Fuzzy Multivariate Outliers with Application on BACON Algorithm

Huda Mohammed Touny
Computer Science
Faculty of Computers and Artificial Intelligence
Cairo, Egypt
hodam.touny@gmail.com

Ahmed Shawky Moussa
Computer Science
Faculty of Computers and Artificial Intelligence
Cairo, Egypt
a.moussa@fci-cu.edu.eg

Ali S.Hadi
Department of Mathematics and Actuarial Science
American University in Cairo
Cairo, Egypt
ahadi@aucegypt.edu

***Abstract*—Depending on a crisp cut-off value to identify outliers is not linguistically meaningful or insightful for reliable decision-making. In this paper, two methods of fuzzy treatment for the Blocked Adaptive Computationally-efficient Outlier Nominator (BACON) algorithm are proposed rather than a crisp cut-off threshold. Experimentation has shown that the proposed fuzzy treatments for BACON provide more meaningful interpretations to the final results than its crisp version and captured the uncertainty at the boundary between the inliers and outliers of the data.**

Keywords—Fuzzy Outliers, Multivariate Outliers, Outlier Detection, BACON, Mahalanobis Distance

I. Introduction

When talking about data, we can mainly address two major facets to describe and model its behavior. Inlier is the first facet which is defined as the "data value falling within the expected range." [1] On the opposite view, an anomaly or an outlier is defined as "Data value falling outside the expected range" [1] which is the facet that helps in data modeling. Authors in the literature have proposed many definitions for an outlier with seemingly no universally accepted definition [2]. Accordingly, Outlier Detection is defined as the process of identifying anomalies in a dataset. Outlier Detection has many synonyms based on the use-case or the application context. For example, in Network Analysis it is called Behavioral Analysis [3]. It can also be addressed with abnormality detection in the context of surveillance cameras. Other synonyms include novelty detection, intrusion detection, noise detection, deviation detection or exception mining [2]. Outlier detection is a critical task for various application domains and has been researched intensively for a long time. The process of outlier detection represents a challenge as it is difficult to accurately define and quantify. Another challenge lies in the customization of outlier detection to the corresponding domain. Thus, many techniques have been introduced for outlier detection, yet they do suffer drawbacks such as labeling a datum that is close to the separating boundary between normal and outlying behavior. Hence, fuzzy based outlier detection techniques were introduced. Most of such techniques mainly depend on fuzzy clustering. The Fuzzy C-Means (FCM) algorithm is the most used [4]. Unlike the conventional crisp clustering methods which restrict each point of the data set to only belong to exactly one cluster, fuzzy clustering techniques rely on the fuzzy set theory, which was proposed by Zadeh [5] in 1965. The theory introduced the idea of uncertainty of a data point belonging to a cluster which was defined by a membership function. There are other fuzzy based outlier detection techniques that may rely on fuzzy association rules or fuzzy reasoning [6]. Other hybrid techniques combine more than one technique to overcome the drawbacks of each technique individually [7].

The difficulty of outlier detection becomes much harder in higher dimensionality than in one dimension. In this context, the direction of an observation is a critical factor to declare an observation as outlier. For example, an outlying observation may lie close to the majority of the data but deviates from the overall data correlation structure. Thus, the Euclidean distance is not a dependable metric. As a consequence, Fuzzy C-Means is not suitable in the situation of experimenting high dimensional data. In [8] Winkler et al. conducted an extensive analysis of FCM in high dimensional data to highlight this weakness. The Mahalanobis distance is a distance metric which is used in the case of analyzing two or more dimensions. An efficient distance-based algorithm to identify outliers in large and multivariate data was proposed in [9] as Blocked Adaptive Computationally-efficient Outlier Nominator (BACON). BACON mainly depends on a statistical hypothesis test value to declare an observation as outlier. The aim of this paper is to propose two fuzzification approaches to the used crisp cut-off threshold in BACON to attain more meaningful notion of outliers.

The paper is organized as follows: Section II briefly discusses the BACON algorithm. Section III presents the related work. Section IV explains both of the proposed fuzzification approaches of BACON. Section V explains the used datasets, whereas Section VI presents the experimental results and Section VII presents a summarized conclusion of the paper.

