International Journal of Scientific Research and Management (IJSRM)
||Volume||07||Issue||01||Pages||EC-2018-158-161||2019||
Website: www.ijsrm.in ISSN (e): 2321-3418
Index Copernicus value (2015): 57.47, (2016):93.67, DOI: 10.18535/ijsrm/v7i1.ec01

A data-oriented approach for outlier detection

Nripesh Trivedi,

Department of mathematical sciences. Indian Institute of Technology, Varanasi, India

Abstract
In this paper, characteristics of data obtained from the sensors (used in OpenSense project) are identified in order to build a data-oriented approach. This approach consists of application of Class Outliers: Distance Based (CODB) and Hoeffding tree algorithms. Subsequently, machine learning models were built to detect outliers in a sensor data stream. The approach presented in this paper may be used for developing methodologies for data-oriented outlier detection.

Keywords— Data characteristics, Nominal attribute, outlier analysis, machine learning, model verification

I. Introduction

Numerous algorithms have been proposed to detect outliers. However, for this paper, the approach adopted is exclusive to the data-set under consideration. Since the approach is meant exclusively for the data-set, it may be termed as a data-oriented approach. This paper may be the first to use to a data-oriented approach for outlier analysis.

II. Outlier detection: a data-oriented methodology

A. Data-set Description

The Data-set used in this paper was made available by Dr. Jean Paul Calbimonte who is an active researcher in the OpenSense project. Size of the data-set is more than 16 million rows. Around 100000 rows of this data-set were used to build machine learning models while other rows of this data-set were used to build a data stream that these machine learning models could use in prediction. The attributes in the data-set are:

- Latitude
- Longitude
- LDSA
- Station

In OpenSense project, sensors were positioned on top of the buses (public transportation) to measure pollution levels in Lausanne city. The term LDSA stands for lung deposited surface area. It is a way to measure the quantity of particles

Table I: Data Attributes and their respective range of values

Latitude	46.5202347 - 46.5218066
Longitude	6.6307456 - 6.6315791
Station	41, 43, 45, 47. 48, 49, 50, 51, 54, 55
LDSA	1 - 2000

10 stations indicated in the table I include both static and mobile stations.

III. Methodology for outlier detection

Uniform Random sampling is a method for summarizing multidimensional data streams [1]. Using this method, data-sets could be sampled. This sampling was applied on the data-set described in Table I. The data-set obtained from sampling (say data-set I) was treated in the following manner:

- CODB Algorithm was applied over data-set I to detect outliers. Further, data-set I and outliers detected in this data were used for building machine learning models that could learn from a data stream and find outliers in it.

Station is a numerical attribute but it may be used as a class (nominal) attribute. The advantage of using Station attribute as a class attribute is that class labels may be assigned to each row in the original data-set and also data-set I (the data-set obtained from sampling). An instance in Data-set I is of the form (Station, Latitude, Longitude, LDSA) where Station is a nominal attribute and other attributes are numerical.

Class Outliers: Distance Based (CODB) Algorithm

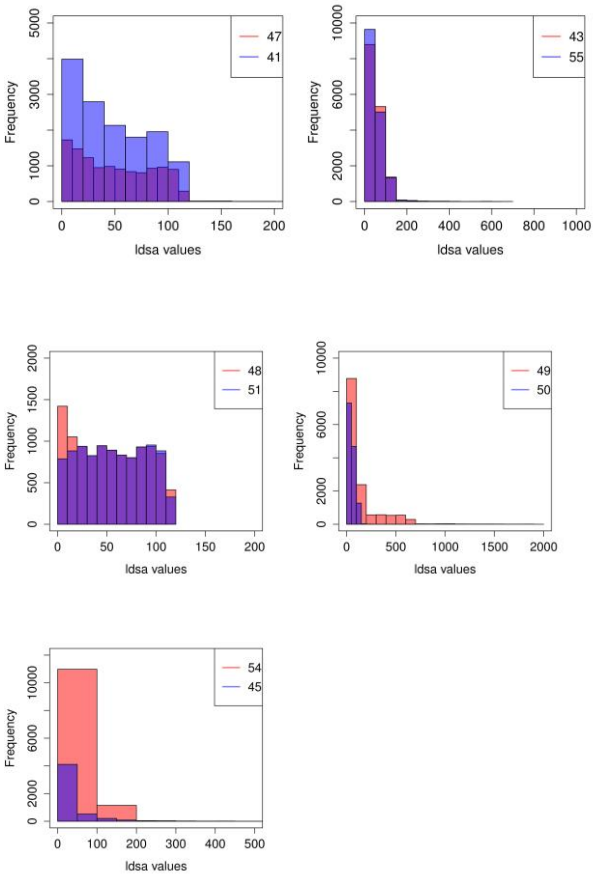
was applied over the data-set I in the following manner:

