



# Adaptive Clustering for Outlier Identification in High-Dimensional Data

Srikanth Thudumu<sup>1</sup>(✉), Philip Branch<sup>1</sup>, Jiong Jin<sup>1</sup>, and Jugdutt Jack Singh<sup>2</sup>

<sup>1</sup> Swinburne University of Technology, Hawthorn, VIC 3122, Australia  
{sthudumu,pbranch,jiongjin}@swin.edu.au

<sup>2</sup> State Government of Sarawak, Kuching, Malaysia  
jack.singh@sarawak.gov.my

**Abstract.** High-dimensional data brings new challenges and opportunities for domains such as clinical, scientific and industry data. However, the curse of dimensionality that comes with the increased dimensions causes outlier identification extremely difficult because of the scattering of data points. Furthermore, clustering in high-dimensional data is challenging due to the intervention of irrelevant dimensions where a dimension may be relevant for some clusters and irrelevant for others. To address the curse of dimensionality in outlier identification, this paper presents a novel technique that generates candidate subspaces from the high-dimensional space and refines the identification of potential outliers from each subspace using a novel iterative adaptive clustering approach. Our experimental results show that the technique is effective.

**Keywords:** Outlier detection · High-dimensionality problem · Adaptive clustering · Big data

## 1 Introduction

Large amounts of data and data sources have become ubiquitous in recent years and become available for analysis in many application domains. This availability is commonly referred to as “big data” comprising large-volume, heterogeneous, complex, unstructured data sets with multiple, autonomous sources growing beyond the ability of available tools. As Gartner [8] noted, big data demands cost-effective novel data analytics for decision-making that infer useful insights. In recent years, the core challenges of big data have been widely established. These are contained within the five Vs of big data volume, velocity, variety, veracity and value. However, such a definition ignores another important aspect: “dimensionality”, that plays a crucial role in real-world data analysis. Research in the data analytics community has mostly been concerned with “volume”, whereas “dimensionality” of big data has received lesser attention [19].

Dimensionality refers to the number of features, attributes or variables within the data. High-dimensionality refers to data sets that have a large number of

independent variables, components, features, or attributes within the data available for analysis. Data with high-dimensionality has become increasingly pervasive, and has created new analytical problems and opportunities simultaneously. The curse of dimensionality often challenges our intuition based on two and three dimensions [3]. Anomaly detection in high-dimensional data sets is computationally demanding and there is a need for more sophisticated approaches that are currently available. An important issue in big data is outlier or anomaly detection, outliers represent fraudulent activities or other anomalous events that are subject to our interest. The “curse of dimensionality”, may negatively affect outlier detection techniques as the degree of data abnormality in fault-relevant dimensions can be concealed or masked by unrelated attributes. When dimensionality increases, the data set becomes sparse, and the conventional methods such as distance based, proximity based, density based and nearest neighbour becomes far less effective [6]. The average distance between a random sample of data points in a high-dimensional space is much larger than the typical distance between one point and the mean of the same sample in low-dimensional space.

While high-dimensionality is one measure of high volume big data, much recent work has focused on finding anomalies using methods that can only draw implicit assumptions from relatively low dimensional data [1]. Furthermore, when the available dimensions of the data are not relevant to the specific test point, the analysis quality may not be credible as the underlying measurements are affected by irrelevant dimensions. This result in a weak discriminating situation where all data points are situated in approximately evenly sparse regions of full dimensional space. However, computing the similarity of one data point to other data point is essential in the outlier detection process.

Clustering in high-dimensional data space is a difficult task due to the intervention of multiple dimensions. A dimension may be relevant for some specific clusters, but unrelated to others. However, clustering is an indispensable step for data mining and knowledge discovery; characterised by unsupervised learning that seeks to detect homogeneous groups of objects based on the values of their attributes or dimensions and grouping them based on similarity, to reveal the underlying structure of data. Conventional methods of clustering attempt to identify clusters constituted of similar samples based on some statistical significance such as distance measurement. The increase in dimensions facilitates similar distance points originated from sparsity triggered by irrelevant dimensions or other noise, aiding to difficulty in identifying accurate and reliable clusters with high quality. The existence of irrelevant attributes or noise in the subspaces critically impacts the formation of clusters. As a result, different subsets of features may be relevant for different clusters, in addition to which diverse correlations among attributes may tend to determine different clusters. Consequently, the curse of dimensionality has become the main challenge for data clustering in high dimensional data sets [7]. This challenge of the clustering process in high dimensional data makes a global dimensionality reduction process inappropriate to identify a subspace that encompasses all the clusters. Nevertheless, in high-

dimensional space, meaningful clusters can be found by projecting data onto certain lower-dimensional feature subspaces and manifolds [9, 10, 12, 17].