II. BACON ALGORITHM

BACON is an iterative distance-based algorithm which was proposed to efficiently identify outliers in large and multivariate data [9]. BACON initially starts with a basic subset of data points which are assumed to be outlier-free, and then it iteratively adds points that have a small Mahalanobis distance to that clean subset. BACON stabilizes when iterations no more change the size of the clean subset. Outliers then are defined to be the points which reside outside of the clean subset. Therefore, it is theoretically possible to consider BACON as a clustering algorithm with the number of clusters being only two. The steps of BACON are listed below:

Step 1: Identify an initial basic subset of $m > p$ observations that can be presumed to be outliers-free, where p is the dimension of the data and m is an integer chosen by the expert. Initial subset is chosen based on one of two methods: Mahalanobis distance given in (1) or Distance from the vector of coordinate-wise median given in (2).

$$d_i\,(\bar{x}, \mathrm{S}) = \sqrt{(x_i - \bar{x})^{\mathrm{T}} \mathrm{S}^{-1} (x_i - \bar{x})}\,, i = 1,..,n, \quad (1)$$

$$\|x_i - \mathcal{M}\|, \quad (2)$$

where $\bar{x}$, and S are the mean and covariance matrix of the n observations and $\mathcal{M}$ is the coordinate-wise median of the observations.

Step 2: Compute discrepancies for all observations as given in (3).

$$d_i\,(\bar{x}_{\mathrm{b}}\,, \mathrm{S}_{\mathrm{b}}) = \sqrt{(x_i - \bar{x}_{\mathrm{b}})^{\mathrm{T}} {\mathrm{S}_{\mathrm{b}}}^{-1} (x_i - \bar{x}_{\mathrm{b}})}\,, i = 1,..n, \quad (3)$$

where $\bar{x}_{\mathrm{b}}$ and S_{b} are the mean-vector and the covariance matrix of the observations in the basic subset.

Step 3: Form a larger basic subset consisting of observations with the smallest discrepancies. This new basic subset may exclude some of the previously included observations, but it must be at least as large as the previous basic subset.

Step 4: Repeat Steps 2 and 3 until the basic subset can no longer grow safely.

Step5: Declare the observations excluded by the final basic subset as outliers.

III. RELATED WORK

Outlier detection has been intensively investigated for a long time in various substantive areas. With the emergence of fuzzy set theory, more studies were introduced to demonstrate the significant role of fuzzy logic in outlier detection. Moreover, researches carried out survey studies of fuzzy logic-based methods to investigate their importance in outlier detection [10]. In [11] Kim et al. presented a fuzzy logic-based outlier detection framework that is applied at the data entry phase for high dimensional cumulative biomedical database. The proposed framework is capable of identifying an outlier data record by testing the vertical and horizontal consistency of the sought records' attributes' values. The vertical consistency concerns about the deviations in the attribute values in person-wise follow-up data, while the horizontal consistency concerns about the deviation of an attribute value from its domain over all the data records. All attributes of the record are assigned an outlierness degree determined by a trapezoid fuzzy membership function. The proposed framework offers a flexibility to tune the behavior of the outlier detection process depending on a changeable cut-off threshold for the degree of outlierness, which in turn affects the outlier detection rate. Experiments were performed on artificial datasets and some data sets were modified by random noise to get outliers. Results showed that the proposed method can detect meaningful outliers at data entry stage. For further work, the authors aim to detect more complicated outliers where attributes seem normal individually but their combination is anomalous.

Cateni, Colla and Vannucci [12] discussed how industrial data exhibit high dimensionality and require conflicting computation requirements as time vs. efficiency. This tradeoff is mainly highlighted in the paper by proposing a Fuzzy Inference System (FIS) that is able to detect anomalies by firstly combining several outlier detection methods, which mainly depend on clustering and neighbor points analysis. The proposed FIS is then fed by these four features as an input and outputs an index within the interval [0, 1] as a risk indicator that the observed pattern is an outlier. The performance of the proposed fuzzy method was compared to Grubbs [13], the comparison was mainly subject to the percentage of the detected outliers for each data variable. The results showed that the proposed fuzzy method outperformed Grubbs by recognizing 100% of the outliers in most of the cases without priori assumptions of the data distribution, yet its computation time is almost 10 times larger than Grubbs. Another drawback of the proposed fuzzy method lies in the assumption that the present outlier must deviate markedly from the center of data as to have a very low membership to the possible cluster to which it could be assigned. Heuristically tuning the parameters of the membership functions is also considered as a drawback.

