A High-Dimensional Outlier Detection Approach Based on Local Coulomb Force

Pengyun Zhu, Chaowei Zhang, Xiaofeng Li, Jifu Zhang, *Member, IEEE*, and Xiao Qin, *Senior Member, IEEE*

Abstract—Traditional outlier detections are inadequate for high-dimensional data analysis due to the interference of distance tending to be concentrated ("curse of dimensionality"). Inspired by the Coulomb's law, we propose a new high-dimensional data similarity measure vector, which consists of outlier Coulomb force and outlier Coulomb resultant force. Outlier Coulomb force not only effectively gauges similarity measures among data objects, but also fully reflects differences among dimensions of data objects by vector projection in each dimension. More importantly, Coulomb resultant force can effectively measure deviations of data objects from a data center, making detection results interpretable. We introduce a new neighborhood outlier factor, which drives the development of a high-dimensional outlier detection algorithm. In our approach, attribute values with a high deviation degree is treated as interpretable information of outlier data. Finally, we implement and evaluate our algorithm using the UCI and synthetic datasets. Our experimental results show that the algorithm effectively alleviates the interference of "Curse of Dimensionality". The findings confirm that high-dimensional outlier data originated by the algorithm are interpretable.

Index Terms—High-dimensional outlier detection, similarity metric, outlier coulomb resultant force, local outlier coulomb force, neighborhood outlier factor

1 Introduction

Outliers data, which clearly deviates from other data, are inconsistent with general patterns or behaviors [1]. Outlier detection techniques are widely used in sensors network [2], fraud detection in financial insurance [3], [4], medical diagnostic detection [5], [6], intrusion detection [7], industrial systems [8], celestial spectroscopy [9] and other fields. More times than not, outliers contain a large amount of valuable information that is not discovered by humans. Because of the data distribution of high-dimensional data being extremely sparse, outlier detection is interfered by the phenomenon of the distance tending to be concentrated, which is referred to as "curse of dimensionality" [10]. Therefore, one research objectives in this study is to tackle a challenging issue - "curse of dimensionality" - in the task of detecting high-dimensional outlier data.

Coulomb force is an electromagnetic force described as an interaction among charges in a vacuum in physics. As a vector, coulomb force contains rich distance and direction information among particles [11]. Between any two charges, the Coulomb force has the same magnitude and opposite direction. In the Cartesian coordinates, the component of Coulomb force vector on any coordinate axis (component force of the Coulomb force) reflects distance and orientation of two charges on the coordinate axis. The Coulomb resultant force is the vector sum of Coulomb force, which multiple charges acting on the same charge. The more concentrated charge is around by others, the more concentrated the direction of the Coulomb force is, and the greater the Coulomb resultant force will be. On the contrary, dispersed charges around a center charge scatters Coulomb force's direction coupled with reduced Coulomb resultant force. And the direction in which the charge is subjected to the Coulomb resultant force is the direction, in which the charge deviates from the center of an electric field.

For high-dimensional datasets, any data object is regarded as a charge in an electrostatic field; attribute dimensions are envisioned as the Cartesian coordinates in the electrostatic field. Therefore, data objects are characterized and modeled by Coulomb force. A Coulomb force vector between any two data objects reflects its similarity information on any attribute dimension. The Coulomb resultant force of a data object reflects the deviation and direction of the data object deviating from a data center. If a data object deviates more in a certain dimension, the projection of the Coulomb resultant force in that dimension (component of the Coulomb resultant force) will be greater, and

- *Pengyun Zhu, Xiaofeng Li, and Jifu Zhang are with School of Computer Science and Technology, Taiyuan University of Science and Technology (TYUST), Taiyuan, Shanxi 030024, China. E-mail: 13253518976@163.com, lxfffeng@126.com, zjf@tyust.edu.cn.*
- *Chaowei Zhang is with the Department of Computer Science, Yangzhou University, Yangzhou, Jiangsu 225127, China. E-mail: cwzhang@yzu.edu.cn.*
- *Xiao Qin is with the Department of Computer Science and Software Engineering, Samuel Ginn College of Engineering, Auburn University, Auburn, AL 36849-5347 USA. E-mail: xqin@auburn.edu.*

Manuscript received 11 October 2021; revised 7 February 2022; accepted 28 April 2022. Date of publication 3 May 2022; date of current version 1 May 2023.
The work of Pengyun Zhu, Xiaofeng Li, and Jifu Zhang were supported in part by the National Natural Science Foundation of China under Grant 62172293 and in part by the Scientific and Technological Innovation Team Foundation of Shanxi Province, China under Grant 201805D131007. The work of Xiao Qin was supported in part by the U.S. National Science Foundation under Grants IIS-1618669, OAC-1642133, and CCF-0845257.
(Corresponding author: Jifu Zhang.)
Recommended for acceptance by Q. He.
Digital Object Identifier no. 10.1109/TKDE.2022.3172167

1041-4347 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

vice versa. We propose to make use of projection as interpretation information of outlier data. Since the Coulomb resultant force of a data object embraces distance and direction, this approach effectively reduces the influence of attribute dimensions with low deviation and enhances the influence of attribute dimensions with high deviation. The higher the dimensions are, the richer the deviation information is. Therefore, the use of outlier Coulomb resultant force to measure the deviation degree of data objects is expected to effectively reduce the interference of the "Curse of Dimensionality". With the data object's outlier Coulomb resultant force in place, we define a neighborhood outlier factor as a new high-dimensional outlier data measurement, and develop a high-dimensional outlier data detection algorithm to alleviate the "Curse of Dimensionality" interference in high-dimensional outlier detection. The key contributions of this paper are summarized as follows:

- We propose a similarity measure of high dimensional data based on Coulomb force.
- A new neighborhood outlier factor is introduced to gauge outliers.
- We define interpretable information of outlier data.
- A novel high-dimensional outlier detection algorithm based on local coulomb force is proposed.

2 Related Work

2.1 Similarity Measure

Measuring the degree of similarity between any two data objects is a core task in data analysis, data mining, and machine learning. Popular similarity measures mainly includes euclidean distance, Manhattan distance, Cosine similarity, Chebyshev distance, Hamming distance, Minkowski distance, Jaccard index, Haversine distance, Srensen-Dice index, and the like. Jaccard index and Srensen-Dice index are widely adopted in image segmentation tasks or text similarity analysis. Hamming distance is a key player in measuring similarities among classification data objects. The Haversine distance is only applicable to datasets where all data objects are on the same sphere. euclidean distance, Cosine similarity, Manhattan distance, Chebyshev distance, Minkowski distance are often used to gauge similarities among numerical data objects.

Euclidean distance is the most commonly employed similarity measure in outlier detection tasks. In a high-dimensional data space, the "concentration effect" (the distances among data objects being very close) makes euclidean distance ineffective in measuring the similarity of high-dimensional data objects. Evidence [1] suggests using the extension of a fractional norm ($Lp(0 < p < 1)$) and euclidean norm to deal with the "concentration effect". Another study [12] shows that "concentration effect" is an intrinsic property of distance, meaning that all p-norms have "concentration effect" that is independent of sample sizes. Therefore, in high-dimensional outlier detection tasks, norm distances (e.g., euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance) are interfered by the "concentration effect" of distance convergence. Cosine similarity is considered to be a common distance measure to alleviate the "concentration effect", but the cosine similarity also has the phenomenon of "concentration effect" [13]. A vector method anchored on the LRF measurement of local gravity ,it assumes that the distances between the current object and its different neighbors do not vary significantly, ignoring the relationship between the magnitude and the direction of the force [14]. This solution results in magnitude and direction of the resultant force, thereby leading to a "concentration effect". In a nutshell, the interference of the "Curse of Dimensionality" caused by the "concentration effect" makes the traditional distance measurements undesirable for detecting high-dimensional outliers.

2.2 High-Dimensional Outlier Detection

Outlier detection, playing a major role in the field of data mining, mainly includes statistics-based [15], [16], distance-based [17], [18], density-based [19], [20], clustering-based [21], [22], subspace-based [23], [24] methods. Traditional approaches anchored on statistics, distance, density, and clustering are adversely affected by the "Curse of Dimensionality", which is the distance tending to concentrate [10], [25], [26]. For example, Markus *et al.* [19] introduced the concept of local outliers and proposed a density-based local outlier detection (LOF) method. Unfortunately, this method cannot identify outliers in complex neighborhood relationship patterns. Tang *et al.* [15] proposed an outlier detection method (RDOS) based on local kernel density estimation using k-nearest neighbors, reverse nearest neighbors, and shared nearest neighbors. The efficiency of this method is low due to the complexity of nearest neighbors search. Yang *et al.* [27] developed the MOD and MOD^{+} methods of the extended reference set using the mean shift of data objects. The mean-shift technique replaces a data object with the mean of its k-nearest neighbors. This study unveils that a large number of outliers in real-world datasets can be effectively uncovered by the mean-shift technique. Dong *et al.* [28] piloted two cutting-edge studies in high-dimensional robustness statistics, namely, the mean estimation of robustness and outlier detection. Importantly, Dong *et al.* developed a new outlier scoring method or QUE-scoring by leveraging the quantum entropy regularization. Wang *et al.* [14] devised a gravitational model (LRF) that roughly distinguishes boundary points, internal points, and outliers based on the theory of universal gravitation. This model uses an average distance as a weighted weight of unit vectors. However, this method ignores the impact of discrepancies among data objects on vectors leading to the gravitation neutralization among the opponent data objects that have different values and same directions. The model is so complex that it can only be roughly distinguish outlier. Inspired by the gravitational model (LRF), Xie *et al.* [29] designed a local gravity-based outlier detection method (LGOD) using the difference in LRF length. This technique is implemented in the following three steps. Local resultant forces are initially calculated for each data objects. Then, the variation of the local resultant forces for each data point are accumulated under different number of neighbors. Finally, the type of data objects is determined by the rate of accumulative variation. However, the method suffers and low accuracy due to the ignorance of LRF's direction information. The only ABOD method that adopts the cosine similarity measure alleviates the "Curse of Dimensionality" [18]. Nevertheless, this method is unable to properly identify outliers in the

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

low-density surrounding area [26]; the cosine similarity also exhibits a "concentration effect" phenomenon [13].