- Since Station attribute is a nominal attribute; Latitude, Longitude and LDSA attributes are independent of the corresponding Station data value (values are in Table I). Since data values of station attribute are independent of each other as station is a nominal attribute, data values of Latitude, Longitude and LDSA attributes for each station are also independent of the corresponding station data values. Each Station value is also independent of all other station value (as stated above). Thus, when CODB algorithm is applied to data-set I, detection of outliers is independent of the station value.
- Data values from to one station were paired with another station since LDSA, Latitude and Longitude attributes have almost similar numerical ranges for both of these stations. Station values paired together are shown in table II while corresponding LDSA, Latitude and Longitude values are shown in figure 1 and Table I respectively. Due to this pairing, CODB algorithm does not need to be applied individually to data values from each station. It could be individually applied to five pairs of stations Moreover, since decision to use class attribute as a nominal attribute is crucial to the application of CODB and also of Hoeffding tree ((VFDT) later in the paper), the data is involved in a central role in the analysis.

Table II: Paired Stations in Data-set I

Pair number	Station number 1	Station number 2
1	49	50
2	45	54
3	41	47
4	43	55
5	48	51

Figure 1: Distribution of LDSA attributes for pairs of stations for data-set I



- CODB algorithm has three components:

1. PCL (T, K)
2. Dev (T)
3. K-Dist (T)

Where T is the instance for which COF (T) (degree of being a class outlier) is evaluated and K is number of nearest neighbors. While evaluating Dev (T) and K-Dist (T), three numerical attributes within the data-set are used, namely, Latitude, Longitude and LDSA. The value of PCL (T, K) is a number between 0 and K and indicates nearest neighbors that belong to the same class (station attribute). CODB algorithm was applied over data values belonging to each of the five pairs of Stations in the manner described above. After running initial experiments on data-set I, maximum value of deviation (Dev (T))) was found to lie in thousands and maximum value of K-Dist (K-Dist (T)) was 0.0. Since maximum value of K-Dist was 0.0, corresponding parameter may be any arbitrary real number between 0 and 1 as suggested in [2]. Number of nearest neighbors (K) were set to 7 as suggested in [2]. The parameters with their corresponding values are shown in the table III below. Given the varying quality of measurements obtained from different

sensors used in the OpenSense project, 10% of the data values in data-set were regarded as outliers.

Table III: Parameters with their respective values

Parameters	Values
α	1000
β	0.1
k	7

IV Outlier Detection In Data Stream

For detecting outliers in the data stream (detailed description of the data stream may be found in data-set description), an algorithm should be chosen that should build machine learning models using a small sample of data and these machine learning models may be used for prediction. Further, the machine learning models should learn while carrying out prediction. Hoeffding tree (VFDT) provides a robust solution [3] for this requirement. Hoeffding tree was trained over data-set I and also over the outliers detected in data-set I. A separate Hoeffding tree algorithm was trained over five pair of stations shown in table (IV), therefore, five machine learning model were obtained. The efficiency of each model is shown in the table below.

Table IV: Efficiency for pairs of station

Pair number	Overall Efficiency
1	92.66%
2	96.21%
3	95.32%
4	95.30%
5	97.73%

Since minority class (outlier class) is 10% of data-set I while majority class (i.e. non- outlier class) is 90% of data-set I, precision and recall must be used to verify the accuracy of the five machine learning models. In order to do so, 10 fold cross validation is applied. Table (V) shows the precision and recall for each pair of stations.

Table V: Precision and Recall value for each pair

Pair number	Precision	Recall
1	87.38%	93.26%
2	81.85%	93.97%
3	83.90%	91.92%
4	92.66%	95.81%
5	84.35%	98.36%

Discussion

This discussion is about the generalizability of the approach. Since, Uniform random sampling was applied to the-data-set, the complete data-set need not be inspected for discovering patterns in the data-set. As described in the methodology for outlier detection section, each row in data-set is independent of all other rows due to treatment of station attribute. Therefore, relation among attributes of the data-set are identified as shown in methodology for outlier detection section, thus making the approach data-oriented. This approach is contrary to the approach adopted by the authors in [4].The authors adopt an algorithmic approach where a novel algorithm, kernel k-means is used for unsupervised clustering rather than the traditional k-means method.

Conclusion and Future work

The approach shown in this paper is specific to OpenSense data. In order to come up with generalized approaches, it is necessary to find common properties in data obtained from various sensor networks. These common properties could be used to design generalized approaches.

References

[1] Fabio Fumarola Donato Malerba Annalisa Appice, Anna Ciampi. Data Mining Techniques in Sensor Networks: Summarization, Interpolation and Surveillance. Springer-Verlag London, 2013.

[2] Nabil M Hewahi and Motaz K Saad. Class outliers mining: Distancebased approach. International Journal of Intelligent Technology, 2(1):55–68, 2007.

[3] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining timechanging data streams. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 97–106. ACM, 2001.

[4] Nripesh Trivedi, Daniel Adomako Asamoah, and Derek Doran. Keep the conversations going: engagement-based customer segmentation on online social service platforms. Information Systems Frontiers, pages 1–19, 2016