The weakness of depending on a binary inclusion or exclusion criterion to declare an observation as outlier is discussed in [14]. The authors highlighted how the classical tests as, in Grubs and Dixon [13, 15] are unstable in detecting outliers as they produce contradictory conclusions under the same circumstances. The study suggested a less crisp decision criterion by proposing a novel treatment of outliers based on fuzzy logic. The proposed treatment depends on 2-input/1-output fuzzy inference engine. The inference engine then maps the inputs to an outlierness degree to be high, intermediate or low. The performance of the proposed system was tested through simulated datasets indicating the efficiency in assigning candidate outliers an outlying degree rather than eliminating them.

A generic fuzzy treatment method for outlier detection was introduced in [7] with an application on penicillin production data. The proposed method depends on a 2-input/1-output fuzzy inference system. The Hampel [16] identifier along with the FCM membership values are fed to the fuzzy inference system as inputs. The outlierness degree of an observation is then computed as an output. Moreover, authors introduced a comparative study with four individual methods. Results indicated that the proposed method provides a satisfactory decrease in the number of false positives and false negatives compared to each method individually.

IV. Proposed Solution

The main aim of this study is to both propose and develop two fuzzy treatment approaches to the candidate outliers obtained by BACON. As mentioned in Section II, BACON is an iterative algorithm. In Step 3, BACON performs a hypothesis test to decide whether a point should be included in the new basic subset. BACON depends on an iteration-based critical value in its hypothesis test which is $C_{npr}\,\chi^2_{p,\propto/n}$, where χ^2 is the $1-\propto$ percentile of the chi square distribution with $\mathcal{P}$ degrees of freedom, $C_{npr} = C_{np} + C_{hr}$ and

$$C_{hr} = max\{0, \left((h - r)/(h + r)\right)\},$$

$h = [(n + \mathcal{P} + 1)/2]$, n is the size of the dataset, $\mathcal{P}$ is the dimension of the dataset, r is the size of the basic subset, and

$$C_{np} = \frac{\mathcal{P}+1}{n-\mathcal{P}} + \frac{1}{n-h-\mathcal{P}} + 1 = 1 + \frac{\mathcal{P}+1}{n-\mathcal{P}} + \frac{2}{n-1-3\mathcal{P}} \qquad (4).$$

The idea of the proposed solution mainly depends on making this cut-off threshold a fuzzy number. Thus, the end decision of a data point should be more linguistically interpretable rather than being binary as an absolute outlier or inlier. This means that a point could be both outlier and inlier but with different belonging degrees to both classes. Therefore, and according to the steps of BACON mentioned in Section II, a point with a discrepancy relatively close to the cut-off threshold will have a more belonging degree to be an outlier than a point with a farther discrepancy from the cut-off threshold as shown in Fig. 1. This in turn will have an impact on the decision making process regarding the risk interpretation of each point with respect to its overlapped belonging degrees in both inlier and outlier classes.

As Fig. 1 shows, in the Mahalanobis distance of Mtcars dataset, BACON will consider observation 33 and 34 absolute outliers with equal outlying degree even though it is obvious how farther observation 34 than observation 33 from the cut-off threshold represented by a red line in Fig. 1. BACON will also consider observations 18, 20, and 25 as absolute inliers despite being relatively close to the cut-off threshold. The proposed solution shall assign two membership degrees to those extreme points with respect to having two classes as inlier and outlier.

The two proposed fuzzy treatment approaches mainly depend on the crisp cut-off threshold, which BACON calculates in the last iteration. In both approaches, the distance domain is used instead of the original data domain. Accordingly, all the used calculations strictly depend on the corresponding distance value for all data points calculated by BACON. Both approaches then consider setting a left and right tolerance limits on the crisp cut-off to form a fuzzy number and two unbalanced classes as presented in Fig. 2.

The two proposed approaches are:

A. *Coefficient of Variation (CoV)*

In this approach, both left and right tolerance limits are set using the coefficient of variation, which is calculated as the standard deviation divided by the mean. The idea behind using coefficient of variation relies on its ability to give a sense of how uncertain the data is, given an expected value, which is the mean in this case.