Subspace-based outlier detection is one of the hotspots of high-dimensional outlier data mining research. This strategy is considered to be an effective way to mitigate the "Curse of Dimensionality" [23], [26], [30]. We classify subspace-based approaches into two categories, namely, sparsity subspace methods [31], [32], [33], [34] and related subspace methods [20], [24], [30], [35]. Subspace outlier detection assumes that outlier data are usually hidden in local abnormal behaviors of low-dimensional subspace, where abnormal behaviors are covered by full-dimensional analysis. For instance, Agarwal *et al.* applied genetic algorithms to search for sparse subspaces. Due to the influence of the initial population, the completeness and accuracy cannot be guaranteed [31]. Zhang *et al.* adopted the concept lattice as a subspace description tool to improve the accuracy and completeness of outlier detection in response to the deficiencies [31]. The findings unfold that the concept lattice structure is complex and the detection efficiency is low [32]. Zhao *et al.* deployed the particle swarm algorithm to search a sparse subspace; a local and context outlier detection algorithm was proposed to effectively improve the outlier detection effect and interpretability [33], [34]. Kriegel *et al.* developed an axis-parallel subspace outlier mining (SOD) method, which only captures the outlier degree of data objects in one related subspace. When the data objects are distributed in two or more subspaces, SOD will be unable to distinguish their outlier degrees [24]. Zhang *et al.* designed a scalable mining algorithm running on the MapReduce programming platform. By representing the attribute dimension and local outlier factor of the relevant subspace as important context information, the interpretability and comprehensibility of outliers are enhanced [23], [36]. Muller *et al.* used the statistical method of the Kolmogorov-Smirnov test to propose a scheme of selecting meaningful subspaces on the data set, solving the comparability problem of outlier factors in different subspaces. The efficiency of this algorithm is low due to its high time complexity at the exponential level of dimensions [35].

In summary, the paradigm distances (e.g., euclidean distance, Manhattan distance, Chebyshev distance, and Minkowski distance) are a cumulative sum of differences in all attribute dimensions. Such distances are ineffective in distinguishing a large number of sample data objects with various directions coupled with an identical Hypersphere. In contrast, the cosine similarity-based and the gravitation-based metrics merely deal with the relative directional information of data objects to detect outliers, thereby discriminating data objects that are in distinct directions on a hypersphere. Since the two metrics never incorporate vector size, it is arduous to differentiate a large group of data samples that share a same direction but have various distance values. Additionally, the distances among data points in a high-dimensional dataset in the real world are inconsistent, but it is assumed by the two metrics that such distances are consistent. Therefore, a phenomenon of "Curse of Dimensionality" or "concentration effect" is formed in high-dimensional datasets, making an obstacle in detecting outlies from high-dimensional data. Although subspace-based outlier detection techniques are an effective way to alleviate the interference of "Curse of Dimensionality", finding meaningful subspace is nontrivial. Moreover, subspace-based approaches exhibits high overhead due to exponential time complexity of dimensionality. Weakly correlated or uncorrelated dimensions may generate too much noise, thereby identifying concealing outliers with only relevant dimensions (subspaces). Subspace outlier detection can be viewed as a local feature selection method or a local dimensionality reduction method, which inevitably lose critical information.

3 Coulomb Force and Coulomb Resultant Force

The Coulomb's law is a law of physics published in 1785 by the French physicist Charles-Augustin de Coulomb [11]. This law asserts that an interaction force between two static charges in a three-dimensional vacuum environment. The interaction force is inversely proportional to the square of the distance, proportional to the product of the electric quantity, and the direction is on the line between the two charges. Charges of the same sign repel each other; charges of different sign attract. The literature [37] unfolds that the size of a Coulomb force depends on dimension, and an interaction force between two static charges in an m-dimensional vacuum environment is inversely proportional to the $m-1$ power of the distance. Referring to the literature [11], [37], we articulate the relevant theorems and concepts follows:

Given any two electric charges p and p' with electric quantity q and q' and the positions r and r', the Coulomb force vector of charge q' acting on charge q is

$$\overrightarrow{F}(p,p') = Eq'\vec{r} = k_m \frac{qq'}{\|r-r'\|^{m-1}}\vec{r}. \tag{1}$$

Where E denotes the electric field intensity at the position of charge p' in the electrostatic field with charge p as a field source charge calculated as $k_m q/\|r-r'\|^{m-1}$; k_m is an electrostatic constant in m-dimensional space; $\vec{r}$ represents the unit vector from charge p' to charge p with r as the starting point computed as $(r-r')/\|r-r'\|$; $\|r-r'\|$is an euclidean distance between charges p and p'.

Coulomb force $\overrightarrow{F}(p,p')$, as a vector, can be expressed as $[f_1(p,p'), f_2(p,p'), \ldots, f_d(p,p'), \ldots, f_m(p,p')]$, where m is the dimension of an electrostatic field. And for any charges p_1 and p_2, the Coulomb force $\overrightarrow{F}(p_0,p_1) \neq \overrightarrow{F}(p_0,p_2)$ and is independent of dimension as long as $p_1 \neq p_2$.

In m-dimensional electrostatic field Q composed of n charges, the sum of a Coulomb force of any charge p subjected to the other charges $p' \in Q \setminus p$ is called Coulomb resultant force defined as

$$\overrightarrow{F_Q}(p) = \sum_{p' \in Q \setminus p} \overrightarrow{F}(p,p') = [f_1(p), \ldots, f_d(p), \ldots, f_m(p))]. \tag{2}$$

Where $f_d(p)$ is a summation of the value on dimension d of Coulomb force of charge p subjected to the other charges $p' \in Q \setminus p$, can be calculated as $\sum_{p' \in Q \setminus p} f_d(p,p')$.

Coulomb force accurately describes a relative positional relationship among charge p and the other charges. Coulomb resultant force reflects a degree which charge p deviates from an electrostatic field. Fig. 1 reveals a Coulomb

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

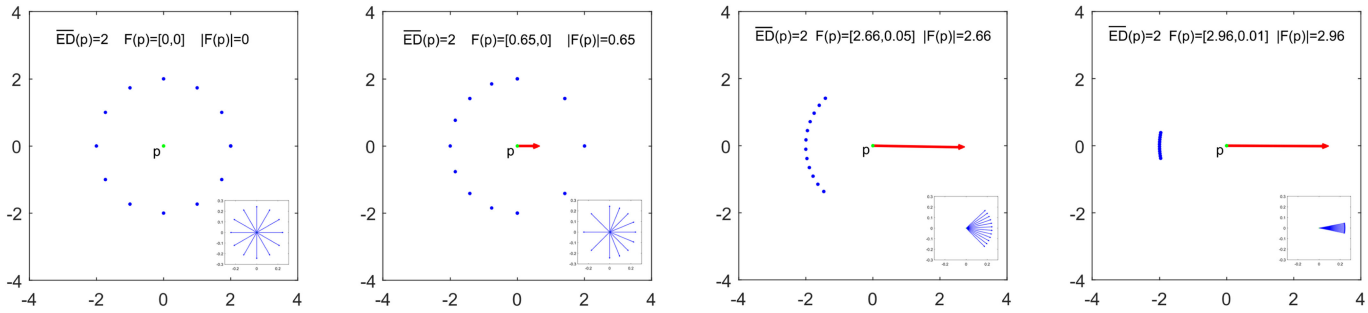


Fig. 1. Coulomb force and Coulomb resultant force.

force and Coulomb resultant force of charge p in four electrostatic field of different distributions, where the solid dots indicate charges, the red arrow signifies Coulomb resultant force, the blue arrow denotes Coulomb force. The bottom right corner of the figure is a schematic diagram of the Coulomb force of the charge p subjected to the blue charges. The blue charges around the green charge gradually concentrate. Under the condition of euclidean distance measure, an average euclidean distance ($\overline{ED}(p) = 2$) is exactly the same between charge p and all the blue charges in the four electrostatic fields. In contrast, under the condition of Coulomb force measure, the Coulomb resultant force is different among p and the blue charges in the four electrostatic fields, and this force is consistent with the distribution of each electrostatic field. That is, the more concentrated the blue charges relative to p, the greater of Coulomb resultant force is, meaning that it always points in the opposite direction to the center of the electrostatic field. Considering a dataset as an electrostatic field and the data objects as charges, the Coulomb resultant force is an adequate indicator to gauge a deviation of a data object from its data center.

4 High-Dimensional Outlier Measures

Traditional distance measurements, which use a scalar as a measure of similarities among data objects, fail to reflect differences of data objects in all dimensions. Not surprisingly, these measurement methods suffer from the "concentration effect" to varying degrees. However, the projection of the vector in dimensions reflects the differences in the dimensions of data objects, and the Coulomb force similarity measurement alleviates the pain of of the "concentration effect". Coulomb resultant force as a measure of the degree of deviation can effectively reduce the influence of attribute dimensions with low degree of deviation while enhancing the influence of attribute dimensions with a high degree of deviation. And the higher dimension, the richer information of deviation. The Coulomb resultant force can effectively reduce interference of the "Curse of Dimensionality" in the detection of high-dimensional outliers, making outlier detection results interpretable.

4.1 Outlier Coulomb Force and Similarity Measures

Let us consider dataset $D = [X_1, X_2, \ldots, X_i, \ldots, X_n]^T$, where n is the number of data objects and X_i is the i-th data object in the data set D. Data object X_i is denoted as $X_i = [x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im}]$, $X_i \in D$, where m is the dimensions of dataset, and x_{ij} is the j-th attribute value of the i-th data object in the dataset. Data X_i in data set D can be regarded as charge p_i in electrostatic field Q; the attribute dimension is envisioned as the Cartesian coordinate system in an electrostatic field. Assuming that the importance of data objects in dataset D is the same, then different data objects can be regarded as charges with the same electric quantity q and different positions r. The attribute value $x_{i1} \sim x_{im}$ of data object X_i describes its position r in the electrostatic field. According to Equation (1), for data objects X_i and X_j in the m-dimensional electrostatic field, data object X_i receiving data object X_j's outlier Coulomb force $\overrightarrow{OF}(X_i, X_j)$ can be defined as

$$\begin{aligned}\overrightarrow{OF}(X_i, X_j) &= Eq'\vec{r} = \frac{k_m q^2}{\left\|X_i - X_j\right\|^{m-1}}\vec{r} \\ &= [f_1(i,j), \ldots, f_d(i,j), \ldots, f_m(i,j)]. \end{aligned} \quad (3)$$

Where $E = k_m q / \|X_i - X_j\|^{m-1}$ represents the electric field intensity at X_j in the electrostatic field data object X_i is a field source charge; m is the data dimension; $\vec{r}$ is a unit vector expressed as $(X_i - X_j)/\|X_i - X_j\|$; $\|X_i - X_j\|$ is the modulus length of the vector $X_i - X_j$; $f_d(i,j)$ is the component force of Coulomb force $\overrightarrow{OF}(X_i, X_j)$ in dimension d; q is the data object's electric quantity configured as a constant; k_m is an electrostatic constant of m-dimensional Coulomb force.

