ISSN 2347 - 3983
Volume 7, No. 11 November 2019
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter297112019.pdf
https://doi.org/10.30534/ijeter/2019/297112019

Comparative Study for Outlier Detection In Air Quality Data Set

Devi Afriyantari Puspa Putri [1], Endah Sudarmilah [2]
[1] Informatics Department, Universitas Muhammadiyah Surakarta, Indonesia, dap129@ums.ac.id
[2] Informatics Department, Universitas Muhammadiyah Surakarta, Indonesia, Endah.Sudarmilah@ums.ac.id

ABSTRACT

Outlier detection considered as non-trivial task because it can be misclassified as noise, also sometimes it can be misclassified as normal data or otherwise. Difficulties level of detecting outliers arose in time series data, it happens because different context has different amount of outlier data set. One of type data which considered as time series is air quality data set, because level of air quality can be different over time. In order to overcome the difficulties of detects outlier in time series data set, this research compare three different techniques of outlier detection are : statistical, density ratio, and forecasting technique, using the same data set about air quality data set in Madrid in three year period. The result of this research concludes that density ratio has outperformed the other two techniques that obtain the highest balanced accuracy and the lowest percentage of FAD. The lowest percentage of FAD means that those technique have the ability to not misclassified outlier data set as normal data or otherwise.

Key words : Air Quality Data Set, Big Data, Density Ratio, Outlier Detection.

1. INTRODUCTION

In rapid development today, data which shared between various stakeholders coming from different sources. Those condition creates data set consists of various format and large in term of size, which can lead to the efforts which takes to process the data becoming non-trivial task. The origins of data which comes based on different types and formatting alongside with the quality which vary from one source and others, it can leads to the appearance of outliers in data set.

Data set can be considered contain outliers if one or more data has different behaviour compare with the dominance of data in data sets[1]. Detecting outliers in data set becoming non-trivial task because the appearance of noise which share similar behaviour with outliers can create confusion and misclassified of detection between them. However, according to[1] stated that the occurrences of outliers in data set lead the interest of analysist because outliers in data set often lead to hidden and useful information, while noise does not contain any useful information [2]. Therefore, their existence often ignored by analyst. In many cases, the evidence of the usefulness of the presence of outlier already proven in many fields. In health area, cancer can be detected using the examination of the presence of unusual cells activity that appeared in human body [3]. In traffic data area, the presence of outlier data in data set can be used to build a decision support system which gave users the recommendation of the best route to choose in order to prevent congestion [4]. In air quality control, outlier detection can be used to build automation measurement tools which useful to inform the citizen about the index of air quality in those area[5]. The existence of those tools can make citizen aware if harmful pollution occurs.

Because of the usefulness along with the difficult work for detecting outlier in large data, it leads many research concern to propose different technique in different areas of research. Those situation lead that outlier detection commonly known as collaborative research area, such as in statistical based, machine learning and data mining. That three techniques widely known as the main three approaches in outlier detection [6].

Even though many hidden and advantageous information can be extracted from the existence of outliers in data set, many studies stated that detecting outlier in data set become a challenging task because different technique of outlier detection does not have the same degree of success rate on different types of data set. [7]. According to that challenge, in order to attempt outlier detection successfully this research discusses three techniques which used in outlier detection in air quality control domain, are : statistical, density ratio, and forecasting technique.

This research concern about air quality in Madrid in three years period from 2016 to 2018, in station Parque del Retiro (28079049) in Comma Separated Values (CSV) format, the data set was taken from Kaggle Website [8].

2. BACKGROUND AND LITERATURE REVIEW

In this part discussed about the increasing demand of big data and the importance of data preparation followed by types outliers may occurred in data set, and labelling data set.

2.1 The Existence of Big Data

In today development, big data considered as hot topic which almost discussed everywhere in this world. Those phenomenon occurs because data which needed in most organisations required in big size of data and those data comes from many sources. The availability of data increase very significantly due to the ease of accessing data itself and the development of internet which become massively increase these days. According to [9] the large amount data which created today, comes from various devices and social network. Another experiment about big data suggested that [10] data has large amount of size effected by the increasing speed of collecting data set. The increasing amount of data also discussed in [11] which stated that in 2.5 quintillion bytes almost 90% of data were produced in the last two year. Based on the advantages which carried by managing big data set, many companies try to collect as many data as possible. Those behaviour causes people to take data based on their quantity over quality [12].