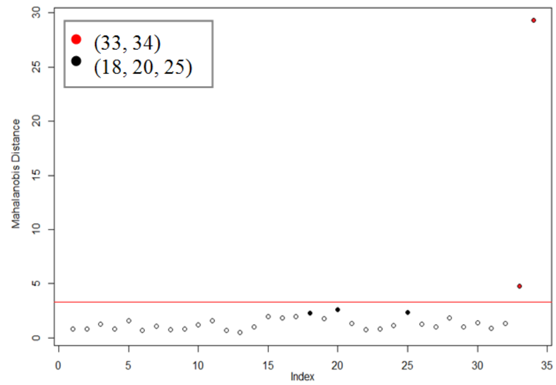

Fig. 1. Mahalanobis distance for Mtcars dataset

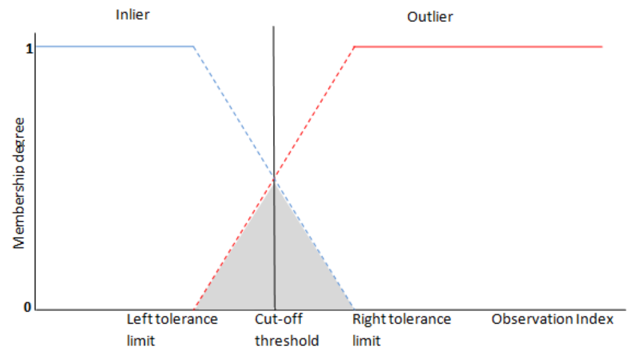

Fig. 2. Proposed Fuzzy Treatment for the crisp cut-off of BACON

B. *Median Absolute Deviation (MAD)*

Using MAD comes from the unreliability of the classical estimates such as the mean in the first approach, since we are dealing with a distance domain which already has outlier values in the first place as shown in Fig. 1, hence the need for more robust measure such as MAD. Considering the non normal distribution of the distances, a heuristic to set the right and left limits of the fuzzified cut-off threshold is given in (5).

$$limit = Cutoff\ threshold \pm \chi^2_{p,\propto} \frac{\text{MAD(distances)}}{\sqrt{n}}, \qquad (5)$$

where χ^2 is the 1 - $\propto$ percentile of the chi square distribution with $\mathcal{P}$ degrees of freedom, whereas n and $\mathcal{P}$ represent the dataset size and dimension, respectively. In the experimentations done $\propto$ was set to 0.05.

V. Dataset

In this section we explain the used datasets in assessing the performance of the proposed fuzzy treatment approaches to the candidate outliers obtained by BACON. Three datasets were used in this study which are Bushfire [17], Philips [18] and Mtcars [19] as the shown Table. I.

In Mtcars dataset, two artificial outliers were planted as points 33 and 34 concerning attributes mpg and disp as shown in Fig. 1. In Bushfire all attributes were involved in the study except the second.

TABLE I.
DATASETS USED IN THE EXPERIMENTS

Dataset	No. of dimensions	Size	Outlying observations
Mtcars	2	34	33,34
Bushfire	4	38	7:11, 32:38, 12,13,31
Philips	9	677	491:565 and other suspicious points

Philips dataset is mainly used because of an effect called masking. Masking is the situation in which there are observations exist as outliers but are not detectable by Mahalanobis distance. These observations remain outliers but in a lower variance subspace. A robust Mahalanobis distance unmasks these observations so they become detectable. In [18] a robust analysis has been conducted on Philips dataset to introduce diagnostic plots to highlight the outlying observations in the presence of the masking effect. As Fig. 3 shows, there are three clusters which are the first 100 points, 491 to 565 and the rest of the observations. Other diagnostic plots for Phillips dataset have been introduced in [20] stating a disagreement concerning the number of the present outlying observations which could possibly include but not restricted to suspicious observations such as 175, 297, 298, 433 [21]. In this study, all the nine attributes of Phillips dataset were involved.

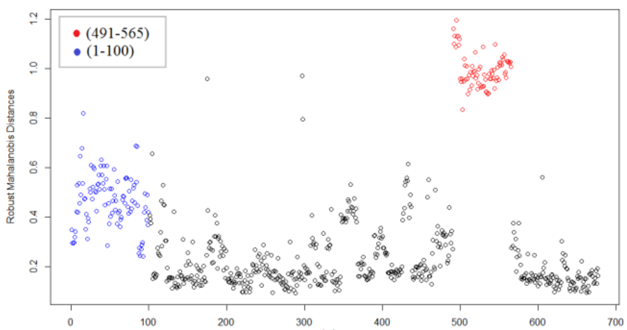

Fig. 3. Robust Mahalanobis distance for Philips dataset

VI. EXPERIMENTS AND RESULTS

In this section we discuss and compare the results of the three algorithms: BACON, Fuzzy BACON V1 (FBACON1), and Fuzzy BACON V2 (FBACON2). BACON remains the fastest along the three datasets as shown in Table. II as both FBACON1 and FBACON2 introduce an additional step to BACON.