In an electrostatic field, the electric field intensity decays with the $m-1$ power of the distance; that is, as the difference among data objects increases, the Coulomb force's range becomes narrow. The smaller the difference among the data objects, the greater the interaction force. Let us put the concept of Coulomb force in the action of measuring outliers. As the discrepancy among data objects is widened, the Coulomb force (repulsion) are expected be lifted up, making a discernible range of the Coulomb force. To effectively measure the similarity of outlier data, we convert an electrostatic field into an abnormal electric field. The electric field intensity increases (relative to the attenuation) with distance; thus, electric field intensity $E = k_m q \|X_i - X_j\|^{m-1}$. Moreover, the values of constants k_m and q have little effect on similarity measures. For the convenience of calculation, we set constants as $k_m = 1$ and $q = 1$. Then we have

$$\overrightarrow{OF}(X_i, X_j) = Eq'\vec{r} = \left\|X_i - X_j\right\|^{m-1}\vec{r}. \quad (4)$$

Equation (4) suggests that the more similar of data objects X_i and X_j, the smaller of outlier Coulomb force $\overrightarrow{OF}(X_i, X_j)$; the more discrepancy between objects X_i and

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

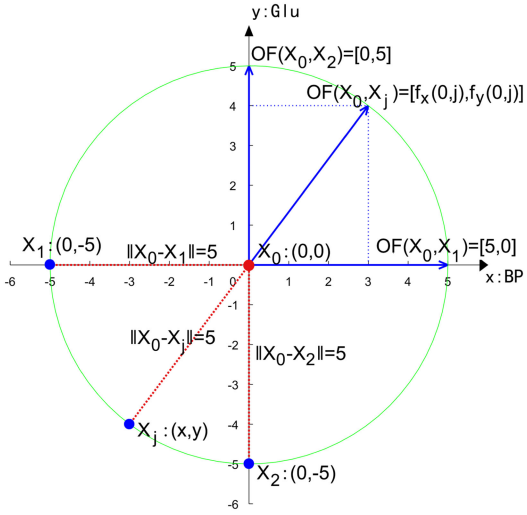


Fig. 2. Outlier Coulomb force measure on each dimension.

X_j, the larger of outlier Coulomb force $\overrightarrow{OF}(X_i, X_j)$. With an increase of the difference, the Coulomb force is increased at a faster pace. From the perspective of vectors, a marginal difference of data object X_i and X_j in dimension results in a small outlier Coulomb force $\overrightarrow{OF}(X_i, X_j)$ in corresponding dimension and vice versa. If and only if the difference is 0 of data object X_i and X_j in any dimension ($X_i = X_j$), the Coulomb force is 0 in any dimension. Further, let's consider data objects X_i, X_j, X_p. If we have $X_i \neq X_j$, then Coulomb force $\overrightarrow{OF}(X_p, X_i) \neq \overrightarrow{OF}(X_p, X_j)$ and is irrelevant to dimensions. Therefore, Coulomb force measure can be used to quantify high-dimension similarity, thereby effectively alleviating the interference of "concentration effect".

Recall that (see Section 3) euclidean distance cannot distinguish the difference of equidistant data objects. In essence, the definition of euclidean distance determines that it can only represent the overall discrepancy among data objects. That is, it accumulates the variance information in each dimension without difference, and the actual disparity in each dimension cannot be distinguished. In a dataset, each dimension (attribute) has a specific meaning; even data objects with the same total difference have distinctive meanings in various dimensions. We advocate for a full consideration of the intrinsic meaning of the corresponding dimensions of data objects to quantify the degree of similarity of the data objects.

Fig. 2 shows the discrepancy measure of the outlier Coulomb force on each dimension, where a solid point represents a data object, a blue arrow represents an outlier Coulomb force, dimension x denotes blood pressure (BP), and dimension y is blood glucose (Glu). Data object X_0 is located at the Coordinate origin, and the blue data object is located on the green circle with data object X_0 as the center and a radius of 5. Obviously, the differences between data objects and data object X_0 in blood glucose and blood pressure on the circle are not in common. For example, data objects X_1 and X_0's blood pressure values have a gap of 5 and their blood glucose values are identical; data objects X_2 and X_0's blood glucose values are different by 5 whereas the same blood pressures are the same. Nonetheless, the actual difference between X_1 and X_2 is significantly different. Under the euclidean distance metric, the euclidean distance is 5 between any data object on the green circle and X_0. That is, all the data objects on the green circle provide exactly the same outlier metric information for data object X_0. Such an outlier measure is undoubtedly unreasonable. In contrast, outlier Coulomb force of X_1 to X_0 is $\overrightarrow{OF}(X_0, X_1) = (5, 0)$, and the value is 5 in blood pressure dimension, meaning that the blood pressure value of the data object X_0 is greater than that of the data object X_1 by 5; its value is 0 in blood glucose dimension, indicating that blood glucose value of X_0 is the same as the value of X_1. Similarly, we have $\overrightarrow{OF}(X_0, X_2) = (0, 5)$, which implies that the blood pressure value of X_0 is the same as that of X_2, and the blood sugar value of X_0 is higher than that of X_2 by 5. That is, the value of the outlier Coulomb force $\overrightarrow{OF}(X_0, X_j)$ in the dimensional blood pressure and blood glucose precisely reflects the similar information of the data object X_0 relative to X_j in the dimensional blood pressure and blood glucose. For any data object X_j on the green circle, we have the outlier Coulomb force $\overrightarrow{OF}(X_0, X_j) = [f_{x(BP)}, f_{y(Glu)}]$, where value $f_{x(BP)}$ is difference between X_0 and X_j in blood pressure dimension. The positive value means that the blood pressure value of X_0 greater than X_j, and vice versa.

In summary, the outlier Coulomb force as a similarity measure can not only reflect the overall similarity of the data object, but also its projection in each dimension reflects the difference in the corresponding dimension of the data object. The reasonable use of the outlier Coulomb force as a high-dimensional similarity measure may effectively alleviate the "concentration effect".

4.2 Coulomb Resultant Force and Outlier Measures

Equations (2) and (4) suggest that in an electrostatic field constructed by data set D, the resultant force of data object X_i subjected to the Coulomb force from the other data objects $X_j \in D \setminus X_i$ is denoted as $\overrightarrow{OF_D}(X_i)$, which is calculated as

$$\begin{aligned}\overrightarrow{OF_D}(X_i) &= \sum_{X_j \in D \setminus X_i} \overrightarrow{OF}(X_i, X_j) \\ &= \sum_{X_j \in D \setminus X_i} \left\| X_i - X_j \right\|^{m-1} \vec{r} \\ &= [f_1(X_i), f_2(X_i), \ldots, f_d(X_i), \ldots, f_m(X_i)]. \end{aligned} \quad (5)$$

Where m is dimension of dataset; $f_d(X_i)$ is the value of outlier Coulomb resultant force $\overrightarrow{OF_D}(X_i)$ in dimension d, and can be computed as $\sum_{X_j \in D \setminus X_i} f_d(i, j), d = 1, 2, \ldots, m$.

In a dataset, a normal data object located in the dense area of a data center has little disparity from the other data objects. Its outlier Coulomb force has a small norm and the directions are scattered, so its outlier Coulomb resultant force is fairly small. An outlier data object that deviates in at least one dimension is quite distinctive from the other data objects. An outlier Coulomb force is large norm and the direction is concentrated, so its outlier Coulomb resultant force is large. Fig. 3 shows the outlier Coulomb force and the outlier Coulomb resultant force of data objects at various locations. In the figure, a solid point represents a data object; green X_0 is a normal data object; red X_5 is an outlier

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

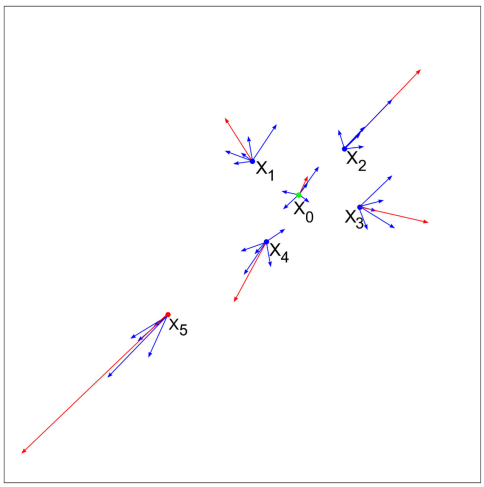


Fig. 3. Outlier Coulomb force and outlier coulomb resultant force.

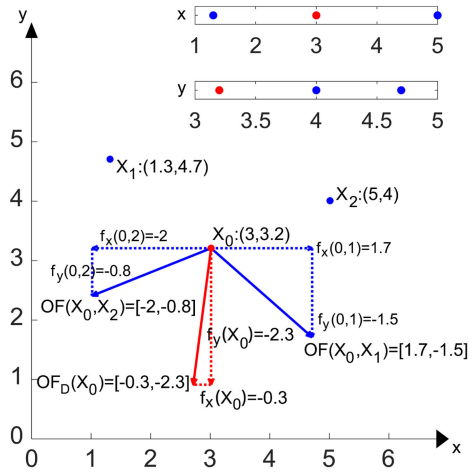


Fig. 4. Irrelevant dimensions and interpretable information.

data object; the blue arrow is an outlier Coulomb force; red arrow is an outlier Coulomb resultant force. Data object X_0 is located in the data center area, its outlier Coulomb force direction is scattered and small, and its outlier Coulomb resultant force is small. Data object X_5 that deviates from the data center receives a concentrated and large outlier Coulomb force direction, and its outlier Coulomb resultant force is large.

From a vector's point of view, outlier resultant Coulomb force $\overrightarrow{OF_D}(X_i)$ of data object X_i is the vector sum of the outlier Coulomb force of other data objects on data object X_i in the dataset. The component force (that is the value of corresponding dimension) of the outlier Coulomb resultant force in any dimension is the sum of the component force of each outlier Coulomb force in the corresponding dimension, which just resembles the degree to which the data object deviates from the data center in the corresponding dimension. The closer the data object is to the data center in dimension d; the smaller the component force of the outlier Coulomb force is in dimension d and it cancels out, so the component force of the outlier Coulomb resultant force is smaller in dimension d. If the data object is farther away from the data center in dimension d, the component of outlier Coulomb force will be strong in dimension d and it superimposes each other; thus, the component force of the outlier Coulomb resultant force is powerful in dimension d. Therefore, for any data object, the value of the outlier Coulomb resultant force in any dimension reflects the degree to which the data object deviates from the data center in the corresponding dimension, its can be used as a basis for analyzing the rationale behind the deviation of the data object. The more the dimension and the deviation of the data object, the greater the size of the outlier Coulomb resultant force. In this case, the data is more likely to be an outlier. Comparing to LRF [14], it takes fully the subtle different between the arbitrary neighbors of the object into consideration and makes the outlier coulomb resultant force has granularity advantage in alleviating the "Curse of Dimensionality".