In this paper we propose a novel method of clustering that can identify possible outliers in the candidate subspaces of high-dimensional data. To effectively detect outliers in high-dimensional space, we integrate a technique based on our previous work [16] that explores locally relevant and low-dimensional subspaces using Pearson Correlation Coefficient (PCC) and Principal Component Analysis (PCA).

The structure of the paper is as follows: Sect. 2 presents the related work. Section 3 discusses the proposed algorithm. Section 4 discusses the proposed adaptive clustering framework for outlier identification in high-dimensional data. Section 5 presents the experimental results, followed by the conclusion and future work.

## 2 Related Work

The curse of dimensionality poses significant challenges for traditional clustering approaches, both in terms of efficiency and effectiveness. Tomašev et al. [18] have proved that hubness-based clustering algorithms perform well, whereas standard clustering methods fail due to the curse of dimensionality. Hubness is the tendency of data points to occur frequently to k-nearest-neighbor lists of other data points in a high-dimensional space. To address the challenges of clustering technique in high-dimensional data, Ertöz et al. [6] presented an algorithm that can handle multiple dimensions and varying densities, which automatically determines the number of clusters. The algorithm is more focused on identifying clusters in the presence of noises or outliers but not particularly on outlier detection. Deriving meaningful clusters from the data set is an important step because outliers are hidden due to the sparsity in high-dimensional space. Agrawal et al. [2] presented a clustering algorithm called CLIQUE that accurately finds clusters in large high-dimensional data sets. Schubert et al. [15] presented a framework for clustering by extracting meaningful clusters from uncertain data that visualizes and understand the impact of uncertainty by selecting clustering approaches with less variability.

Furthermore, subspace clustering is another technique that is proposed to address the limitations of traditional clustering, which aims to find clusters in all subspaces, but, it is not effective or scalable in case of increasing dimensionality. Liu et al. [13] proposed identifying subspace structures from corrupted data by an objective function that finds the lowest rank representation among all the candidates and can represent the data samples as linear combinations. Elhamifar and Vidal [4] proposed a method for clustering based on sparse representation from multiple low-dimensional subspaces. They have also proposed sparse subspace clustering algorithm [5] to cluster data points that fall in a union of low-dimensional subspaces. Zimek et al. [20] have discussed some important aspects of the ‘curse of dimensionality’ in detail by surveying specialized algorithms for outlier detection. Many researchers addressed important issues but

the key issue of computationally feasible algorithms for anomaly detection in high dimensional space is still largely open. This paper attempts such an algorithm where outliers are derived from low-dimensional subspaces using a novel iterative clustering technique.

### 3 Proposed Algorithm

The objective of the proposed algorithm is the identification of outliers from the resulting candidate subspaces in the high-dimensional data. Details of elicitation of candidate subspaces are presented in [16]. However, we have included the approach in Algorithm 1 from steps for discovering candidate subspaces in high-dimensional data. The contribution of this paper is the technique based on adaptive clustering approach in the identification of fine-grained outliers from the candidate subspaces of high-dimensional data.