Possessing data without prior knowledge cause the data become unstructured data with different formatting levels, as discussed in [13] which stated that more than 80% of data exist today are unstructured data, while the rest known as structured data. Valuable information may be hindered from researchers if data set which possessed in company contains from unstructured data set. According to [14] stated that big data can be classified as 3V model, contains Volume, Velocity and Variety. Big data brings many valuable information that people can extract. However, in order to get high degree of success rate on outlier detection area, it is necessary to prepare data set before being processed into next step.

2.2 Data Preparation

In air quality data set, before information can be extracted, it is necessary to perform data preparation in initial step, as the preparation of data consider as important task. Based on [15] suggest that data preparation process belongs to important part to examine the data set because it will escalated the quality of data from unstructured data into semi-structured or structured data set. There are four steps which needed in order to prepare the data set, are : data transforming, data integration, data reduction, and data cleaning.

Data transforming is needed in the first step to prepare data, because as discussed above that large data set may contain from various format because it comes from various sources. Therefore, data need to be transform into one type of data or homogenous because existing machine cannot process data which have different type of formatting [16]. In the next process, data integration need to be performed to collect the data set which come from various sources. This process considered essential because it can improve the valuable information gathered. It followed by the proses of reducing data which aim to lessen the amount of data. Based on research of [17] concludes that some advantages can be obtained in the data reduction process. One of the advantages are, the processing of data set become faster and the speed to collect valuable information becoming shorter.

Data cleansing, become the next process needed. It is known that data which contain form unnecessary information likely being discarded in data set. The ordinary process of cleaning data set [18] contains of detects, repair, and remove data. In regard of task which used in extracting information form data set, it can be said that data preparation are the tidiest and challenging task in large data set.
According to [19] in early development, it suggests that only majority type of data which will be preserved in data set, while data that has minority part can be classified as noise and will be removed. However, nowadays it discovered that data set which has minority behaviour in their environment may contain interesting information than normal data. As discussed in part 1 that outlier contain useful information while noise does not have any important information. Therefore, noise in data set will be removed in cleansing part, while outlier remain stayed in data set. The discussion about an outlier will be presented in chapter 2.4

2.3 Air Quality Data Set

Based on [20] through analysing air quality data set, it can be used to examine the association between health level in human body and the level pollutant in particular area. Therefore, it is important to analysed the pollutant level in one area to refrain human body from disease an alert them. In general, the pattern of air quality data changing over time, the concentration level of pollutant during the day used to be higher compare in the night. Therefore, it is necessary to divide the data set in each hour, because the behaviour of air quality data set belongs to time series data set.

Time series data already drawn many interest of researchers because the characteristics owned by time series shows the various record and has more sensitivity based on time which differs from static data set that showed constant movement [21]. Several approaches already applied to get a good accuracy in time series data set. One of approach [22] stated that the segmentation in time consider as important task to be applied, because it creates the higher accuracy in data set. The process in dealing with time series data set should consider the change of data over time because it's the nature of time series data which always has changing point value according to time.

Therefore, in dealing with air quality data set in this research, it is best to fragment the data set in every hour. The segmentation of time should be performed because the behaviour of any elements which effect the level of pollutant

differ between hour. The data set in this research are using air quality data set in Madrid between 2017 and 2019 as stated in chapter 1.

There are 14 gaseous pollutant which collected in Madrid air quality data set, however, only NO_2 which used in this research. The threshold between normal and outlier in this data set has been set based on the Air Quality Index (AQI) level [23]. The threshold value of normal data set is below 361 part per billion (ppb) of the concentration of NO_2 which has the similarity about 150 in AQI which considered tolerable for insensitive groups as stated on [23]. Otherwise, the value of data set above 361 considered as an outlier because beyond that index the air quality becoming unhealthy for people's live.

The fragmentation in Air Quality data set only concerned about differentiation in time, while in weekend and weekdays share the similar distribution pattern of data which can be seen in figure 1. Therefore, in this research does not concern about day fragmentation.