Fig. 4 depicts the filtering on irrelevant dimensions and interpretable information of outlier Coulomb resultant force. A solid point represents a data object; a blue arrow denotes an outlier Coulomb force; a blue dotted arrows is a component force of outlier Coulomb force; red arrow represents an outlier Coulomb resultant force; red dotted arrow is a component force of outlier Coulomb resultant force. The upper right corner is the projection in dimension x and y of data object. It also shows that the outlier Coulomb resultant force is $\overrightarrow{OF_D}(X_0) = (-0.3, -2.3)$. In x-dimension, X_0 is located near the center of X_1 and X_2, its component force $f_x(0,1) = 1.7$ and $f_x(0,2) = -2$ in x-dimension are close in size and opposite in direction, so the component force of outlier Coulomb resultant force $\|f_x(X_0)\| = 0.3$ is smaller in x-dimension; in the y-dimension, X_0 is located on the same size of X_1 and X_2, its component force in y dimension is $f_y(0,1) = -1.5$, $f_y(0,2) = -0.8$, and the directions are consistent and superimposed on each other so component force of outlier Coulomb resultant force $\|f_x(X_0)\| = 2.3$ is greater in y dimension. Because $\|f_y(X_0)\| \gg \|f_x(X_0)\|$, y-dimension becomes a key reason for the deviation of the data object. And the direction of the outlier Coulomb resultant force is consistent with the direction of the data object deviating from the data center. That is, in x-dimension, X_0 is slightly deviated from the left side of the center of X_1 and X_2. Its value f_x (f_x is component force of outlier Coulomb resultant force in x-dimension) is negative and modulus is small. In the y-dimension, X_0 is significantly deviated from the below side of the center of X_1 and X_2. Value f_y is positive and modulus is large. After weakening the influence of irrelevant dimension x and amplifying the impact of the relevant dimension y, the size (modulus) of outlier Coulomb resultant force $\overrightarrow{OF_D}(X_0) \approx 2.32$ is total deviation of data object X_0 relative to the data center.

Amid outlier detection, there are a large number of irrelevant dimensions in the high-dimensional datasets. Due to the influence of irrelevant dimensions, the euclidean distance measures are interfered by the "Curse of Dimensionality" - a phenomenon of the distance tending to concentrate. The outlier Coulomb resultant force is the vector sum of the outlier Coulomb force, reflecting the deviation information of each dimension of the data object. Thus, a data object has a low degree of deviation and small component force in the irrelevant dimension. In the dimension pertaining to outlier information, the data object has high deviation degree and large

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

component force. A high dimension gives rise to rich deviation information. Essentially, the modulus of the outlier Coulomb resultant force is the total deviation of the data object after weakening the irrelevant dimension and amplifying the relevant dimension. The outlier Coulomb resultant force is slated to autonomously filter and weight dimensions, thereby exhibiting distinctive advantages compared with the local dimension reduction approaches. The local dimension reduction approaches such as the subspace-based schemes are implemented by losing partial information. Therefore, compared with the traditional paradigm distance, the outlier Coulomb resultant force is primed to effectively alleviate the impact of "the curse of dimensionality". Moreover, the outlier Coulomb resultant force has better applicability and generalization capability than the cosine similarity-based and the gravitation-based metrics.

In a nutshell, an outlier Coulomb resultant force reflects the deviation of data objects from a data center. Importantly, the Coulomb resultant force not only makes outliers interpretable but also alleviates the "Curse of Dimensionality" in high-dimensional datasets.

5 High-Dimensional Outlier Detection

5.1 Neighborhood Outlier Factor

The distribution of high-dimensional data is sparse and complex from a global perspective. It is arduous to detect strong and weak outliers because these outliers are likely to be treated as noise. Proximity-based methods believe that outlier data are significantly different from the other data objects in their neighborhood, and nearest-neighbor-based methods enjoy an advantage of a more refined granularity on outlier detection. k-neighborhoods of data object X_i can be defined as the set of k data objects $N_k(X_i)$ that have the smallest values among the norms of data objects' outlier Coulomb forces. According to Equation (5), the resultant force of a data object subject to the outlier Coulomb forces of its k-neighborhoods is expressed as a local outlier Coulomb force, denoted as $\overrightarrow{LCF_{N_k}}(X_i)$. Thus, we have

$$\overrightarrow{LCF_{N_k}}(X_i) = \sum_{j=1}^{k} \overrightarrow{OF}(X_i, X_j), X_j \in N_k(X_i). \tag{6}$$

In the k-neighborhood, the more data objects deviate from the center of its k-neighborhoods, coupled with a sparser k-neighborhood cluster, then the norm of the data object's outlier Coulomb force will be enlarged and the direction of its outlier Coulomb force will become increasingly consistent. Such a trend ultimately leads to a greater local outlier Coulomb force, vice versa. the greater its local outlier Coulomb force is and vice versa. The greater the value of the local Coulomb force is in any dimension d, the stronger the deviation from its k-neighborhoods is in dimension d. When the number of dimensions and the degree of deviation become large, the modulus of the local outlier Coulomb force will surge. Therefore, the local outlier Coulomb force reflects the degree of a data object deviating from its k-neighborhoods, and the value of each dimension resembles the deviation of the data object in the corresponding dimension, recorded as attribute deviation $AD(X_iw)$ expressed as

$$AD(X_iw) = f_w(X_i) / \sum_{d=1}^{m} |f_d(X_i)|. \tag{7}$$

Where $|\cdot|$ is an absolute value; m is data dimension; $f_w(X_i)$ is the value of local Coulomb force of data object X_i in dimension w. If $|AD(X_iw)| > \alpha * (1/m)$, attribute w is marked as a main deviation attribute. The positive value of $AD(X_iw)$ means that the attribute value of the data object in dimension w relative to the reference set is too large; a negative value implies a small attribute value. α is deployed as an input parameter to adjust explainable information in the NHOD algorithm. The value of α, which restricts the number of qualified dimensions, imposes no performance impact on the NHOD algorithm with respect to outlier detection. The number of output dimensions that embrace explainable information abates with an increasing value of α, and vice versa. In other words, the abundant deviation information will be offered by a low value of α, which makes it possible to obtain meaningful information of outliers. Additionally, users are equipped to stipulate the α value, which is in a window between 3 and 5 according to the degree of the preciseness on interpretable information.

Let the average size of the local Coulomb force of a data object and its k-neighborhoods be a *neighborhood outlier factor*, which is the average deviation degree of data object X_i and its k-neighborhoods X_j relative to X_j and its k-neighborhoods. Neighborhood outlier factor NOF is calculated as

$$NOF(X_i) = \frac{1}{k+1} \sum_{X_j \in N_k(X_i) \bigcup X_i} \left\| \overrightarrow{LCF_{N_k}}(X_j) \right\|. \tag{8}$$

Where k is the number of neighbors.

The neighborhood outlier factor expands the effective area of a data object's outlier measure, which reflects the average deviation degree between a given data object and its k neighborhoods in respective KNN reference domains, where k signifies the number of neighbors. To discriminate the impact of sorting outliers in various neighborhood areas, the k value is kept as a constant across all the data objects. For outliers deviating in sparse areas, their k-neighborhoods are sparse and deviating from the respective k-neighborhoods. If the average value of the local Coulomb force of outlier data objects is high, so does the neighborhood outlier factor. When it comes to normal data objects undeviating in dense areas, their k-neighbors are dense and centering relative to their respective k-neighborhoods. In this case, the average value of the local Coulomb force of normal data objects is small, leading to a small neighborhood outlier factor. Therefore, the neighborhood outlier factor (NOF) is slated to measure the deviation degree of outliers.

5.2 NOF-Based Outlier Detection

Now we articulate the basic steps of high-dimensional outlier detection using Neighborhood outlier factor. The detection process starts by performing min-max normalization on the data, followed by creating k-neighborhood index table T for all data objects from the dataset D (see Setp 1). Next, the outlier Coulomb force of each data object and its nearest neighbors are calculated according to Equation (4)

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

in Steps 4-5. In Step 7, the local Coulomb force of the each data object is computed using Equation (6). Step 10 applies Equation (8) to quantify the neighborhood outlier factor of each data object. The neighborhood outlier factor of all data objects are sorted in descending order to get table $NOF_{Var}List$ (see Step 12), where the Top-N data objects with the largest neighborhoods outlier factor are elected to be marked as outliers (see Step 13). Steps 16-19 employs Equation (7) to calculate the attribute deviation degree AD of the Top-N outliers, followed by emitting the deviation attribute. The pseudocode of the NHOD algorithm is described as Algorithm 1.

Algorithm 1. NHOD : High-Dimensional Outlier Detection Based on Neighborhood Outlier Factor

Input: Dataset D, Neighborhood size k, AD threshold α.
Output: Outliers and its outlier attributes.
Start NHOD
1: Create k-neighborhoods index table T for all data objects from the dataset D;
2: **for all** data object X_i in D **do**
3: **for** j in 1 to k **do**
4: Find the j-th nearest neighbor X_j form T;
5: According to Equation (4), Calculate the $\overrightarrow{OF}(X_i, X_j)$;
6: **end for**
7: According to Equation (6), Calculate the $\overrightarrow{LCF_{N_k}}(X_i)$;
8: **end for**
9: **for all** data object X_i in D **do**
10: According to Equation (8), Calculate the $NOF(X_i)$;
11: **end for**
12: Sort $NOF(X_i)$ in descending order to obtain the $NOF_{var}List$;
13: The top-n largest data objects in $NOF_{Var}List$ are marked as $outliers$,and put them into $outlierlist$;
14: **for all** outlier data object X_i in $outlierlist$ **do**
15: **for all** dimension w **do**
16: According to Equation (7), Calculate the $AD(X_iw)$;
17: **if** $AD(X_iw) > \alpha * (1/m)$ **then**
18: Mark w as the outlier attribute of data X_i;
19: **end if**
20: **end for**
21: **end for**
End NHOD

Complexity Analysis of the NHOD Algorithm. The time complexity of finding k-nearest neighbor is $O(n^2)$. The time complexity of calculating k outlier Coulomb forces for each data object is $O(kn)$. The time complexity of computing local Coulomb force is $O(kn)$. It takes $O(kn)$ to determine the neighborhoods outlier factor. The time complexity of assessing the attribute deviation factor is $O(mN)$, which can be ignored. The algorithm's total time complexity is $O(n^2 + 3kn)$, where $k \ll n$. Therefore, the overall time complexity of the NHOD algorithm is $O(n^2)$.