---

#### Algorithm 1. Fine-grained Outliers in High-dimensional data:

---

- 1: Apply Standardization or Normalization
  - 2: **for**  $i = 1$  to no. of dimensions **do**
  - 3:   calculate correlation  $r = \frac{n(\sum ab) - (\sum a)(\sum b)}{\sqrt{[n \sum a^2 - (\sum a)^2][n \sum b^2 - (\sum b)^2]}}$
  - 4: **end for**
  - 5: Calculate positive correlation to *CORR*
  - 6: Calculate negative correlation to *UNCORR*
  - 7: Apply PCA  $X = W \cdot \sum \cdot W^T$  on CORR and generate  $PC1_{corr}$  and  $PC2_{corr}$   
by selecting two highest variances
  - 8: **for**  $i = 1$  to no. of dimensions in *UNCORR* **do**
  - 9:   Apply PCA on  $PC1_{corr}$ ,  $PC2_{corr}$  and  $i^{th}$  dimension of *UNCORR*
  - 10:   Save to resultant subspaces  $CS_i$
  - 11: **end for**
  - 12: **for**  $i = 1$  to no. of candidate subspaces in *CS* **do**
  - 13:   Apply Clustering on each  $CS_i$
  - 14:   Generate optimal  $j$  clusters using Elbow criterion
  - 15:   **for**  $j = 1$  to no. of clusters in each *CS* **do**
  - 16:     Calculate centroid of each cluster  $(x_c, y_c)_i$
  - 17:     **while**  $k < threshold$  **do**
  - 18:       Calculate the distance of centroid and each point in the cluster  
 $D_i = \sqrt{(x_c - x_i)^2 + (y_c - y_i)^2}$
  - 19:       Calculate the mean of all the distances  
 $D_{mean} = \sum(D_i)/N$
  - 20:       Use  $D_{mean}$  as the equivalent radius to formulate a circle
  - 21:       Exclude data points within the circle
  - 22:     **end while**
  - 23:   **end for**
  - 24: **end for**
  - 25: **for**  $i = 1$  to no. of candidate subspaces in *CS* **do**
  - 26:   Append remainder data points
  - 27: **end for**
  - 28: Calculate the occurrences of data points in each CS
-

Algorithm 1 provides a step-by-step approach to the technique. Initially, a standardization technique is applied as a pre-processing step to rescale the range of features of input data set if the features of input data consist of large variances between their ranges. To check the correlation among the dimensions of the input data, a Pearson Correlation Coefficient (PCC) is applied to measure the strength of a linear association among the available dimensions. Highly correlated dimensions are combined to form a correlated subspace, and all the uncorrelated dimensions to an uncorrelated subspace, respectively. PCC calculates correlation coefficient of any two dimensions and generates a series of values between +1 to -1. Therefore the correlation coefficient of every dimension with all the other available dimensions in the dataset is calculated and summed up resulting in a final score. If the resultant final correlation score of any dimension is greater than zero, then that particular dimension belongs to CORR subspace or else it belongs to UNCORR subspace.

Principal component analysis (PCA) is applied on the correlated subspace to identifying two highest variances, called principal components, along which the variation in the data is maximal. The resultant principal components are iteratively combined with each dimension of uncorrelated subspace to populate Candidate Subspaces(CS). Every derived candidate subspace is applied with a K-means clustering technique. To find the optimal number of clusters, Elbow model is applied [11]. Based on the result, every CS generates the required number of clusters. In every cluster, a centroid is calculated along with the mean of the distances of available data points to the centroid which we call an “Equivalent Radius” (ER). A circle is formulated in the cluster, and the data points falling within the circle in each cluster are excluded, and the remainder of data points are carry forwarded to the next stage. A new centroid is calculated again in the next stage based on the remaining data points; mean of the distances among each available data points to the new centroid is calculated for a new ER. Then the data points falling within the circle established on the new ER are excluded again. The remainder of the data points is carried forward to the next stage. This process is repeated until the number of data points drops below a certain threshold. Once the threshold is reached, the data points in each CS are calculated for the number of occurrences.

## 4 Adaptive Clustering Framework

This section discusses the proposed framework based on the adaptive clustering approach. Figure 1 delineates the process of outlier identification from the candidate subspaces of the high-dimensional data.

### 4.1 Local Relevancy and Low-Dimensionality

The local relevant subspaces are defined by applying PCC to the data set that differentiates the correlated and uncorrelated dimensions as given in (1), for all the available dimensions  $1...n$  in the data set where no two columns are

equal ( $a \neq b$ ). The resultant correlated dimensions are referred to a correlated subspace. Each dimension that is in the uncorrelated subspace is referred to a low-dimension.

$$r_{a \neq b} = \frac{n(\sum ab) - (\sum a)(\sum b)}{\sqrt{[n \sum a^2 - (\sum a)^2][n \sum b^2 - (\sum b)^2]}} \tag{1}$$

PCA is applied on the subspace of correlated dimensions using eigen decomposition or singular value decomposition and we call this subspace as locally relevant subspace.

$$X = W \cdot \sum \cdot W^T \tag{2}$$

### 4.2 Candidate Subspaces

The principal components resulted from the correlated subspace are combined with each of the low-dimension available from the uncorrelated subspace are the candidate subspaces of the original data. The intention behind combining every low-dimension of uncorrelated subspace with the principal components of the correlated subspace is to reveal the hidden outliers masked by the curse of dimensionality. Furthermore, data points appearing in more than one CS have the highest probability of being an anomaly or outlier.