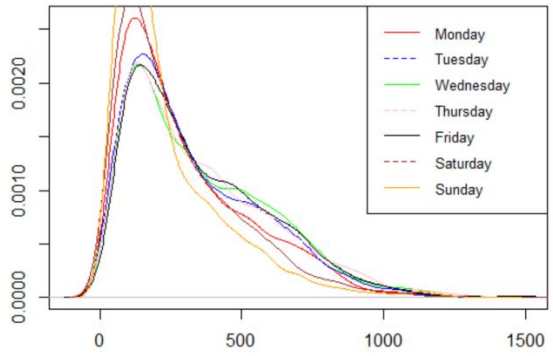

Figure 1: Distribution Pattern in Days

2.4 Outliers of Data Set

As discussed in chapter 1, outliers should be preserved in data set because it brings important information. Based on [24] stated that data set can be classified as an outlier, whenever data set acts differently compare to the others data set that has major occurrences, and the distance of that data occurs far away compare to normal data. The illustration of outliers data in data set can be seen in figure 2 [1].

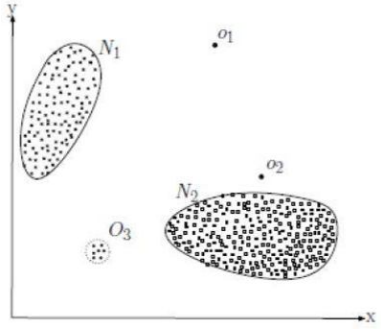

Figure 2: Outlier appearance in data set
(Source : [1])

According in figure 2 it can be seen that data consist of two normal groups and three outliers groups distribution. The outliers groups are O_1, O_2, and O_3 because the appearance of them considered as minority compare to two other groups and their location is far away compare to N_1 and N_2.

In general, there are three major types of outliers data are : point , contextual, and collective outliers. Fraud detection, and cancer detection can be classified as point outliers [6]. Data set can be considered as contextual outlier depends on the situation, the illustration can be seen in figure 3[6].

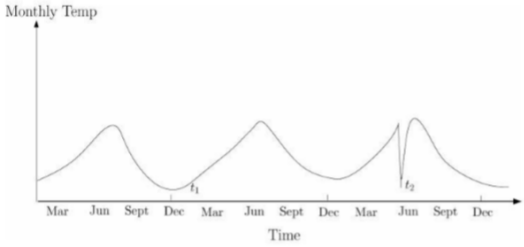

Figure 3: Contextual Outlier in Data Set
(Source : [6])

Based on figure 3, condition in t_1 considered as normal data set because that temperature occurred in winter, while in t_1 considered as normal data set because that temperature occurred in winter, while in t_2 with the same degree of temperature considered as an outlier in data set because it happens during summer, and the neighbourhood on those data set does not shared the similarity. Contextual outliers in data set typically occurs in time-series data set. While collective outlier determine the outliers data based on time-period [6] which typically occurs in stock market.

Based on the explanation above and the observation of data set in air quality data set, this research can be considered into contextual outliers because its occurrences has heavily dependent into the time of data set. The example of different distribution on Madrid air quality data set presented in figure 4.

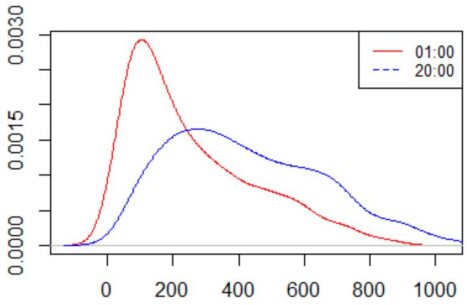

Figure 4: Distribution Pattern based on hour

2.5 Labelling Data Set

In outlier detection context, the availability of labelling data set also takes essential part to determine which techniques

need to be applied in detecting outliers data set. Three types of methods has been known in machine learning to work with data set, are : unsupervised, semi-supervised and supervised learning.

Unsupervised Learning assumes that data set in real words does not have any labels on it. It does not divide data into training and testing data set. In order to detect outlier in data set which does not have any labels, it used assumptions that data gather together within close distance considered as normal data. Some techniques which used this learning behaviour are clustering approach and distance approach.

Semi-supervised has combine the techniques between supervised and unsupervised learning [24]. This techniques used only one label of data as training data which is normal data. This learning approach assumed data that have far differences compare to training data can be concluded as outliers.

In supervised learning, data can used this learning method if the availability of data consists of two classes data which is outlier and normal data set. Similar with semi-supervised learning, this method needs training and testing data to perform outlier detection. However, there are major issue in supervised learning regarding the outlier detection. It is because the occurrences of imbalanced data set because labelling data in outlier class can be considered as non-trivial task since it does not occur as many as normal class [25].

Based on the discussion of the learning method available in outlier detection, this research will mainly perform techniques which has approached in unsupervised, and semi-supervised learning because the appearances of outlier in data set in real world become too wide to define. The label data can be obtained using the threshold on AQI index level [23], and already discussed in part 2.3.