6 Experimental Analysis

6.1 Datasets and Evaluation Index

We make use of both synthetic and UCI high-dimensional datasets to evaluate the NHOD outlier detection algorithm driven by neighborhood outlier factor catering for high-dimensional data. Apart from the NHOD algorithm, the other competitive outlier data mining methods tested in our experiments include ABOD [18], LOF [19], LGOD [29], MOD$^+$ [27], RDOS [15], SOD [24]. More specifically, MOD$^+$ utilizes the parameter I to denote the number of iterations, the initial value of which is set to 3. SOD involves the size l of reference sets and the significance of an attribute α, where l and α are specified to $k-2$ and 0.8 by the source, respectively. RDOS orchestrates the Gaussian kernel to yield an optimized bandwidth of h$\approx \widehat{\sigma} n^{-1/5}$, where $\widehat{\sigma}$ represents the standard deviation of a dataset. A total of eight synthetic datasets [38] and six real-world UCI high-dimensional datasets. UCI[1] and ODDS[2] are used as empirical data. More specifically, the original MUSK dataset embraces several musk and non-musk classes. The non-musk classes are combined to forge data inliers, whereas the musk classes are assembled as data outliers without down sampling. The arrhythmia dataset is a multi-class classification dataset, in which the smallest classes are blended to construct the outlier class and the rest of the classes are merged to form the inlier class. The MNIST dataset is comprised of handwritten digits that are size-normalized and centered in a fixed-size image. In this dataset, digit-zero class is considered as inliers, whereas the images sampled from the digit-six class are adopted as outliers. The Human Activities and Postural Transitions (HAPT) dataset contain six physical activities coupled with postural transitions that occurred between the static postures. Thus, we regard the six physical activities as nominal transitions while treating the postural transitions as the anomalous state. The InternetAds dataset consists of images that are classified as ads or non-ads from web pages. The goal is to automatically remove ads from web pages while retaining regular images. In this study, the Ads are envisioned as data outliers. The KDDCup99 dataset captures different types of network intrusions or attacks. We adopt U2R as data outliers and normal ones as data inliers. The information about the synthetic datasets is tabulated in Table 1; the summary of the UCI datasets is shown in Table 2. Accuracy and ROC_AUC [39] are applied as the outlier data mining performance metrics. Min-max normalization is the key for data preprocessing.

TABLE 1
Synthetic Datasets

Dataset	Attributes	Instances	Outliers
DS1	2	1043	43
DS2	2	876	77
DS3	2	1372	72
DS4	2	1037	37
DS5	2	1034	36
DS6	2	2042	64
DS5	2	1020	26
DS6	2	1242	50

A testbed furnishing the experimental setup is powered by the Linux operating system running on NVIDIA GeForce GTX 850M CPU with 8GB main memory. We implement the outlier detection algorithms using Python as a programming language.

1. http://archive.ics.uci.edu/ml/index.php
2. http://odds.cs.stonybrook.edu/

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
UCI Datasets

Dataset	Attributes	Instances	Outliers
KDDCup99	41	60839	246
Mnist	100	7603	700
Musk	166	3062	97
Arrhythmia	274	452	66
HAPT	561	10929	518
PInternetAds	1556	3224	454

6.2 Parameter k

To verify the influence of different parameter k on the performance of the NHOD algorithm, our use the six UCI datasets as an input. Table 3 outlines the experimental results in terms of AUC, Accuracy(ACC), and efficiency(s).

The Table 3 reveals that with an increasing k, the AUC and accuracy measures gradually climb. When k is approaching $\sqrt{m}$, the AUC and accuracy of NHOD's detection results tend to be stable. For example, when the AUC and Accuracy results are flattened, the k values of datasets *Musk*, *PInternetAds*, and *Mnist* become $k = 50$ ($\sqrt{m} = 55$), $k = 65$ ($\sqrt{m} = 57$), and $k = 80$ ($\sqrt{m} = 87$). This trend is expected because as the number of k-nearest neighbors rises, the dimensionality of the local Coulomb force is stretching out. In case of abundant dimensional deviation information provided, the neighborhood outlier factor progressively approaches a true deviation. When the number of neighbors reaches $\sqrt{m}$, the newly added data in the neighborhood are insufficient to offer adequate outlier information to change the neighborhood outlier factor.

Table 3 shows that the effect of parameter k to the NHOD algorithm on the efficiency is nebulous. As the value of k increases, NHOD only ends up with one neighbor for each data object.In this case, the operational efficiency of the NHOD algorithm is insignificantly affected by parameter k.

6.3 Synthetic Datasets

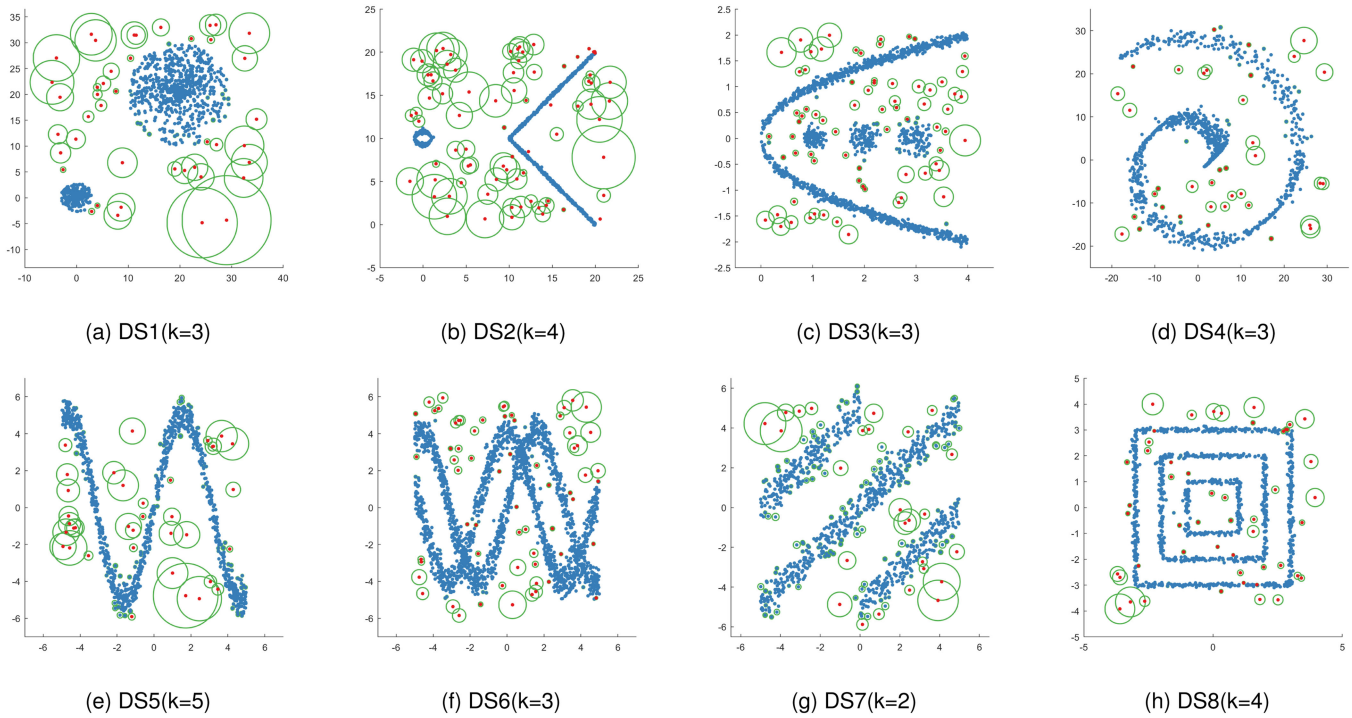
We rely on eight synthetic datasets to assess the accuracy of NHOD under various data distributions, the results of which are plotted in Fig. 5. The data objects identified as normal data are marked in blue; data identified as outliers are noted in red. The size of the green circle mimics the size of the neighborhood outlier factor of a data object.

Fig. 5 unveils that the NHOD algorithm is capable of completely and correctly identifying outliers under various complex data distributions such as datasets DS1, DS2, DS3, DS5, DS6, DS7. Individual outliers wrapped by ordinary data in datasets DS4 and DS8 are overlooked by NHOD. The reason is two-fold. First, the data objects within the k neighborhood are almost symmetrically distributed around the outliers. Given such a data distribution, the outlier Coulomb force acting on an outlier cancel out each other, thereby making the local outlier Coulomb force weak. Second, the k-nearest neighbors of an outlier data object are all normal data coupled with light local outlier Coulomb force. Therefore, the neighborhood factor of these outlier data objects is small. Hence, these two reasons make it futile, if not impossible, to detect outliers from datasets DS4 and DS8.