TABLE II.
COMPUTATION TIME FOR BACON, FBACON1 AND FBACON2 IN SECONDS

Dataset	BACON	FBACON1	FBACON2
Mtcars	0.345201	1.241204	1.1694
Bushfire	0.314	1.280603	1.233403
Philips	0.3588	8.939511	7.79645

In Mtcars dataset, BACON detected the two planted outliers which are observation 33 and 34. In Bushfire, BACON detected 7 – 13 and 31 – 38 as outlying observations. Bushfire is an ideal example for investigating the labeling of observations that lie on the decision boundary such as 13 and 31. In the current research, we contend that it is intuitive and rational that if an observation lies close to the decision boundary then it should be considered inlier and outlier yet with different belonging degrees to both classes. BACON as discussed in Section IV depends on a test to whether declare an observation to be an outlier. The decision boundary in the last iteration of BACON is highlighted as a red line in Fig. 4 with both observations 13 and 31 being close to it. Yet, they both are labeled as definite outliers as if they lie as far as observation 9 or 10 from the boundary decision as shown in Fig.4.

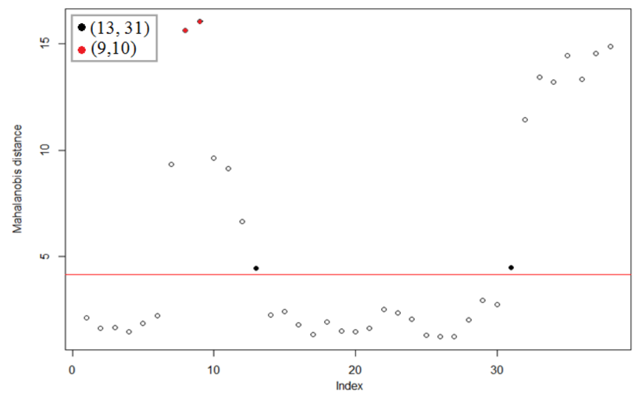

Fig. 4. Mahalanobis distance for Bushfire dataset

The calculated Mahalanobis distances for observations 13 and 31 in the last iteration of BACON are 0.447325 and 4.479342, respectively, and they were tested against the cut-off value of 4.169758 as shown in Fig. 4. On the other hand both of the proposed solutions have assigned membership degrees to observations 13 and 31 in both outlier and inlier classes. Table.III shows the output of FBACON1 and FBACON2 compared to BACON for bushfire dataset. Although FBACON1 and FBACON2 assign both observations higher outlier membership degrees, it remains obvious that both observations are not definite outliers and should not be treated as observation 10 for example. The Results in Table III show how FBACON2 strongly demonstrates the fuzziness of both observation 13 and 31 giving very close membership degrees in inlier and outlier classes as shaded in yellow. The computation time of BACON and both FBACON1 and FBACON2 is shown in Fig. 5 and it shows that FBACON2 slightly outperforms FBACON1.

TABLE III. OUTPUT FOR BACON, FBACON1 AND FBACON2 FOR BUSHFIRE DATASET

Observation	Original BACON output	FBACON1 membership degree		FBACON2 membership degree	
		inlier	*outlier*	*inlier*	*outlier*
1	definite inlier	definite inlier		0.9600174	0.03998259
2	definite inlier	definite inlier		definite inlier	
3	definite inlier	definite inlier		definite inlier	
5	definite inlier	definite inlier		definite inlier	
6	definite inlier	definite inlier		0.9376714	0.06232861
12	definite outlier	definite outlier		definite outlier	
13	definite outlier	0.3520339	0.6479661	0.4376028	0.5623972
14	definite inlier	definite inlier		0.9328029	0.06719707
15	definite inlier	definite inlier		0.8969458	0.1030542
16	definite inlier	definite inlier		definite inlier	
18	definite inlier	definite inlier		definite inlier	
21	definite inlier	definite inlier		definite inlier	
22	definite inlier	definite inlier		0.869288	0.130712
23	definite inlier	definite inlier		0.911155	0.08884496
24	definite inlier	definite inlier		0.9752324	0.02476755
28	definite inlier	definite inlier		0.9793917	0.02060825
29	definite inlier	definite inlier		0.7771846	0.2228154
30	definite inlier	definite inlier		0.8210952	0.1789048
31	definite outlier	0.3349662	0.6650338	0.4304053	0.5695947