3. RELATED WORK

Many research and methods have addressed in this area to handle the problem which happens in detection of outlier in time series data. According to [26] outlier detection in time series data is non-trivial task because it needs to handle many aspects, especially the condition in specific time-frame. Those research used local outlier approach, and stated that there are several types of approach which can be used to detects anomaly on local outlier, the research discuss the limitation of each approach.

In windows based approach, it stated that in this approach window size in data set need to determine carefully since it is the most important items which contributed in the percentage of accuracy obtain in the final measurement. The narrower the window size the higher degree of accuracy obtained. Besides that, the time computation using windows based approach turned out to be expensive and require longer process compare to others. Another limitation found are if the distance from normal data and centre point has same value with outlier data and centre point, it can be mistakenly detect outlier as normal data or otherwise.

Another approach using Euclidean distance in density approach turns out to have the best degree of accuracy among others methods used. Meanwhile, prediction based using Autoregressive in moving average (ARIMA) or autoregressive (AR) is critical to any shifting point in the middle of data set. It also considered oversensitive to detect outliers so the result using this approaches have low percentage in accuracy.

Other research about traffic data conduct [27] that also used windows size of time, in regards to achieve greater accuracy, it is necessary to determine windows size very carefully. Another approach in outlier detection proposed by [28] in detecting outlier in stock market using three steps which applied. The initial step to perform is create the model in data set, in the research it used density function to generated data. It is also make an assumption that data stayed in gaussian model. The next step are build a model in time series data using AR calculation, and finally it will give a score for every data to decide whether data belongs to outlier or normal data. Other techniques [29] in traffic data need data transformation to handle the limitation of wrongly detects data into normal and outlier data set.

Based on the previous researches using different techniques, this research compared three different techniques, are: statistical, density ratio, and forecasting technique. In this research data set fragmented in an hour window time, without differentiation of days as discussed in part 2.3. The further explanation about three chosen techniques described in chapter 4.

4. ALGORITHM FOR OUTLIER DETECTION

4.1 Statistical Technique

In order to perform statistical technique this research use adjusted boxplot technique. The underlying reason regarding the chosen technique because adjusted boxplot does not assume data has normal distribution which become the limitation of original boxplot proposed by Tukey [30]. Adjusted boxplot considered better choice compare to original boxplot because data in real word does not always contains normal distribution[31]. In air quality data set used in this research does not have normal distribution as seen in figure 2.1. Therefore, it is suggested to use adjusted boxplot in predicting outlier data set. This technique belongs to unsupervised learning because no assumption has been made

regarding prior data set. In order to detects outliers adjusted boxplot using fence with two conditions, are :

- Medcouple > 0 use range "$Q3 + (1.5\ e3\ MC * IQR)$ and $Q1 - (1.5\ e{-}4\ MC * IQR)$"
- Medcouple < 0 use range "$Q3 + (1.5\ e4\ MC * IQR)$ and $Q1 - (1.5\ e{-}3 MC * IQR)$"

Every data set which lies beyond the fence considered as outlier data set, while data set inside fence classified as normal data. A package called "robustbase" has been used in R Software to implement adjusted boxplot. Data preparation using this statistical techniques illustrated in table 1. The design of package and workflow in statistical technique from data set to evaluation measurement can be seen in figure 5.

4.2 Density Ratio Technique

Technique used density ratio approach based on paper [32], which called densratio afterward. Densratio technique can be considered as semi-supervised learning, because it only used one labelled data as training data set as discussed in chapter 2.5. According to [6] semi-supervised learning proven to be more applicable in detecting outliers in data set compare to other methods. It is happen because in real word data, normal distribution can be found easily while outlier data set rarely found, and it can leads to imbalance data set.

In this method, to detect outlier data in every data set, density ratio estimation has been made. This technique, works by comparing the ration density between training and testing data. The formula used for densratio technique written below :

$$w(x) = \frac{Ptr\ (x)}{Pte\ (x)} \qquad (1)$$

Based on the equation (1), it can be explained that the estimation ratio of data set ($w(x)$) can be obtained by divide the density of training data set ($Ptr\ (x)$) and testing data set ($Pte\ (x)$). The data point considered as an outliers data set if the ratio of density has low value or closed to zero while normal data has high value of density usually closed to one.