TABLE 3
UCI Datasets Outliers Detection of NHOD

	k	Mnist	Musk	KDDC up99	PInter netAds	HAPT	Arrhy thmia
ACC	5	0.8932	0.9406	0.9931	0.8352	0.9543	0.8363
	20	0.9008	0.9732	0.9937	0.8480	0.9561	0.8496
	35	0.9006	0.9895	0.9936	0.8545	0.9555	0.8584
	50	0.9000	1.0000	0.9934	0.8575	0.9555	0.8540
	65	0.8990	1.0000	0.9932	0.8606	0.9524	0.8761
	80	0.8964	1.0000	0.9931	0.8539	0.9593	0.8628
	95	0.8937	1.0000	0.9954	0.8511	0.9564	0.8628
	110	0.8935	1.0000	0.9954	0.8480	0.9575	0.8496
AUC	5	0.8232	0.6323	0.8131	0.7961	0.9691	0.7943
	20	0.8587	0.9815	0.9314	0.7654	0.9730	0.8132
	35	0.8757	0.9988	0.9561	0.7321	0.9731	0.8254
	50	0.8795	1.0000	0.9649	0.7277	0.9728	0.8286
	65	0.8823	1.0000	0.9800	0.7139	0.9714	0.8311
	80	0.8820	1.0000	0.9857	0.7075	0.9718	0.8230
	95	0.8828	1.0000	0.9898	0.7071	0.9719	0.8118
	110	0.8812	1.0000	0.9895	0.7055	0.9721	0.8027
Time (s)	5	1.4703	0.9855	6.3509	9.7709	10.92	0.1873
	20	2.1775	1.3422	9.1654	14.524	16.51	0.3032
	35	2.6805	1.7672	12.131	18.380	21.24	0.3863
	50	3.3194	2.0127	14.724	22.266	26.45	0.5150
	65	4.0035	2.4057	17.671	25.989	32.30	0.5634
	80	4.5491	2.7851	20.914	30.142	37.17	0.6613
	95	5.4666	3.2721	23.073	34.820	44.99	0.8465
	110	6.4843	3.8794	25.729	38.854	51.80	0.8841

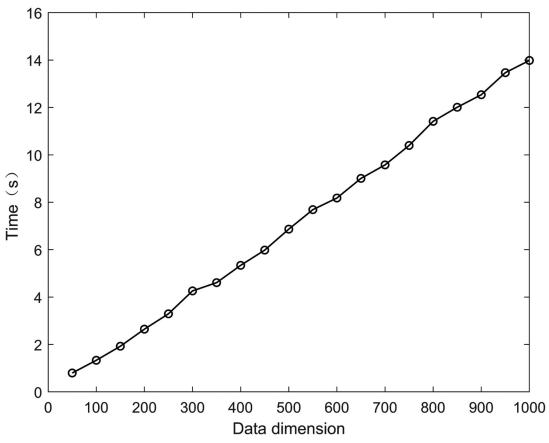
Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.



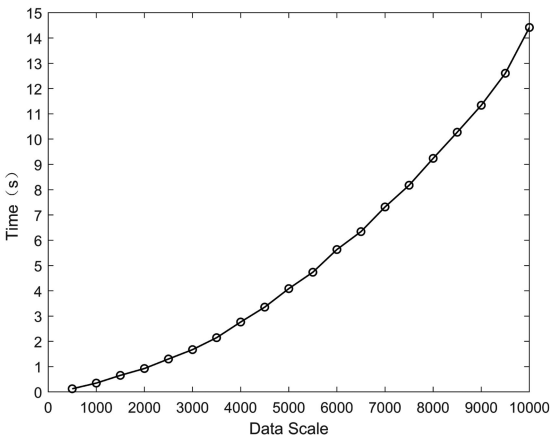
(a) DS1(k=3) (b) DS2(k=4) (c) DS3(k=3) (d) DS4(k=3)

(e) DS5(k=5) (f) DS6(k=3) (g) DS7(k=2) (h) DS8(k=4)

Fig. 5. Detection result of the algorithm in the synthetic data sets.



(a) Data dimension



(b) Data scale

Fig. 6. The impact of data dimension and scale on the efficiency.

6.4 Data Dimension and Data Scale

To investigate the impact of data dimensions on NHOD's efficiency, we vary the number of dimensions in the UCI's PInternetAds data. The NHOD's efficiency in the PinternetAds case is plotted in Fig. 6a. Similary, we manipulate the number of data objects in the UCI's KDDCup99 dataset to study the impact of data size on the efficiency. Fig. 6b unravels the NHOD's efficiency performance in the KDDCup99 scenario.

Fig. 6a shows that the NHOD algorithm's detection time almost linearly goes up with an increase of data dimensions. As the dimensionality expands, the vector operation time in NHOD is linearly enlarged. Therefore, the efficiency of NHOD is not exponentially affected by dimensionality, suggesting that NHOD is adept at processing high-dimensional datasets. Fig. 6b illustrates that the detection time of NHOD surges in a non-linear manner with an increasing data size. When data size grows, the similarity k-neighbor search time is non-linearly extended during the course of constructing a nearest-neighbor-index table T.

6.5 UCI High-Dimensional Datasets

Now we are positioned to evaluate the detection accuracy of NHOD, which is compared against the six mainstream outlier detection algorithms handling the six UCI datasets. The experimental and comparison results are tabulated in Table 4.

Table 4 unveils that the NHOD algorithm exhibits a clear edge over the other competitive algorithms in the cases of the six UCI high-dimensional datasets. NHOD's AUC measures are the highest in the six test cases among all the competitors. When it comes to ACC

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

TABLE 4
Comparison of AUC and Accuracy

	Algorithm	Mnist k=88	Musk k=55	PInter netAds k=57	KDDC up99 k=247	Arrhy thmia k=21	HAPT k=105
AUC	NHOD	88.27	100.0	72.18	98.93	81.33	97.21
	LGOD	63.08	70.95	67.89	81.33	69.96	80.09
	LOF	79.24	35.47	62.60	68.84	75.92	53.25
	RDOS	74.25	44.12	71.78	86.16	76.41	66.13
	MOD^+	84.13	96.25	59.86	98.42	79.70	94.92
	ABOD	36.53	3.447	53.77	75.44	76.84	92.73
	SOD	55.21	88.16	59.49	94.81	71.20	89.97
ACC	NHOD	89.37	100.0	86.55	99.51	84.96	95.70
	LGOD	85.95	95.56	81.07	99.40	81.42	93.41
	LOF	88.19	93.86	80.30	99.19	83.19	90.92
	RDOS	87.90	94.12	80.82	99.57	83.19	92.19
	MOD^+	88.93	95.89	78.92	99.47	84.96	95.17
	ABOD	82.06	93.66	76.35	99.60	77.53	94.22
	SOD	85.03	94.58	79.53	99.28	80.53	93.38

readings, NHOD is a winner in five out of the six scenarios. In particular, the best results of NHOD are obtained from the PInternetAds dataset - the 1556 dimensional UCI dataset.

The results under a single parameter k value may not fully reflect the intrinsic detection performance of all the algorithms. To address this concern, we vary parameter k and gauge the overall performance of NHOD and its alternatives driven by the six UCI high-dimensional datasets. Fig. 7 plots the AUC and accuracy measures of all the seven outlier detection algorithms.

Fig. 7 illustrates that the outlier detection results of our NHOD on the six UCI high-dimensional datasets are superior to those of the other algorithms. For instance, the AUC and accuracy readings of NHOD on the datasets MUSK, Mnist, HAPT and Arrhythmia are significantly higher than those of the other six contenders. In the PInternetAds scenario, our NHOD's accuracy performance is consistently

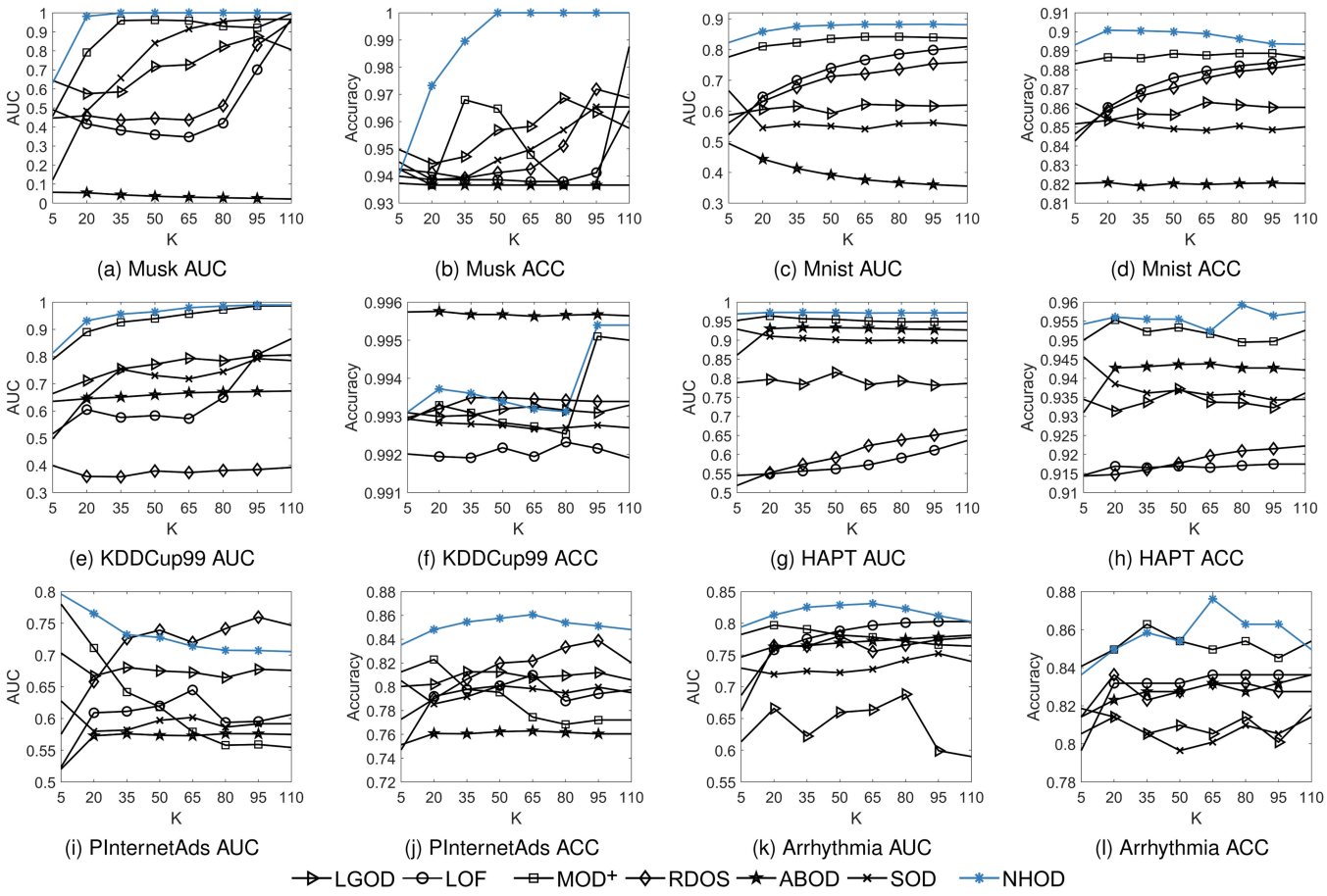


Fig. 7. AUC and ACC curves of 7 algorithms on UCI data sets.