### 4.3 Adaptive Clustering

A clustering on each CS is applied to exclude the data points falling within the definition. Section 5 discusses the importance of repetitive application of this technique and the reason we call as ‘‘Adaptive Clustering’’ on candidate subspaces of high-dimensional data.

**K-Means Clustering.** The proposed technique uses a k-means clustering algorithm that flows a simple and easy way to classify a given dataset through a certain number of clusters (K- clusters) fixed a priori [14].

$$\beta = \sum_{i=1}^k \sum_{j=1}^m (\|a_i^{(j)} - c_j\|)^2 \tag{3}$$

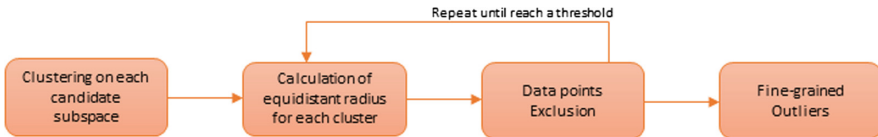


Fig. 1. Outlier identification from candidate subspaces

**Optimal Number of Clusters.** The number of clusters should match the data in the CS. An unfitting selection of the number of clusters may undermine the whole process. The best approach is to use Elbow criterion that interprets and validates the consistency within cluster analysis to find the optimal number of clusters [11]. The Elbow model is applied to each CS to deduce the optimal number of clusters in each CS.

**Equivalent Radius (ER).** The centroid for each cluster in each CS is computed. Then the centroid is used for estimating the mean of the distances between each data point within the cluster to its centroid. The resultant mean value is used to formulate a circle in the cluster. This process is repeated until the total number of data points in each CS are less than the given threshold.

$$D_i = \sqrt{(x_c - x_i)^2 + (y_c - y_i)^2} \quad (4)$$

Calculate the mean of all the distances

$$D_{mean} = \sum(D_i)/N \quad (5)$$

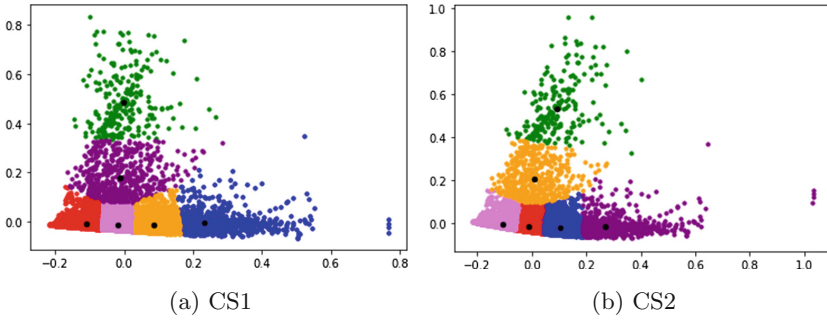
Use  $D_{mean}$  as the equivalent radius to formulate a circle.

**Data Points Exclusion.** The data points inside the circle definition based on the calculation of ER are excluded, and the data points outside the circle are carried out to the next stage. A new centroid is calculated based on the new set of data points and latest ER is used to form another circle. This process of calculation of the new ER is carried out until a specific condition or threshold is reached. If the data points are less than the given threshold, the ER before the given limit is taken into consideration, and the resulting data points from each CS where the threshold is reached are analysed.

**Fine-Grained Outliers.** The calculation of the number of occurrences of each data point in all the CS are calculated based on the final iteration. The more number of times a particular data point appears, the more likely that data point is an outlier. This process is referred to as fine-graining of outliers. The next step is to trace back the fine-grained outliers to its original index.

## 5 Experimental Evaluation

We used a data set with 19 dimensions and 21000 rows, of which 17 are correlated, and 2 are uncorrelated when analysed with PCC. To verify the effectiveness of outlier identification, we have purposefully introduced synthetic anomalies into the data. The combination of correlated subspace with every dimension from uncorrelated subspace with the application of PCA results in two candidate subspaces, as seen in Fig. 2. We applied the proposed technique of adaptive clustering to both candidate subspaces to fine-grain the outliers in each CS.