In order to perform ratio estimation using densratio technique used statistical approach called "unconstrained Least-Square Importance Fitting" (uLSIF) [32]. uLSIF has chosen because that algorithm provide many variation of cross-validation (CV). The advantage of variety CV has been applied to determine the best model used in make the data prediction on testing data. As stated above, this density ratio technique already assume that only normal data that can be classified as training data set, while mix data set (normal and outlier) considered as testing data set.

Densratio technique implemented in R software used package called "densratio"[33]. Using this technique, it is necessary to split the data into training and testing data set. In order to create the model selection in density ratio technique, every data points in training data set assume as the centre of data set. because all data in training data assumes as normal data. Based on that condition, it can assumes that whenever data points in testing data has large deviation with the centre can be classified as outlier data. That is happen because the result between division of training data and testing data which has greater distance produced small estimation. Distance calculation in densratio technique based on Euclidean distance from centres (training data set) and testing data set. The data preparation using density ratio technique illustrated in table 1. The design of package and workflow in this technique from data set to evaluation measurement can be seen in figure 5.

Table 1: Data Preparation of Each Technique

Requirement	Technique		
	Adjusted Boxplot	**Densratio**	**Forecating**
Separate data into training and testing data	-	+	-
Extract day,hour, and adding GS	+	+	+
Data set formatting	+	+	+
Data set Cleansing	+	+	+
Fragment the data set	+	+	+
Change type of data	-	-	+
Labelling data set	+	+	+

4.3 Forecasting Technique

This technique has similar approach with adjusted boxplot that concludes as unsupervised learning, because it does not need to divide data set into training and testing data. The underlying reason this technique chosen because it can be used to predict the occurrence of data in the future. This technique assumed data set as an outlier if the data has big difference from the previous data set. In performing outlier detection using forecasting technique in R environment used package called "tsoutlier" [34]. Forecasting technique in tsoutlier package already applied ARIMA model and combine the AR model. The data preparation using this techniques illustrated in table 1. The design of package and workflow in forecasting technique from data set to evaluation measurement can be seen in figure 5.

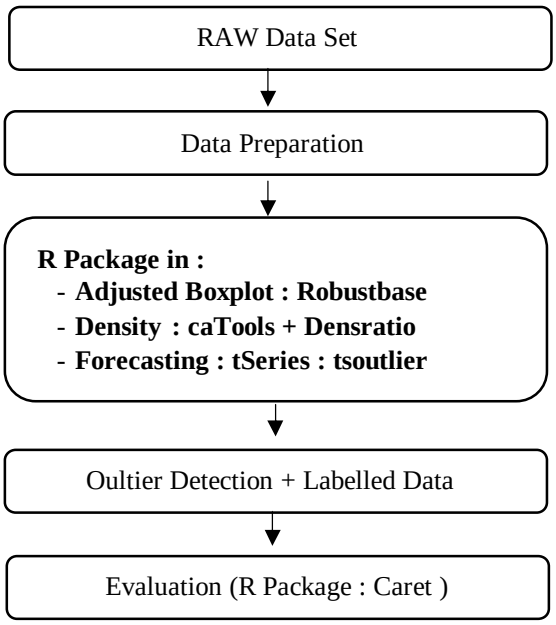

Figure 5: Workflow and Packages used in R for each technique

5. RESULTS OF OUTLIER DETECTION

This chapter discussed about result and measurement used to calculate the accuracy based on selected technique which discussed in chapter 4.

5.1 Evaluation Measurement

The ability of each technique to detects outlier in data set presented in accuracy. The accuracy of data set obtained from comparing GS and prediction yield from each technique. According to [35] it is essential to detect outlier as real outlier. Therefore, it is considered to obtain high degree accuracy in rate detection and low accuracy in false alarm detection (FAD).

In order to detects FAD, it is important to used false positive rate formula which described in (2) equation. In order to handle imbalanced data set, balanced accuracy (BA) need to be performed in order to measure accuracy. Based on [36], BA important to yield unbiased result in imbalanced data set. The formula of BA presented in equation (3).

$$false\ positive\ rate : \frac{Negatives\ incorrectly\ classified}{Total\ Negatives} = \frac{FP}{FP + TN} \quad (2)$$

$$balAcc : \frac{\sum_{i=1}^{n} \frac{sensitivity + specificity}{2}}{n} \quad (3)$$

5.2 Result in Adjusted Boxplot

The result and measurement of adjusted boxplot in order to detects outlier in air quality data set presented in this chapter.

In this research detect outlier data during days in one hour time-periods using NO_2 value. Results of outlier detection depicted in figure 6.