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

TABLE 5
Operating Efficiency (Second)

Data	k	NHOD	LGOD	LOF	RDOS	MOD$^+$	ABOD	SOD
Mnist	20	2.177507	2.868339	2.245455	247.6401	57.54033	5.457108	132.7187
	50	3.319377	7.390770	2.861935	655.5250	74.42497	29.57536	275.8146
	80	4.549132	8.29132	3.580716	896.7818	98.38920	86.14654	275.3254
	110	6.484384	13.09863	4.050871	1224.564	119.2220	187.8086	311.6721
Musk	20	1.342234	1.233319	1.266015	68.48176	19.81137	3.921095	15.41665
	50	2.012786	2.217966	1.612974	163.8340	32.90055	19.00353	28.60350
	80	2.785114	3.109390	2.087422	255.7436	44.70804	54.23021	46.32081
	110	3.879442	3.621915	2.380119	366.0291	57.70655	104.1743	53.68291
Arrhythmia	20	0.303236	0.410921	0.237109	16.98327	2.191044	0.845630	0.635928
	50	0.515040	0.679534	0.242470	37.71199	4.847150	4.047778	0.913496
	80	0.661307	1.052483	0.311947	62.46220	7.424192	9.979809	1.175369
	110	0.884188	1.222829	0.325049	79.13076	10.18690	18.57160	1.404357
HAPT	20	16.50771	16.57615	12.62589	491.5706	422.2832	39.43093	157.9644
	50	26.44543	33.18015	17.37579	1096.184	564.3697	191.1373	228.5061
	80	37.16597	42.19232	23.40660	1992.361	678.4121	470.5699	386.3448
	110	51.80267	67.70377	27.26643	2206.133	810.8721	879.1067	456.1787
KDDCup99	20	9.165429	12.78756	5.545149	1741.313	1272.072	33.83005	4297.303
	50	14.72409	19.33905	8.414259	3096.946	1207.378	189.6546	6064.178
	80	20.91463	27.59461	11.04316	5099.771	1318.015	478.2393	10305.34
	110	25.73000	35.21548	13.93077	6690.433	1411.122	915.8654	12925.63
PInternetAds	20	14.52464	8.709991	12.59053	361.1563	183.8378	31.12583	33.12030
	50	22.26658	16.80934	15.86165	758.2603	316.5951	157.9205	41.41011
	80	30.14293	24.98660	19.26432	1018.732	457.3972	364.7743	61.66943
	110	38.85441	35.01385	22.58502	1696.061	583.9409	657.3483	69.47249

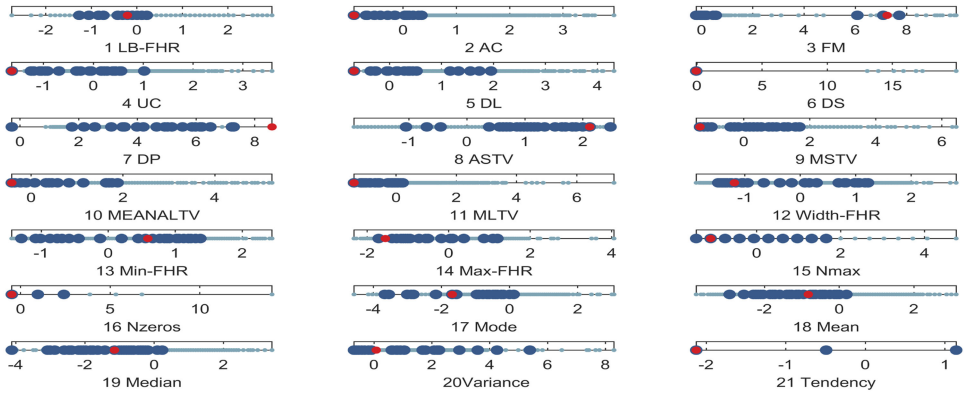
superior to those of the other contenders, whereas the AUC measure achieves a maximum value. In the KDDCup99 case, the AUC of NHOD is constantly higher than those of the competitors; the accuracy of NHOD is a competitive runner-up following the ABOD algorithm. The outlier detection results of NHOD are immensely better than those of the alternative solutions such as LOF, RDOS, and MOD$^+$ powered by the euclidean distance measurement. These existing algorithms are adversely affected by the "Curse of Dimensionality" - a phenomenon of the distance tending to concentrate. Our NHOD mitigates this problem by applying local Coulomb force to measure the deviation of data objects, thereby weakening the scale of dimensions that are not related to outlier information. Meanwhile, NHOD expands the scale of dimensions that contain outlier information; as a result, NHOD effectively alleviates the interference of "Curse of Dimensionality". LGOD algorith employs vector measurement m while ignoring the most important directionality of the vector. The high-dimensional outlier detection algorithm ABOD uses cosine similarity as a similarity measure, which has a "concentration effect". ABOD ignores the distance information among data objects. The subspace-based method SOD is inadequate to detect outlier data objects distributed in two or more subspaces. Therefore, our experimental results conclude that the NHOD algorithm remarkably outperform all the tested alternative algorithms in terms of detecting outliers from high-dimensional datasets.

Table 5 compares the efficiency of NHOD against the six mainstream outlier detection algorithms. The results confirm that NHOD has a detection efficiency edge over the RDOS, ABOD, MOD$^+$ and SOD algorithm. NHOD's detection efficiency performance is well aligned with those of the LOF and LGOD algorithms. Moreover, NHOD is slightly worse than LOF and is marginally better than LGOD. The majority of time consumed in NHOD, LOF, and LGOD is concentrated on finding k nearest neighbors. The RDOS algorithm needs to spend an excessive amount of time in finding k nearest neighbors, reverse neighbors, as well as shared neighbors. ABOD is also inefficient because it ought to calculate the cosine similarity between any two vectors in the k neighborhood; MOD$^+$'s inefficiency is due to an iterative search for k nearest neighbors; the time consumed to find a subspace in the SOD algorithm is an exponential order of the dimensions, making SOD an inefficient solution.

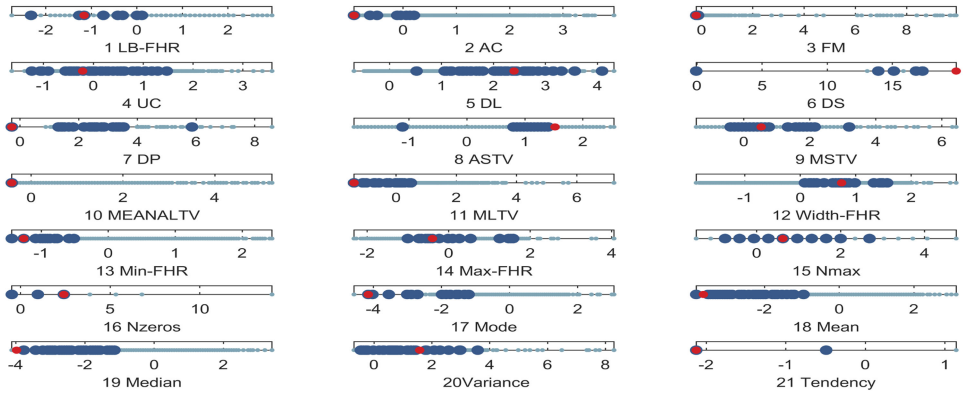
6.6 Interpretability

Recall that (see Sections 4.2 and 5.1) the NHOD algorithm not only can quickly and effectively locate outlier data objects in high-dimensional datasets, but also offers good interpretability - pinpointing the causes of abnormal data objects. To verify the interpretability of NHOD, we undertake the experiments using the UCI Cardio dataset, which contains 21 attributes, 1831 diagnosis records and 176 patients. UCI Cardio consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by the expert obstetricians. The specific meanings of the 21 attributes are LB-FHR baseline (beats per minute), AC-accelerations per second, FM-fetal movements per second, UC-uterine contractions per second, DL-light decelerations per second, DS-severe decelerations per second, DP-prolongued decelerations per second, ASTV-percentage of time with abnormal short

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.



(a) Patient 1723



(b) Patient 1782

Fig. 8. The projections of patients 1723 and 1782 and their reference sets in each dimension.

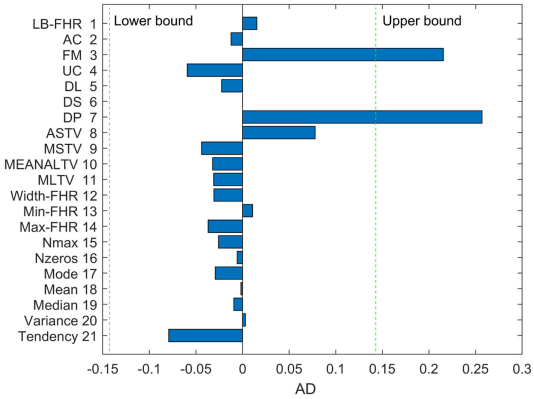
term variability, MSTV-mean value of short term variability, ALTV-percentage of time with abnormal long term variability, MLTV-mean value of long term variability, Width-width of FHR histogram, Min-minimum of FHR histogram, Max-Maximum of FHR histogram, Nmax-histogram peaks, Nzeros-histogram zeros, Mode-histogram mode, Mean-histogram mean, Median-histogram median, Variance-histogram variance, Tendency-histogram tendency.

When we have k=43 and α=3, the NHOD algorithm ranks the neighborhood outlier factors of 1831 diagnostic records. Among the top 176 diagnostic records, the attribute values of patients 1723 and 1782 with the highest degree of outlier (the largest neighbor outlier factor) are $obj1723$=[-0.2067, -0.9200, 7.2373, -1.6359, -0.6836, -0.0614, 8.5873, 2.1186, -1.3255, -0.4205, -1.3981, -1.1881, 0.5915, -1.5554, -1.0974, -0.4883, -1.6823, -0.8285, -1.1478, 0.0776, -2.1266], $obj1782$=[-1.1590, -0.9199, -0.2036, -0.2160, 2.3951, 19.9003, -0.2782, 1.5203, 0.5274, -0.4204, -1.3980, 0.7446, -1.2616, -0.4003, 0.6226, 2.4144, -4.1470, -3.6377, -3.9755, 1.5806, -2.1266]. Fig. 8a shows the projections of the attribute values of patient 1723 and its reference set in each dimension. Fig. 8b depicts the projections of the attribute values of patient 1782 and its reference set in each dimension. Fig. 9 illustrates the attributes deviation degree of the patient 1723 and the patient 1782 in their corresponding dimensions.