**Fig. 2.** Original candidate subspaces

In this section, we present the results of three experiments we have conducted to explain the effectiveness of the adaptive clustering approach in identifying outliers. Figure 2a represents the first candidate subspace and Fig. 2b represents the second candidate subspace.

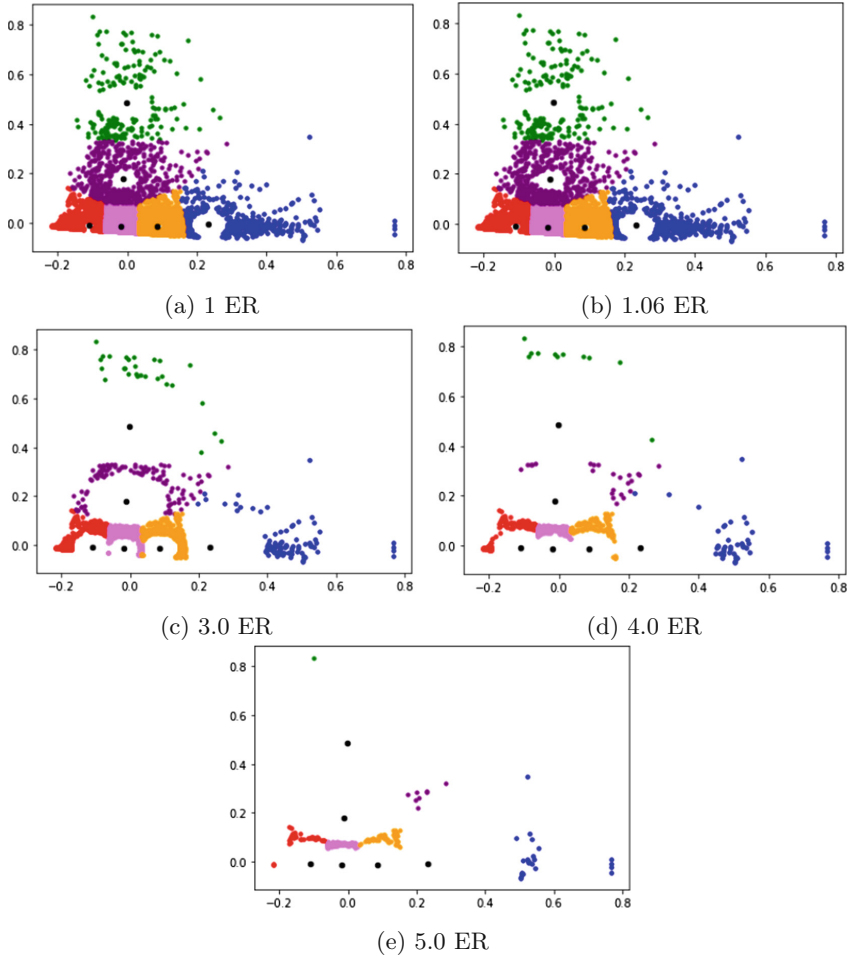
**Table 1.** Equivalent radii and the associated data points

Figure	Equivalent radius	Iterations	No. of data points
a	ER*1	0	15498
b	ER*1.06	0	14951
c	ER*3.0	0	2086
d	ER*4.0	0	569
e	ER*5.0	0	219

### 5.1 Data Points Exclusion Using a Large ER

Identifying anomalous data points from the candidate subspaces is difficult and may not reveal real anomalies as there are many data points in each CS as depicted in Fig. 2. Hence, an efficient technique is required to filter the possible outliers in each CS. In this experiment, we present a technique that finds outliers and evaluates the technique’s effectiveness in outlier identification by taking one candidate subspace CS1 and a large ER, that excludes data points within the circle definition from every cluster. As mentioned in Sect. 4.3, an ER is computed from the mean of the distances of data points available within the cluster to its centroid. The computed ER is used to define a circle, and the data points within the circle definition are excluded from the CS. The motivation behind the proposed equivalent radius is to deselect the nearest points as to reveal hidden outliers.