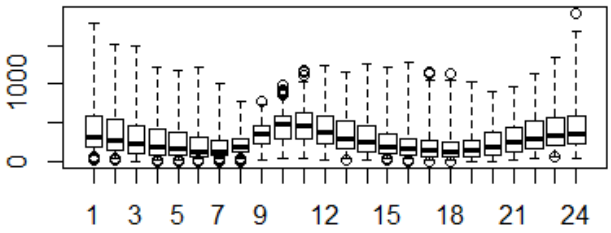

Figure 6: Result of Adjusted Boxplot in Each Hour

According to figure 6, it can be seen that data lies outside fence described as a circle and considered as outlier in data set. It can be seen that at 09:00 there is one data set considered as outlier.

5.3 Result in Density Ratio

In this chapter described the result of outlier detection in densratio technique which described in figure 7. In this research the setting of threshold set in 0.5, it means that data set that has value higher than 0.5 determined as normal data set, while below 0.5 classified as outlier. In densratio the subjectivity of outlier quite high because the setting of threshold highly depend on user point of view.

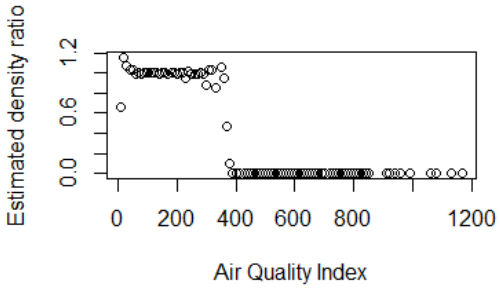

Figure 7: Densratio Technique at 05:00 A.M

5.4 Result in Forecasting Technique

In forecasting technique, outlier data set need to manually labelled since the output of this techniques represented in diagram and whichever the outlier comes out it highlighted in thick red dot. The illustration of this technique can be seen in figure 8. It can be seen in figure 8, there are three outliers lies in data set. The type of three outlier which occurred described in figure 9.

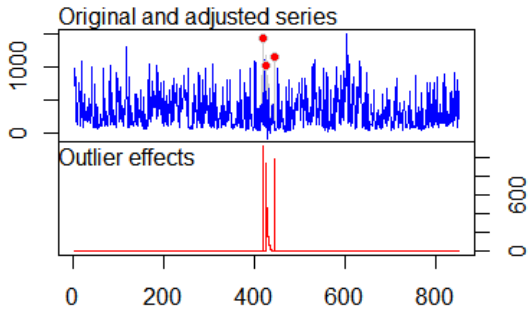

Figure 8 : Outlier detection at 03:00 A.M in forecasting

```
Outliers:
  type ind time coefhat tstat
1   AO 417  417  1123.3 4.682
2   TC 423  423   944.0 4.415
3   AO 443  443   995.6 4.151
```

Figure 9 : List of outlier detection at 03:00 A.M

Based on figure 8 forecasting technique able to detect outlier because there are huge differences value between the outlier data set and previous data set.

5.5 Comparison of Each Technique

The measurement of comparison data between data set and GS based on FAD and BA depicted in figure 10. It can be seen that densratio technique achieved the best technique to detects outlier in air quality data set compare to other techniques. It can be concludes because densratio achieved highest rate of BA and lowest percentage of FAD. which mean it has the lowest probability to misclassified normal data set as outlier data set or vice versa.

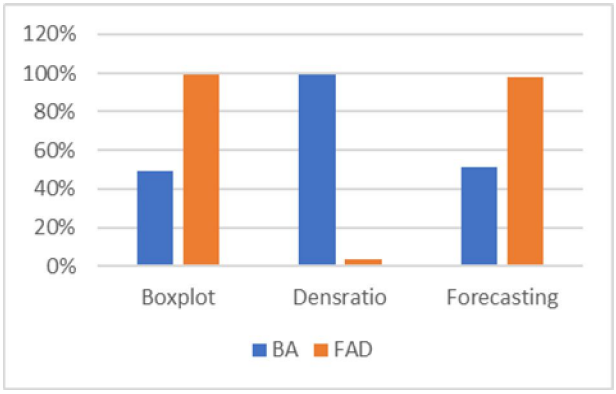

Figure 10 : Comparison of each technique

6. CONCLUSION

According to the result described in chapter five, it can be seen that density ratio has the best accuracy and the lowest probability to misclassified the data set compare to two other techniques. Such result can be achieved because in air quality data set in Madrid which collected through website [8] the occurrence of outlier in that data set has high percentage of occurrence compare to normal data set. The two techniques are adjusted boxplot and forecasting technique failed to detect the outliers because it based on unsupervised learning, which in their examination heavily depend on the occurrence of majority data set.