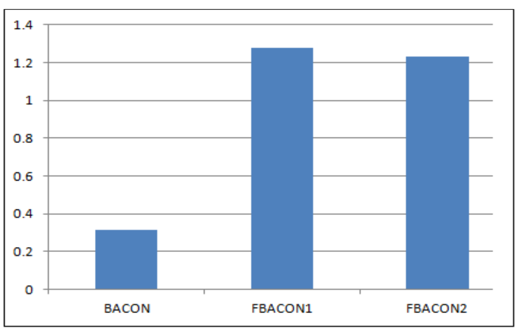

Fig. 5. Computation time for BACON, FBACON1 and FBACON2 for Bushfire dataset

In Philips dataset, BACON found 92 outlying observations [9] including a cluster containing the observations from 491 to 565. BACON also detected other outlying observations which are 16, 297, 298, 429, 430, 431, 432, 433, 435, 436, 70, 95, 104, 116, 120, 175, and 605. In BACON three suspicious observations are labeled to be definite outliers as shown in Fig. 6. The 3 points are 120, 435, and 436. The latter two overlap as they closely lie on the decision boundary. All three observations have been tested in the last iteration of BACON against the value of 5.5737 as cut-off threshold to be declared as definite outliers, although the difference between the testing value and their Mahalanobis distance is small compared to points 297 and 298. The Mahalanobis distance of observations 120, 435 and 436 are 5.602519, 5.65154 and 5.645079, respectively. The three observations should have a degree of membership in both inlier and outlier classes, thus declared as fuzzy outliers rather than definite outliers. The proposed membership degrees of FBACON1 and FBACON2 to the three suspicious observations are shaded in yellow in Table. IV. Moreover, Table. IV shows the assigned membership degrees of FBACON1 and FBACON2 to other observations which were declared to be definite outliers by BACON. Although the results of FBACON1 and FBACON2 do not markedly deviate from each other, they have different computation times as shown in Fig. 7 which shows again that FBACON2 takes less time than FBACON1.

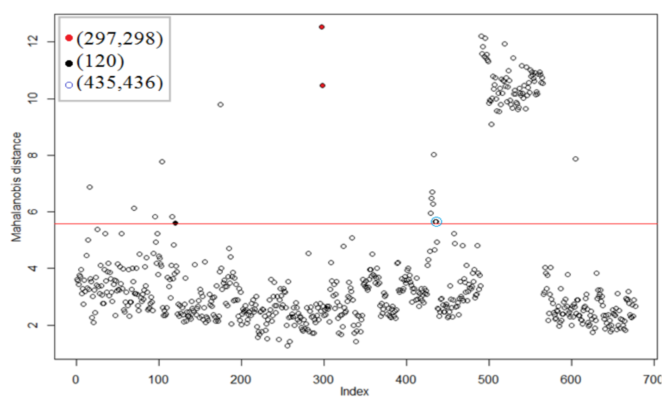

Fig. 6. Mahalanobis distance for Philips dataset

TABLE IV. OUTPUT FOR BACON, FBACON1 AND FBACON2 FOR PHILIPS DATASE

Observation	Original BACON output	FBACON1 membership degree		FBACON2 membership degree	
		inlier	*outlier*	*inlier*	*outlier*
14	definite inlier	0.9303739	0.06962613	0.9960101	0.00399
26	definite inlier	0.6481621	0.3518379	0.6707582	0.329242
35	definite inlier	0.7620384	0.2379616	0.8020018	0.197998
55	definite inlier	0.7654098	0.2345902	0.8058874	0.194113
70	definite outlier	0.09180652	0.9081935	0.02955305	0.970447
95	definite outlier	0.2996337	0.7003663	0.2690759	0.730924
96	definite inlier	0.9887479	0.01125213	definite inlier	
98	definite inlier	0.7582134	0.2417866	0.7975934	0.202407
116	definite outlier	0.3051567	0.6948433	0.2754412	0.724559
120	definite outlier	0.4783641	0.5216359	0.4750644	0.524936
334	definite inlier	0.8612514	0.1387486	0.9163457	0.083654
429	definite outlier	0.2177979	0.7822021	0.1747593	0.825241
435	definite outlier	0.4415288	0.5584712	0.4326113	0.567389
436	definite outlier	0.4463832	0.5536168	0.4382061	0.561794
437	definite inlier	0.9736967	0.02630326	definite inlier	
458	definite inlier	0.7629163	0.2370837	0.8030136	0.196986

