE FOR DETECTION OF OUTLIERS IN ONE-DIMENSIONAL NUMERICAL DATA USED FOR POINT OUT ANOMALOUS CONSUMPTION

Davi Mantovani RICCI
Daimon – Brazil
davi.ricci@daimon.com.br

Paulo Henrique BAUMANN
Daimon – Brazil
paulo@daimon.com.br

Fabio ROMERO
Daimon – Brazil
fabio.romero@daimon.com.br

André MEFFE
Daimon – Brazil
andre@daimon.com.br

Armando H. S. G. JESUS
Equatorial Group – Brazil
armando.jesus@cemar-ma.com.br

Eliezer S. OLIVEIRA
Equatorial Group – Brazil
eliezer.oliveira@cemar-ma.com.br

ABSTRACT

This paper presents a simple technique for detection of outliers in one-dimensional numerical data. This was developed to point out anomaly in the consumption information in the DSO's database. This, in turn, was inspired from concepts like DBSCAN and K-Means grouping techniques. It has been shown plausible for data whose distribution curve is skewed, positively or negatively.

INTRODUCTION

This work presents one of the tools implemented in the BIAT (BDGD Integrity Assurance Tool) Software, a result of the Research and Development Project of ANEEL (Brazilian Electricity Regulatory Agency), proposed by the Equatorial Group (CEMAR and CELPA) and executed by Daimon Engineering and Systems. This tool can be defined as a binary classifier that assesses if a given value from a set of integers representing the consumption (kWh) is an outlier or not. The BIAT Software is composed of a set of intelligent algorithms, whose purpose is to detect anomalies in the database of assets and customers of Distribution System Operators (DSOs), named Geographic Database of the Distribution System (BDGD). As required by ANEEL, the DSOs must send the BDGD annually to update the SIG-R (Regulatory Geographic Information System). The SIG-R was created with the objective of improving the methodologies of supervision and regulation of the generation, transmission and distribution of electric energy in Brazil. Some of the tables from such database contains the consumption of consumer units in only one month of the year and a reasonable number of registers shows some implausible information.

Due to the simple appearance of the problem, the methods of Tukey [1] and Adjusted Boxplot for Skewed Distributions [2] were first tested. However the poor thresholds provided resulted in a high false positive rate. Then, an investigative activity was carried out on the consumption data according to their categories. To better understand its patterns, the data was segregated in many categorical groups according to the intersection of three variables, namely voltage level (medium voltage or low voltage), consumption class (residential, commercial, industrial, rural, and others) and phase configuration (single, two or three-phase, with or without neutral).

Among the analyses that have been done, it is important to highlight the one that resulted in the conception of the binary classifier. It consists of extracting a sample of the unique values from each categorical group (i.e. there are no repeated values in the sample). For some samples, the distance between a value and its closest neighbor (*connectivity*) showed some important disruptions along the range, in a way that it was possible to form clusters of related points. The cluster whose range of values had the greater representativeness was taken as the profile of the categorical group, while the others were considered outliers.

To group adjacent points, two clustering methodologies were adapted for one-dimensional data and to act as binary classifiers, the DBSCAN (Density-based spatial clustering of applications with noise) [3] and the K-Means [4]. The determination of the initialization parameters of these methodologies was a very complex task, which makes the automation of the tool unfeasible. Nevertheless, based on the respective operational concepts, an efficient methodology was developed.

THE THEORY

DBSCAN

The DBSCAN methodology adapted for one-dimensional data is summarized in classifying which of the closest pairs of points are connected or not from a connectivity parameter R_O. In order to accomplish that, the values of the sample are arranged in ascending order so that the adjacent points of the consequent arc are the closest points. An exhaustive scan of adjacent pairs of points is then carried out to verify which are connected or not. If the distance between two points is less than or equal to R_O then they are connected, otherwise they are not connected. Through this procedure, one or more clusters are formed. The largest group in terms of number of points was considered the profile of the categorical group from where the values were extracted. The values from the other clusters (if any) were considered outliers.