Based on that case, if the majority of data set belongs to outlier data, it can labelled those data as normal data, even though it is outlier data set. Therefore, based on the experiment it can concludes that unsupervised learning can work best if the outlier in data set occurs in rare frequency. On the other hand, density ratio can work best in detecting outliers because it based on semi-supervised learning which the data in training set only consists of normal data set, therefore if the occurrence of data has big deviation compare to training data set considered as outlier. Even though the occurrence of outlier in data set has big frequency, it does not effects the performance of the algorithm. Based on the experiment it can concludes that density ratio technique which based on semi-supervised learning perform better compare to unsupervised learning, event in data set which consist of outlier data set in majority.

ACKNOWLEDGEMENT

This research was funded by Indonesia Endowment Fund for Education and as a part of master dissertation work which submitted in The University of Manchester under School of Computer in 2017 using different data set.

REFERENCES

[1] Singh, K. and Upadhyaya, S. **Outlier detection: applications and techniques**, *International Journal of Computer Science Issues*, vol. 9(1), pp.307-323. 2012.

[2] Ben-Gal, I. **Outlier detection**, *Data mining and knowledge discovery handbook*, pp.131-146, 2005. https://doi.org/10.1007/0-387-25465-X_7

[3] Alshalalfa, M., Bismar, T.A. and Alhajj, R. **Detecting cancer outlier genes with potential rearrangement using gene expression data and biological networks**, *Advances in bioinformatics*, 2012 https://doi.org/10.1155/2012/373506

[4] Pham, T.L., Germano, S., Mileo, A., Küemper, D. and Ali, M.I. **Automatic configuration of smart city applications for user-centric decision support**, *In Innovations in Clouds, Internet and Networks (ICIN)*, 2017 20th Conference on (pp. 360-365). IEEE. 2017 https://doi.org/10.1109/ICIN.2017.7899441

[5] Torres, J. M., Nieto, P. G., Alejano, L., & Reyes, A. N. **Detection of outliers in gas emissions from urban areas using functional data analysis**. *Journal of hazardous materials*, 186(1), 144-149, 2011. https://doi.org/10.1016/j.jhazmat.2010.10.091

[6] Chandola, V., Banerjee, A. and Kumar, V. **Anomaly detection: A survey**, *ACM computing surveys (CSUR)*, 41(3), p.15, 2009.

https://doi.org/10.1145/1541880.1541882

[7] Markou, M. and Singh, S. **Novelty detection: a review—part 1: statistical approaches**, *Signal processing*, 83(12), pp.2481-2497, 2003. https://doi.org/10.1016/j.sigpro.2003.07.018

[8] Soluciones, D. **Air Quality in Madrid**, Version 5, 2019. from https://www.kaggle.com/decide-soluciones/air-quality-madrid.

[9] Erraissi, A. and Belangour, **A. Meta-modeling of Big Data management layer**, *International Journal of Emerging Trends in Engineering Research*, Vol.7, No.7, pp. 36-43, 2019. https://doi.org/10.30534/ijeter/2019/01772019

[10] Kumar, N,P., Sastry, J., Rao, K,R,S. **Mining Negative Frequent Regular Itemsets from Data Streams**, *International Journal of Emerging Trends in Engineering Research*, Vol.7, No.8, pp. 85-98, 2019. https://doi.org/10.30534/ijeter/2019/02782019

[11] Dobre, C. and Xhafa, F. **Intelligent services for big data science**, *Future Generation Computer Systems*, 37, pp.267-281, 2014. https://doi.org/10.1016/j.future.2013.07.014

[12] Kaisler, S., Armour, F., Espinosa, J.A. and Money, W. **Big data: Issues and challenges moving forward**, *In System Sciences (HICSS)*, 2013 46th Hawaii International Conference on (pp. 995-1004). IEEE, 2013. https://doi.org/10.1109/HICSS.2013.645

[13] Das, T.K. and Kumar, P.M. **Big data analytics: A framework for unstructured data analysis**, *International Journal of Engineering Science & Technology*, 5(1), p.153, 2013.