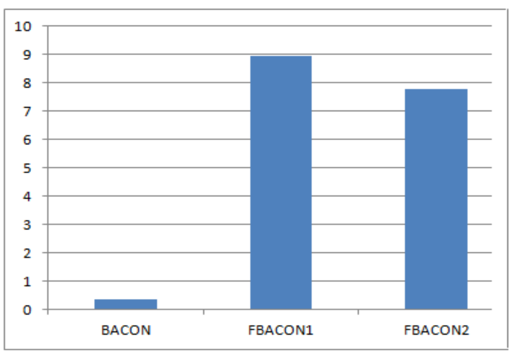

Fig. 7. Computation time for BACON, FBACON1 and FBACON2 for Philips dataset

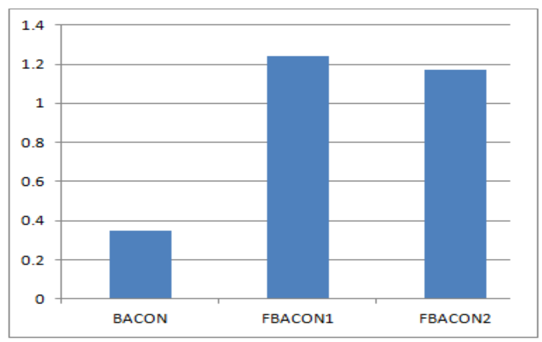

Fig. 8. Computation time for BACON, FBACON1 and FBACON2 for Mtcars dataset

In Mtcars dataset, BACON identified the planted outlying observations as definite outliers which are 33 and 34. The last iteration of BACON relied on the value of 3.318423 as a cut-off to test the Mahalanobis distances of observations 33 and 34. The calculated Mahalanobis distances of observations 33 and 34 are 4.792307 and 29.30728, respectively. Although observation 33 is closer to the decision boundary than observation 34 as shown in Fig.1, BACON declares them both as equally outlying observations. It is also obvious that other observations which BACON considers inliers are also close to the decision boundary in which their degree of belonging to the outlier class should be considered. The membership degrees of all observations of Mtcars dataset are shown in Table. V. In the case of Mtcars dataset, FBACON1 introduces more interpretable results than FBACON2 while the computing time of FBACON2 remains better than FBACON1 as shown in Fig. 8.

VII. CONCLUSION

Using a crisp cut-off threshold in outlier detection can only be useful when all outlying observations being positioned far from it. If an observation lies relatively close to the cut-off threshold then it is more meaningful to be labeled as a fuzzy outlier and to have membership degrees in both inlier and outlier classes. In this paper we proposed two fuzzification approaches to the candidate outliers produced by BACON named as FBACON1 and FBACON2. Experiments showed the ability of the proposed approaches to handle the fuzzy nature of candidate outlying observations rather than declaring them as definite outliers and capturing their uncertainty through introducing more meaningful interpretations. This, in turn, shall bring new insights to the fuzzy nature of decision-boundary points. The insights may also lead to an optimization of the parameter tuning process to precisely tune the limits of decision boundary and, consequently, the number and nature of the outliers.

TABLE V. OUTPUT FOR BACON, FBACON1 AND FBACON2 FOR MTCARS DATASET

Observation	Original BACON output	FBACON1 membership degree		FBACON2 membership degree	
		inlier	*outlier*	*inlier*	*outlier*
3	definite inlier	0.9749073	0.02509269	definite inlier	
5	definite inlier	0.888199	0.111801	definite inlier	
10	definite inlier	0.9812789	0.01872109	definite inlier	
11	definite inlier	0.8885347	0.1114653	definite inlier	
15	definite inlier	0.8119575	0.1880425	definite inlier	
16	definite inlier	0.8340223	0.1659777	definite inlier	
17	definite inlier	0.8074302	0.1925698	definite inlier	
18	definite inlier	0.7426551	0.2573449	definite inlier	
19	definite inlier	0.8568625	0.1431375	definite inlier	
20	definite inlier	0.6631987	0.3368013	definite inlier	
21	definite inlier	0.9529676	0.04703239	definite inlier	
24	definite inlier	0.9990268	0.000973238	definite inlier	
25	definite inlier	0.7222767	0.2777233	definite inlier	
26	definite inlier	0.9686658	0.03133423	definite inlier	
28	definite inlier	0.8373078	0.1626922	definite inlier	
30	definite inlier	0.9349432	0.06505679	definite inlier	
32	definite inlier	0.9507306	0.04926945	definite inlier	
33	definite outlier	0.1644011	0.8355989	definite outlier	
34	definite outlier				