K-Means

The K-Means adapted for one-dimensional data has as initialization parameter the number N of groups desired. The algorithm then calculates the N percentiles of the values and takes them as initial means. The distances between each point and each mean are calculated. Based on this, the points are rearranged so that they are allocated in the group that has the closest mean, and.then the means are recalculated. The process repeats itself until the means do not suffer further changes. Just like the procedure used in the previous technique described, the largest group is taken as the profile, and the values from the other groups are dubbed outliers.

BIAT Binary Classifier

For the application of the method, a curve is obtained from the arrangement of the sample values in ascending order. This curve is then approximated by an arc formed by the union of two line segments. The first step is to unite the extreme points of the arc by a line segment. Then, the farthest point of the arc to that line is taken as the divider point, splitting the set in two. The next step is to define the aforementioned two segments: From the first point to the divider point there is a line segment, and from the divider point to the last point of the arc, there is another line segment. The context is illustrated in Figure 1.

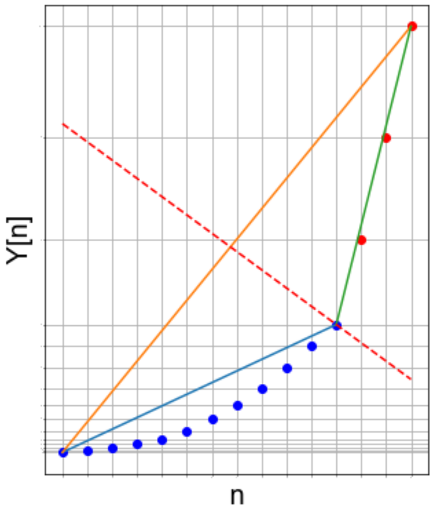

Figure 1 - Division inferred by BIAT Binary Classifier.

The slopes of the straight line segments used in the approximation represent the means of the variation rates of the two arcs segments resulted from the division. One feature noted in the data analysis was that for cases where the slopes of the straight segments are significantly different from each other, the range of lower slope values had greater representativeness (largest number of points), indicating that the values that define the profile are closer together. The algorithm acts recursively on the longest arc. The stopping criterion is defined by equations (1) and (2).

$$\Delta Y_{MEAN} = \left(\frac{1}{N} \sum_{j=1}^{N} \left| \Delta Y_j \right| \right) \qquad (1)$$

$$R = \frac{N_{Sample}}{N_{Data}} \qquad (2)$$

Equation (1) calculates the mean connectivity of the arc subject to the next iteration and has a similar meaning to that of connectivity R_O, but with greater adaptive capacity. Equation (2) is the ratio of the sample size and the mass to which the sample was extracted, used to preserve the representativiness of the sample.

APPLICATION INTO COMPANY

It is convenient to highlight some details that have made the problem challenging despite its simple appearance. The first is that the database available contains the consumption of only one month of the year for all the consumers associated to the electric power distributor, therefore avoiding the usage of historical information to discern which values are plausible and which are not. Another issue is the target audience of the software, that are not expected to have deep knowledge in statistics to adjust a complex technique. Thus, the simpler and more automated the tool, the more suitable for its target audience. The tool was developed so that the data quality managers of the BDGD simply run it on the desired data and then execute a manual inspection of the positives. As expected, if the parameters of the technique are made too tight aiming no false negatives, there will probably be a high false positive rate, increasing the volume of data do be double-checked. A tradeoff must be found, and for sake of parsimony, the authors think it is preferable to guarantee the non-occurrence of false positives with a low false negative rate.

Insertion of the tool in the management of the data's quality register

The team of DSO responsible for managing the BDGD data quality runs the tool on the customers' consumption table. . It groups consumers categorically (voltage level, consumption class and phase configuration). On each categorical group, it extracts the sample of unique values. On these, the binary classifier is applied. So, it returns the set of consumers whose consumption value was identified as implausible. The managers of the BDGD then carry out a manual inspection, quantifying the false and true positives. True positives are fixed and adequate.