[14] Kulkarni P., Akhilesh K.B. **Big Data Analytics as an Enabler in Smart Governance for the Future Smart Cities**, *In: Akhilesh K., Möller D. (eds) Smart Technologies*. Springer, Singapore, 2020.

[15] Zhang, S., Zhang, C. and Yang, Q. **Data preparation for data mining**, *Applied Artificial Intelligence*, 17(5-6), pp.375-381, 2003. https://doi.org/10.1080/713827180

[16]Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R. and Shahabi, C. **Big data and its technical challenges**, *Communications of the ACM*, 57(7), pp.86-94, 2014. https://doi.org/10.1145/2611567

[17] Aggarwal, C.C.. **Data mining: the textbook**. *Springer*, 2015.

[18] Rahm, E. and Do, H.H. **Data cleaning: Problems and current approaches**, *IEEE Data Eng. Bull.*, 23(4), pp.3-13, 2000.

[19] Jiang, S.Y. and An, Q.B. **Clustering-based outlier detection method**, *In Fuzzy Systems and Knowledge Discovery*, Fifth International Conference on Vol. 2, pp. 429-433, 2008. https://doi.org/10.1109/FSKD.2008.244

[20] Venkatram, A., Isakov, V., Thoma, E., & Baldauf, R. **Analysis of air quality data near roadways using a dispersion model**. *Atmospheric Environment*, *41*(40), 9481-9497, 2007.

[21] Liao, T.W. **Clustering of time series data—a survey**, *Pattern recognition*, 38(11), pp.1857-1874, 2005. https://doi.org/10.1016/j.patcog.2005.01.025

[22] Esling, P. and Agon, C. **Time-series data mining**, *ACM Computing Surveys (CSUR)*, 45(1), p.12, 2012. https://doi.org/10.1145/2379776.2379788

[23] AirNow.. **Air Quality Guide for Nitrogen Dioxide**, 2016.

[24] Gao, J., Cheng, H. and Tan, P.N. **Semi-supervised outlier detection**, *In Proceedings of the 2006 ACM symposium on Applied computing*, pp. 635636, 2006.

[25] Chawla, N.V., Japkowicz, N. and Kotcz, A. **Special issue on learning from imbalanced data sets**, *ACM Sigkdd Explorations Newsletter*, 6(1), pp.1-6, 2004.

[26] Golmohammadi, K. and Zaiane, O.R. **Time series contextual anomaly detection for detecting market manipulation in stock market**, *In Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference*, pp. 1-10, 2015.

[27] Li, X., Li, Z., Han, J. and Lee, J.G. **Temporal outlier detection in vehicle traffic data**, *In Data Engineering, 2009. ICDE'09. IEEE 25th International Conference*, pp. 1319-1322, 2009. https://doi.org/10.1109/ICDE.2009.230

[28] Yamanishi, K. and Takeuchi, J.I. **A unifying framework for detecting outliers and change points from non-stationary time series data**, *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 676-681, 2002.

[29] Ngan, H.Y., Yung, N.H. and Yeh, A.G. **Outlier detection in traffic data based on the Dirichlet process mixture model**, *IET intelligent transport systems*, 9(7), pp.773-781, 2015.

[30]Tukey, J.W. **Exploratory Data Analysis**, *Addison-Wesley, Reading, Massachusetts*, pp. 39–49, 1977.

[31] Hubert, M. and Vandervieren, E. **An adjusted boxplot for skewed distributions**, *Computational statistics & data analysis*, 52(12), pp.5186-5201, 2008. https://doi.org/10.1016/j.csda.2007.11.008

[32] Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M. and Kanamori, T. **Inlierbased outlier detection via direct density ratio estimation**, *In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference*, pp. 223-232, 2008. https://doi.org/10.1109/ICDM.2008.49

[33] Makiyama, Koji. **densratio: Density Ratio Estimation version 0.0.3**, 2016.

[34] de Lacalle, J.L. **tsoutliers: Detection of Outliers in Time Series**, r package version 0.6, 2015.

[35] Zhang, Y., Meratnia, N. and Havinga, P. **Outlier detection techniques for wireless sensor networks: A**

survey, *IEEE Communications Surveys & Tutorials*, 12(2), pp.159-170, 2010. https://doi.org/10.1109/SURV.2010.021510.00088

[36] Garcıa, V., Mollineda, R.A. and Sánchez, J.S. **Index of balanced accuracy: A performance measure for skewed class distributions**, *4th IbPRIA*, pp.441-448, 2009. https://doi.org/10.1007/978-3-642-02172-5_57