**Fig. 3.** Exclusion of data points using large radii

Table 1 presents the number of remaining data points after the exclusion of data points from the definition of a circle formed from the respective ER. The increase in ER leads to a decrease in the number of data points remaining. However, this approach is not effective when finding the outliers in each cluster of the CS. Furthermore, the increase in ER caused the circle to grow bigger, excluding even the possible outliers that may be hidden in the clusters. Figure 3 shows the exclusion of data points when the ER is increased progressively. Figure 3a represents the exclusion of data points when the computed mean is taken 1 ER, however, when we multiply 1 ER to 1.06 ( $ER \cdot 1.06$ ) as in Fig. 3b, 3.0 as in Fig. 3c, 4.0 as in Fig. 3d and 5.0 as in Fig. 3e, the declination of data points or irrelevant grouping is observed. To address this issue and to fine-grain the outliers,

we calculated the ER iteratively, defining a new circle after each exclusion and presented in the following section.

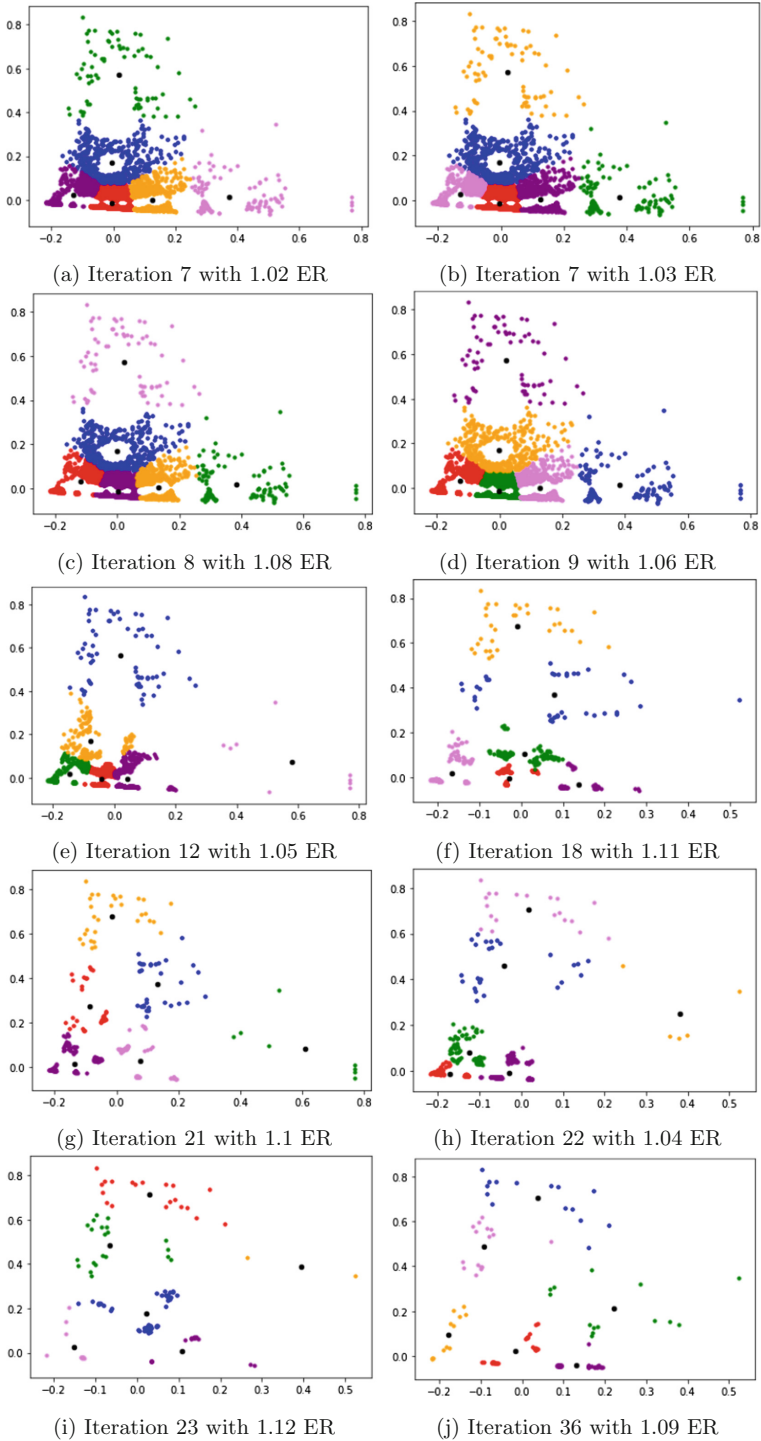
## 5.2 Data Points Exclusion with Iterative ER

In the second experiment, we evaluate the behaviour of outlier identification by iteratively calculating the ER depending on the new set of data points after the exclusion from the previous circle. The process of computing a new ER that forms a dependent circle is terminated when the total data points in the CS becomes lesser than the given threshold of 100 data points.