The failure in database record is initially identified by the anomaly in the consumption value as a function of the categorical group. However, the consumer registry is not limited to the categorical variables and monthly value of electric energy consumption. It also contains its georeferenced information, which allows an intrinsic mitigation in the process of altering the database records, which begins in the measurement of the electrical energy consumed and ends in the registration in the BDGD. The investigation based on the inconsistency of consumption may point to other issues besides the inconsistencies of consumption values such as an infringing commercial activity.

VALIDATION OF DATA – IN PROGRESS

In binary classifiers, either the classification is right or wrong. Because of that, the performance of the classifier is measured as a function of the amounts of hits and misses in reject or not the null hypothesis "The value Yi is not an outlier" for a representative sample. Given these figures, it is quantified the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

The analysis of the performance of BIAT's binary classifier is verified on a sample set. The tool operates grouping the sampled values categorically (consumption class, number of phases and voltage level). Then, the consumers are rearranged into two subsets, one with outlier consumption and other with regular consumption. The number of registers for the second subset is much larger than the number of the first subset. However, in order to avoid the impact of such prevalence in the analysis that could skew the conclusions, it is performed a stratified sampling with each stratum having the same size. The information used to stratify is the classification made by the tool.

The reference values to assess if the tool has made the right classifications for each consumption of the sample is based on the monthly consumption data coming from the DSO's billing system for the last 12 months. Using these data, a model is elaborated for each consumer unit to estimate its average consumption. Using the model, consumption prediction intervals were calculated, whose probability of containing the real value is 80%. These intervals are taken as limits to quantify successes and failures in the process of classification (as credible or improbable) of consumption information by BIAT.

The chosen figures of merit were sensitivity, precision and accuracy. The sensitivity is the ratio of positives detected (true positives) to the total positives (false negatives and true positives). Sensitivity is depreciated by the occurrence of false negatives. Precision is defined by the ratio of positives detected (true positives) to the number of objects denoted as positive, whether false or true. Precision is depreciated by the occurrence of false positives. Sensitivity and precision were chosen as performance metrics because they point the way to reduce database records inconsistencies with efficiency (i.e. with minimal false positives). On the other hand, accuracy shows a general performance measure. It is calculated by the ratio of total hits (true positives and negatives) to the total objects evaluated.

For a solid validation of the binary classifier performance, it has to be compared with algorithms that, for the context, are defined as baselines through a hypothesis test named Sign Test [5]. The baselines chosen were Zero Rule and Random Guessing. The Zero Rule was chosen because it is the simplest classification method, whose mechanism is to reject the null hypothesis for all instances or not to reject it due to its prevalence. If the prevalence of positives is greater than 50% of the population, then the null hypothesis is rejected for all objects in the sample, otherwise it is not rejected. This has no significant predictive ability, but it is useful to be used as a benchmark in binary classifier validation tests. Its for such application can be found in [6]. It is assumed that the occurrence of database records inconsistencies is less than non-occurrence, so it is estimated that the number of positives is less than the number of negatives. Because of this, Zero Rule will not reject the hypothesis for all instances and its misclassifications will be only false negatives. So this is a good comparator if the statistic evaluated is the sensitivity. However, the accuracy, as described, is the ratio of all hits (whether true or negative) and all objects evaluated, so this statistic is influenced by both the nature of classification errors (false positives and false negatives). That said, Random Guessing was chosen so that it was also considered false positives in the comparative process, even though it is a counter-indicated methodology for this situation [7]. This also consists of a simple mechanism, since from the proportion on the sample under study Random Guessing randomly ranks the instances.

The Sign Test aims to compare the random variables over a same group of instances, which were obtained by different processes or at different instants. For this case, performance statistics are calculated for all classifiers in question. The test of the signal evaluates whether the contrast in the performance is associated with the different operational mechanics that constitute them or it is reduced to a random event.