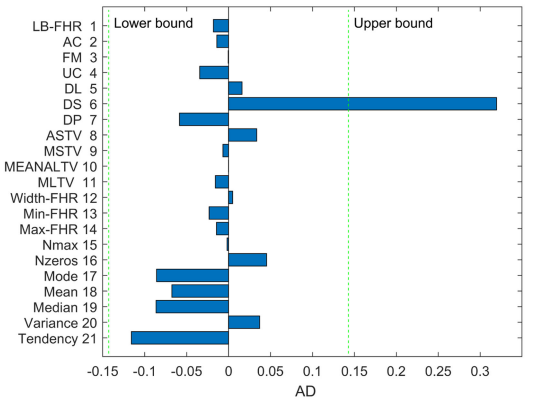
Fig. 8 unfolds the projections of patients 1723 and 1782 and their reference sets in each dimension. The red points are the projections of patient 1723 and patient 1782 observations; the dark blue points represent the projection of the reference set observations. The results demonstrate that the observed values of attributes 3 and 7 of patient 1723 are significantly greater than those of the other diagnostic records in the reference set, whereas the deviation of other attributes from the diagnostic records in the reference set is opaque. Moreover, the observed values of attributes 6 of patient 1782 are unequivocally greater than those of the other diagnostic records in the reference set, whereas the deviation of the other attributes from the diagnostic records in the reference set is obscure.

Fig. 9 shows the degree of deviation AD of the attributes of patients 1723 and 1782, which the abscissa is the value of the attribute corresponding to AD, and the dashed line is the attribute selection limit. The results suggest that the degree of deviation of patient 1723 on the selected attributes 3 and 7 is large and positive, and the degree of deviation of other dimensions is insignificant, which is basically the same as the actual deviation of patient 1723 plotted in Fig. 8a. Therefore, the observation values of patient 1723 in attributes 3 and 7 are used as the main root of the patient, and it can be seen that the observation values are larger than normal. The degree of deviation of 1782 in the selected attribute 6 is large and positive, and the degree of deviation in other dimensions is negligible, which is essentially identical to the actual deviation of 1782 illustrated in Fig. 8b.

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.



(a) Patient 1723



(b) Patient 1782

Fig. 9. Projection length of the local Coulomb force on each dimension.

Therefore, the observation value of patient 1782 in attribute 6 can be used as the main explanation of the patient, and the results indicate that the observation value is larger than the normal level. In a nutshell, the attribute deviation degree AD can be used as an effective basis for analyzing the reasons for the deviation of outliers.

7 Conclusion and Future Work

Detecting outliers from high-dimensional data has been a challenging and rewarding research direction in the field of data mining. We mixed the characteristics of outlier data objects with the physics-related concepts in a distinctive and valuable way, converting data objects into electric charges in the electrostatic field. We proposed a novel similarity measurement approach anchored on Coulomb force catering for high-dimensional data. Compared with the traditional distance measures, our similarity measure is adroit at effectively alleviating the "concentration effect". Importantly, we introduced a new high-dimensional outlier data object neighborhood outlier factor, based on which a high-dimensional outlier detection algorithm - NHOD - was devised. This algorithm is conducive to mitigating the interference of "curse of dimensionality", thereby offering high detection efficiency and interpretability. Recognizing that the NHOD algorithm is insensitive to outlier data objects located in a symmetric center of normal data, we plan to embark on a sensitivity analysis of outlier data objects in the center of symmetry. A second research direction will be the development of a parallel outlier detection algorithm impelled by the Coulomb force concept for high-dimensional datasets.

References

[1] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proc. 24rd Int. Conf. Very Large Data Bases*, 1998, pp. 392–403.
[2] M. Safaei *et al.*, "A systematic literature review on outlier detection in wireless sensor networks," *Symmetry*, vol. 12, no. 3, 2020, Art. no. 328.
[3] E. W. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, 2011.
[4] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia-Social Behav. Sci.*, vol. 62, pp. 989–994, 2012.
[5] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, 2020, Art. no. 2809.
[6] T. Andrysiak, "Sparse representation and overcomplete dictionary learning for anomaly detection in electrocardiograms," *Neural Comput. Appl.*, vol. 32, no. 5, pp. 1269–1285, 2020.
[7] D. Jin, Y. Lu, J. Qin, Z. Cheng, and Z. Mao, "SwiftIDS: Real-time intrusion detection system based on LightGBM and parallel intrusion detection mechanism," *Comput. Secur.*, vol. 97, 2020, Art. no. 101984.
[8] B. Wang and Z. Mao, "Detecting outliers in industrial systems using a hybrid ensemble scheme," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 8047–8063, 2020.
[9] J. Zhang, S. Zhang, K. H. Chang, and X. Qin, "An outlier mining algorithm based on constrained concept lattice," *Int. J. Syst. Sci.*, vol. 45, no. 5, pp. 1170–1179, 2014.
[10] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statist. Anal. Data Mining: ASA Data Sci. J.*, vol. 5, no. 5, pp. 363–387, 2012.
[11] H. J. Eom, "Electrostatics," in *Primary Theory of Electromagnetics*. Berlin, Germany: Springer, 2013, pp. 21–57.
[12] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 873–886, Jul. 2007.
[13] M. Radovanović, A. Nanopoulos, and M. Ivanović, "On the existence of obstinate results in vector space models," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 186–193.
[14] Z. Wang *et al.*, "Clustering by local gravitation," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1383–1396, May 2018.
[15] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, 2017.
[16] M. Goldstein and A. Dengel, "Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm," in *Proc. 35th German Conf. Artif. Intell.*, 2012, pp. 59–63.
[17] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 145–160, Feb. 2006.
[18] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2008, pp. 444–452.
[19] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.
[20] F. Wang and X. Zhang, "Adaptive finite time control of nonlinear systems under time-varying actuator failures," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 9, pp. 1845–1852, Sep. 2019.
[21] A. Elmogy, H. Rizk, and A. M. Sarhan, "OFCOD: On the fly clustering based outlier detection framework," *Data*, vol. 6, no. 1, 2021, Art. no. 1.
[22] E. Bigdeli, M. Mohammadi, B. Raahemi, and S. Matwin, "A fast and noise resilient clustter-based anomaly detection," *Pattern Anal. Appl.*, vol. 20, no. 1, pp. 183–199, 2017.
[23] J. Zhang, X. Yu, Y. Xun, S. Zhang, and X. Qin, "Scalable mining of contextual outliers using relevant subspace," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 3, pp. 988–1002, Mar. 2020.

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.

[24] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, 2009, pp. 831–838.
[25] A. Smiti, "A critical overview of outlier detection methods," *Comput. Sci. Rev.*, vol. 38, 2020, Art. no. 100306.
[26] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier detection: Methods, models, and classification," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–37, 2020.
[27] J. Yang, S. Rahardja, and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognit.*, vol. 115, 2021, Art. no. 107874.
[28] Y. Dong, S. Hopkins, and J. Li, "Quantum entropy scoring for fast robust mean estimation and improved outlier detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 545.
[29] J. Xie, Z. Xiong, Q. Dai, X. Wang, and Y. Zhang, "A local-gravitation-based method for the detection of outliers and boundary points," *Knowl.-Based Syst.*, vol. 192, 2020, Art. no. 105331.
[30] J. Li, J. Zhang, X. Qin, and Y. Xun, "Feature grouping-based parallel outlier mining of categorical data using spark," *Inf. Sci.*, vol. 504, pp. 1–19, 2019.
[31] C. C. Aggarwal and S. Y. Philip, "An effective and efficient algorithm for high-dimensional outlier detection," *VLDB J.*, vol. 14, no. 2, pp. 211–221, 2005.
[32] J. Zhang, Y. Jiang, K. H. Chang, S. Zhang, J. Cai, and L. Hu, "A concept lattice based outlier mining method in low-dimensional subspaces," *Pattern Recognit. Lett.*, vol. 30, no. 15, pp. 1434–1439, 2009.
[33] X. Zhao, J. Zhang, and X. Qin, "LOMA: A local outlier mining algorithm based on attribute relevance analysis," *Expert Syst. Appl.*, vol. 84, pp. 272–280, 2017.
[34] X. Zhao, J. Zhang, X. Qin, J. Cai, and Y. Ma, "Parallel mining of contextual outlier using sparse subspace," *Expert Syst. Appl.*, vol. 126, pp. 158–170, 2019.
[35] E. Müller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *Proc. IEEE 27th Int. Conf. Data Eng.*, 2011, pp. 434–445.
[36] J. Zhang, X. Yu, Y. Li, S. Zhang, Y. Xun, and X. Qin, "A relevant subspace based contextual outlier mining algorithm," *Knowl.-Based Syst.*, vol. 99, pp. 1–9, 2016.
[37] S. Theisen, "Hunting for extra dimensions," in *Einstein Online Band 04*, 2010, pp. 01–1027. [Online]. Available: https://www.einstein-online.info/en/spotlight/extra_dimensions_force/
[38] J. Ha, S. Seok, and J.-S. Lee, "A precise ranking method for outlier detection," *Inf. Sci.*, vol. 324, pp. 88–107, 2015.
[39] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.



Pengyun Zhu received the BS degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2019. He is currently working toward the PhD degree in the School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China. His current research interests include data mining and parallel computing.



Chaowei Zhang received the PhD degree in computer science and software engineering from the Auburn University, Auburn, Alabama, in 2021. He is currently with the Department of Computer Science, Yangzhou University. His research interests include natural language processing, data mining, storage systems, and parallel computing.



Xiaofeng Li received the MS degree in computer science and technology from Southwest Jiaotong University, Chendu, China, in 2004. She is currently working toward the PhD degree in big data and intelligent manufacturing with the Taiyuan University of Science and Technology, Taiyuan, China. Her current research interests include data mining and machine learning.



Jifu Zhang (Member, IEEE) received the BS and MS degrees in computer science and technology from the Hefei University of Tchnology, Hefei, China, 1983 and 1989, respectively, and the PhD degree in pattern recognition and intelligence systems from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a professor with the School of Computer Science and Technology, Taiyuan University of Science and Technology. His research interests include data mining, parallel and distributed computing, and artificial intelligence.



Xiao Qin (Senior Member, IEEE) received the BS and MS degrees in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 1992 and 1999, respectively, and the PhD degree in computer science from the University of Nebraska-Lincoln, Lincoln, Nebraska, in 2004. He is currently an Alumni professor and the director of Graduate Programs with the Department of Computer Science and Software Engineering, Auburn University. Prior to joining Auburn University in 2007, he had been an assistant professor with the New Mexico Institute of Mining and Technology for three years. He won an NSF CAREER award in 2009. His research interests include parallel and distributed systems, storage systems, fault tolerance, real-time systems, and performance evaluation. He is a member of the ACM.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: Auburn University. Downloaded on November 18,2026 at 18:38:52 UTC from IEEE Xplore. Restrictions apply.