When the ER is computed to form a circle followed by the exclusion of data points within the respective circle definition, a centroid is calculated again based on the new set of data points upon which new ER is calculated, defining a respective circle area. The process was repeated until the total number of data points became less than 100. The process was terminated, and the most recent CS with before the threshold is benchmarked. If the ER is smaller with the fewer data points or no data points are excluded even with new iterations, an increase in ER value progressively is needed. Originally, second experiment has generated many graphs; however, we present the less results that exhibit the effectiveness of the technique. Table 2 presents the number of data points and iterations processed, along with the remaining data points within the circle definition formed by the respective ER. The results show that outliers can be effectively filtered when the ER is increased progressively and iterated until the threshold is met. Figure 4 shows the exclusion of data points when the ER is increased progressively. Figure 4a shows the remainder of 4413 data points when the computed ER is multiplied with 1.02 ( $ER \times 1.02$ ) after 7 iterations, Fig. 4b shows the remainder of 4304 data points when the ER is increased to  $ER \times 1.03$  after 7 iterations, we observe a slight decline of data points when there is an increase in ER. The data points decreased slightly with more iterations, Fig. 4c

**Table 2.** Iterative ER and the associated data points

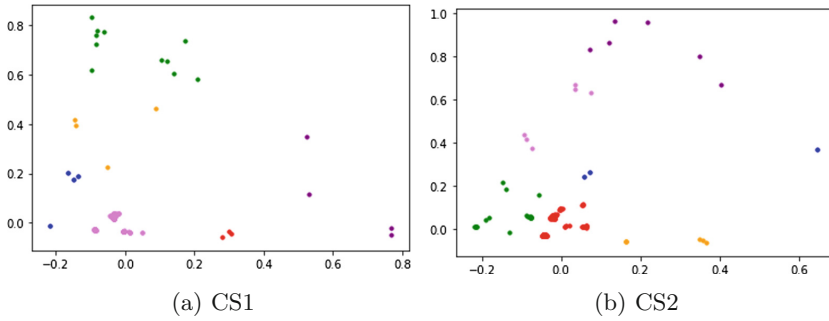
Figure	Equivalent radius	Iterations	No. of data points
a	$ER \times 1.02$	7	4413
b	$ER \times 1.03$	7	4304
c	$ER \times 1.08$	8	3895
d	$ER \times 1.06$	9	3963
e	$ER \times 1.05$	12	1869
f	$ER \times 1.11$	18	502
g	$ER \times 1.1$	21	359
h	$ER \times 1.04$	22	608
i	$ER \times 1.12$	23	116
j	$ER \times 1.09$	36	123



**Fig. 4.** Exclusion of data points using ER iteration approach

consists of 3895 data points with 8 iterations, however, Fig. 4d has 3963 data points with 1.06 ER even after 9 iterations. Hence, the increase in ER filters out more data points than more iterations. We continued the experimentation to observe the decline of the grouping of data points, at 12<sup>th</sup> iteration with ER\*1.05, we observed that there is a slight decline of irrelevant grouping as observed in Fig. 4e.

We continued observing the data points Fig. 4f with 18 iterations and ER\*1.11, Fig. 4g with 21 iterations and ER\*1.1, Fig. 4h with 22 iterations and ER\*1.04, Fig. 4i with 23 iterations and ER\*1.12, and Fig. 4j with 36 iterations and ER\*1.09. Finally, when the threshold of less than 100 data points is reached, the iteration stops and resulting in Fig. 4i with the remainder of 116 data points and Fig. 4j with 123 data points. When compared the two final CS with the results generated when using one large ER presented in Sect. 5.1, the iterative approach of calculating ER outperforms the first experiment with one large ER.



**Fig. 5.** Final candidate subspaces

### 5.3 Calculation of Occurrences for Fine-Grain Outliers

To identify the synthetically introduced outliers, we integrated a technique for the occurrence calculation of each data point in the final candidate subspaces. The more times a particular data point appears in all of the candidate subspaces, the more likely the data point is an outlier. We call the most appeared data points fine-grained outliers. To verify the synthetically introduced outliers are in the final subspace, we traced back each data point to its original index location before evaluating the occurrence in each CS. As observed in Fig. 5, the final data points in CS1 (Fig. 5a) and CS2 (Fig. 5b), 90% of synthetically introduced outliers have appeared in both the candidate subspaces, and 10% of them appeared once. However, it is to be noted that all the introduced outliers are observed in the final candidate subspaces.