ANALYSIS AND RESULTS

In this context, an outlier detector, besides being precise and sensitive, it also needs to be autonomous and easy to operate. Given this, besides a real scenario extracted from the DSO's database, three other scenarios were derived with the purpose of highlighting the disadvantages of the rejected techniques and advantages of the elaborated technique.

The real scenario (A) is formed by all the low-voltage single-phase residential consumers, whose consumption range varies from 1 to 366,110 kWh and consists of 1,782,860 records. The first derived scenario (B) was obtained by removing from A the records ranging from 50 kWh to 100 kWh (exclusion of 559,842 records). The second derived scenario (C) was obtained by purging from scenario A the largest outliers, whose consumption ranges from 10,356 kWh to 366,110 kWh (exclusion of 9 records, or 0.0005%). Finally, the third derived scenario

(D) was obtained from the intersection of the scenarios B and C.

The BIAT Binary Classifier was subjected to a calibration process, in which the default parameters were estimated for the tool. They were adjusted aiming high precision and sensitivity for all categorical classes, prioritizing those of greater representativeness. The average connectivity adopted was $\Delta\text{kWh} = 1.5\,\text{kWh}$ and the adopted ratio was $R = 0.05$. The profile consumption interval resulted from the analysis of the real scenario by BIAT's classifier (using default parameters) was used to calibrate the initialization parameters of the other methodologies. Given this, a connectivity value of $R_O = 8\,\text{kWh}$ was used for DBSCAN and 4 clusters for K-Means, among which only one is taken as profile and the others are rejected. Standard Box Plot (or Tukey method) and Adjusted Box Plot techniques do not require boot parameters, since they point out the profile's limits from percentiles.

Erro! Fonte de referência não encontrada. shows the intervals calculated by the Standard Box Plot (or Tukey method) and Adjusted Box Plot techniques for each scenario. The records not contained in the intervals are rejected and pointed out as anomalous.

Table 1 – Intervals [kWh]

Scenario	Standard Box Plot		Adjusted Box Plot	
A	-86.0	290.0	0.59	437.02
B	-184.5	403.5	-117.28	472.64
C	-86.0	290.0	0.59	437.02
D	-184.5	403.5	-117.28	472.64

The narrow ranges previously presented lead to relatively high rates of positives (alleged outliers), as shown in **Table** 2, which shows the percentage of the total number of records of each scenario thus classified. Narrow bands and exuberant amounts of positives in the context are strong indications of low precision. Although the methodology is highly sensitive and autonomous, its usage is undesirable because its low precision and hence large number of positives (among them many false positives) makes assiduous positive verification activity unfeasible.

Table 2 – Positive Rate [%]

Scenario	Standard Box Plot	Adjusted Box Plot
A	6.54	2.25
B	4.14	2.61
C	6.54	2.25
D	4.14	2.61

Erro! Fonte de referência não encontrada. presents the intervals of the profiles offered by the clustering techniques acting as binary classifiers. In Table 4, it is shown the proportions of positives detected by each classifier in each scenario.

Table 3 – Intervals [kWh]

Scenario	K-Means		DBSCAN		BIAT	
A	1	2564	1	2170	1	2197
B	1	2608	101	2170	1	2197
C	1	1000	1	2170	1	2392
D	1	1026	101	2170	1	2392

Despite the relationship of inverse proportionality noted between the mean value of connectivity for a group of records contained in a given range of consumption and their respective representativeness in the group, connectivity disturbances may be common.

In the operational perspective of DBSCAN, a disturbance can mean the division of records into different groups. This, in turn, can lead to countless false positives in the classification process.

Since DBSCAN connectivity was set at 8 kWh, any connectivity value above that level would lead to an unwanted profile split. This situation can be verified by comparing the scenario A with scenarios B and D, where the excluded range (50 to 100 kWh) causes a split in the profiles obtained by DBSCAN (i.e. instead of 1 kWh to 2,170 kWh, it became 101 kWh to 2,170 kWh). This change increased the positive rate from 0.019% to 32.289%. On the other hand, DBSCAN still has the advantage of not having strong influences due to excessively anomalous values, which can be verified by comparing scenario A and C, whose largest outliers were excluded.