REFERENCES

[1] Van den Broeck J, Cunningham SA, Eeckels R and Herbst K, "Data cleaning: detecting, diagnosing, and editing data abnormalities," PLOS Medicine, 2005.

[2] Hodge, V.J. and Austin, J, "A survey of outlier detection methodologies, " Artificial Intelligence Review, vol. 22, pp. 85-126, 2003.

[3] Yeung DY and Ding Y, "Host-Based Intrusion Detection Using Dynamic and Static Behavioral Models, "Pattern Recognition, 2003.

[4] Binu Thomas and Raju G, "A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining," FCM International Journal of Recent Trends in Engineering, vol. 1, May 2009.

[5] L.A. Zadeh," Fuzzy sets, Inform. and Control", vol. 8, pp. 338-353, 1965.

[6] Ruiz, M. and Martin-Bautista, Maria and Sánchez, Daniel and Vila, M. and Calvo-Flores, Miguel, "Anomaly detection using fuzzy association rules. International Journal of Electronic Security and Digital Forensics," vol. 6, pp. 25-37, 2014.

[7] Tanatavikorn, Harakhun and Yamashita, Yoshiyukim, "Fuzzy Treatment Method for Outlier Detection in Process Data," Journal of chemical engineering of Japan, vol. 49, pp. 864-873, 2016.

[8] Winkler R, Klawonn F and Kruse R, "Fuzzy C-Means in High Dimensional Spaces," International Journal of Fuzzy System Applications, vol. 11, 2010.

[9] N. Billor, A. Hadi and P. Velleman, "BACON: Blocked Adaptive Computationally-Efficient Outlier Nominators," Computational Statist & Data Analysis, vol. 34, pp. 279-298, 2000.

[10] Deepa Verma, Rakesh Kumar and Akhilesh Kumar, "Survey Paper on Outlier Detection Using Fuzzy Logic Based Method," International Journal on Cybernetics & Informatics, April 2017.

[11] Y. K. Kim, S. Y. Lee, S. Seo and K. M. Lee, "Fuzzy logic-based outlier detection for bio-medical data," International Conference on Fuzzy Theory and Its Applications, pp. 117-121, 2014.

[12] Cateni, Sivia & Colla, Valentina and Vannucci, Marco, "A fuzzy logic-based method for outliers detection," Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, pp. 605-610, 2007.

[13] F. E. Grubbs, "Sample criteria for testing outlying observations," The Annals of Mathematical Statistics, vol.21 , pp. 27–58, 1950.

[14] G. E. D'Errico and N. Murru, "Fuzzy treatment of candidate outliers in Measurements,"Advances in Fuzzy Systems, 2012.

[15] W. J. Dixon, "Analysis of extreme values," The Annals of Mathematical Statistics, vol. 21, pp. 488–506, 1950.

[16] K. Pearson, "Outliers in process modeling and identification, "IEEE Transactions on Control Systems Technology, vol. 10 , pp. 55–63, 2002.

[17] N.A. Campbell, "Bushfire mapping using noaa avhrr data". Technical report, CSIRO, 1989

[18] J. Rousseeuw and K. V. Driessen, "A fast algorithm for the Minimum Covariance Determinant estimator". Technometrics, vol. 41, pp. 212–223.

[19] Henderson and Velleman, "Building multiple regression models interactively". Biometrics, vol. 37, pp. 391–411, 1981.

[20] Willems G, Joe H, Zamar R, "Diagnosing multivariate outliers detected by robust estimators". Journal of Computational and Graphical Statistics vol. 18, pp. 73–91, 2009.

[21] C. Béguin and B. Hulliger, "Robust multivariate outlier detection and imputation with incomplete survey data", EUREDIT Deliverable D4/5.2.1/2 Part C, 2003.