## 6 Conclusion and Future Work

This paper introduces Adaptive Clustering that identifies the outliers from the candidate subspaces of the high-dimensional data. To reduce the effect caused by the curse of dimensionality PCC and PCA are integrated to define locally relevant and low-dimensional subspaces. An equivalent radius in each cluster of the candidate subspace is calculated based on the mean of the distances between the centroid and the data points. An iterative application of equivalent radius is computed and used to exclude data points of no interest. To demonstrate that iterative calculation of equivalent radius is more effective, we evaluated the results from both large equivalent radii and iterative calculations of ER and showed that the iterative approach outperforms the other approach. Finally, the resultant data points in each candidate subspace are computed for the number of occurrences. The more times a data point appears, the more likely it is an outlier. In our future work, we will evaluate the performance and accuracy of the proposed technique by analysing the trade-off with respect to the volume and dimensionality to develop a big data framework.

## References

1. Aggarwal, C.C., Philip, S.Y.: An effective and efficient algorithm for high-dimensional outlier detection. *VLDB J.* **14**(2), 211–221 (2005)
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications, vol. 27. ACM (1998)
3. Christiansen, B.: Ensemble averaging and the curse of dimensionality. *J. Clim.* **31**(4), 1587–1596 (2018)
4. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2790–2797. IEEE (2009)
5. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2765–2781 (2013)
6. Ertöz, L., Steinbach, M., Kumar, V.: Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Proceedings of the 2003 SIAM International Conference on Data Mining, pp. 47–58. SIAM (2003)
7. Gan, G., Ng, M.K.P.: Subspace clustering with automatic feature grouping. *Pattern Recogn.* **48**(11), 3703–3713 (2015)
8. Gartner, I.: Big data definition. <https://www.gartner.com/it-glossary/big-data/>. Accessed 6 Sept 2019
9. Jing, L., Ng, M.K., Huang, J.Z.: An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. Knowl. Data Eng.* **8**, 1026–1041 (2007)
10. Jing, L., Ng, M.K., Xu, J., Huang, J.Z.: Subspace clustering of text documents with feature weighting  $K$ -means algorithm. In: Ho, T.B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 802–812. Springer, Heidelberg (2005). [https://doi.org/10.1007/11430919\\_94](https://doi.org/10.1007/11430919_94)
11. Ketchen, D.J., Shook, C.L.: The application of cluster analysis in strategic management research: an analysis and critique. *Strateg. Manag. J.* **17**(6), 441–458 (1996)

12. Li, T., Ma, S., Ogihara, M.: Document clustering via adaptive subspace iteration. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 218–225. ACM (2004)
13. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 171–184 (2012)
14. Mucha, H.J., Sofyan, H.: Nonhierarchical clustering (2011)
15. Schubert, E., Koos, A., Emrich, T., Züfle, A., Schmid, K.A., Zimek, A.: A framework for clustering uncertain data. *Proc. VLDB Endow.* **8**(12), 1976–1979 (2015)
16. Thudumu, S., Branch, P., Jin, J., Singh, J.J.: Elicitation of candidate subspaces in high-dimensional data. In: 2019 IEEE 21st International Conference on High Performance Computing and Communications. IEEE (2019, in press)
17. Tomasev, N., Radovanovic, M., Mladenic, D., Ivanovic, M.: The role of hubness in clustering high-dimensional data. *IEEE Trans. Knowl. Data Eng.* **26**(3), 739–751 (2014)
18. Tomašev, N., Radovanović, M., Mladenčić, D., Ivanović, M.: Hubness-based clustering of high-dimensional data. In: Celebi, M.E. (ed.) *Partitional Clustering Algorithms*, pp. 353–386. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-09259-1\\_11](https://doi.org/10.1007/978-3-319-09259-1_11)
19. Zhai, Y., Ong, Y.S., Tsang, I.W.: The emerging “big dimensionality” (2014)
20. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.: ASA Data Sci. J.* **5**(5), 363–387 (2012)