Table 4 – Positive Rate [%]

Scenario	K-Means	DBSCAN	BIAT
A	0.013	0.019	0.02
B	0.018	32.289	0.03
C	0.172	0.018	0.01
D	0.231	32.289	0.02

The K-Means technique, in turn, is based on measures of central tendency. Because of this characteristic, extremely anomalous values in the mass provide significant influence on the profile limits and they can be verified by comparing the limits obtained by the technique in the scenario A with the ones obtained in scenarios C and D. On the other hand, anomalies in connectivity do not cause significant influences, as it can be seen comparing the scenario A and B.

The BIAT binary classifier carries the advantage of the cluster techniques presented, which can be noticed in all scenarios, since the intervals of the profiles were minimally altered, as well as the respective positive rates.

CONCLUSION

The operational philosophy of the BIAT Binary Classifier was inspired by the DBSCAN and K-Means techniques, since the former suggested the concept of connectivity and the latter the use of measures of central tendency for clustering, but applied on the connectivity values and not on the consumption values.

The union of the concept of measures of central tendency and the concept of connectivity was possible through the mechanics of approximation of the characteristic curve to an arc constituted of two straight segments.

Through the previously presented results, which were obtained in absurd but possible hypothetical situations, the BIAT Binary Classifier was shown to be stable, precise and sensitive. This makes the methodology suitable for the detection of anomalies in the information of the consumption in DSO's database, since the precision and sensitivity of the technique do not depend on the statistical knowledge of the operator.

Some boundary conditions are interesting for the technique to perform well. The first of them is related to the similarity between the characteristic curve (curve obtained by the arrangement of the unique values in ascending order) and the respective approximation arc (arc consisting of two segments of lines), where the greater the similarity, the better the performance. Another condition is the inverse relationship between the average connectivity of the characteristic curve and the representativeness of the respective range of values in the data. The last condition is the difference between the inclinations of the segments (mean of the connectivities), where the greater it is, the better the performance.

Within this, it can be stated that for cases where the probability distribution is skewed, positively or negatively, the method keeps effective.

It should be noted that the outlier definition also depends on the aspects of the context in which the data are involved. This can be exemplified by comparing the results offered by the Adjusted Box Plot and the BIAT Binary Classifier. The Adjusted Box Plot methodology is a percentile-based technique adapted for data with skewed distribution curves. However, because of the characteristics of the context, adopting it as a binary classifier resulted in low precision results.

The precision, sensitivity and accuracy of the methods will be estimated as described in section VALIDATION OF DATA – IN PROGRESS, but the conclusion of this step depends on the historical consumption of the units sampled by the commercial sector of the DSO, an activity still in progress.

REFERENCES

[1] H. Beyer, "Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S.", *Biom. J.*, vol. 23, n^o 4, p. 413–414, 1981.

[2] M. Hubert e E. Vandervieren, "An Adjusted Boxplot for Skewed Distributions", *Comput Stat Data Anal*, vol. 52, n^o 12, p. 5186–5201, ago. 2008.

[3] M. Ester, H.-P. Kriegel, J. Sander, e X. Xu, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996, p. 226–231.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations", presented in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 1967.

[5] W. J. Dixon e A. M. Mood, "The Statistical Sign Test", *J. Am. Stat. Assoc.*, vol. 41, n^o 236, p. 557–566, 1946.

[6] C. Nasa e S. Suman, "Evaluation of Different Classification Techniques for WEB Data", *Int. J. Comput. Appl.*, vol. 52, n^o 9, p. 34–40, ago. 2012.

[7] J. Brownlee, "Do Not Use Random Guessing As Your Baseline Classifier", *Machine Learning Mastery*, 04-mar-2016. .