

Similarity Metrics: Chebyshev Coulomb Force and Resultant Force for High-Dimensional Data

JIAN YING LIU, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China and School of Mathematics and Computer Science, Shanxi Normal University, Taiyuan, China

CHAOWEI ZHANG, Department of Computer Science, Yangzhou University, Yangzhou, China

MIN ZHANG, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China

XIAO QIN, Computer Science, Auburn University, Auburn, Alabama, USA

JIFU ZHANG, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China

The similarity metric has garnered widespread attention thanks to its potential applications in the fields of data mining, machine learning, and so on. Due to the interference of “distance concentration” caused by “Curse of dimensionality,” however, existing similarity metrics are inadequate in high-dimensional data analysis. In this study, we propose two innovative similarity metrics—Chebyshev Coulomb force and Chebyshev Coulomb resultant force—anchored on Chebyshev p -norms. In the initial phase, we eliminate dependency relationships among attributes by applying a metric matrix—and the theoretical analysis reveals that the Chebyshev p -norms is capable of mitigating the effect of “distance concentration” among high-dimensional data objects. Next, we devise two similarity metrics—Chebyshev Coulomb force and Chebyshev Coulomb resultant force—by adopting the metric matrix and Chebyshev p -norms. Chebyshev Coulomb force and Chebyshev Coulomb resultant force, being effective in characterizing the similarity among data objects, quantify the deviation of data objects from their respective dataset centers. Additionally, the two metrics alleviate the interference of “distance concentration.” Importantly, the discrepancy of data objects in attribute dimensions is captured by Chebyshev Coulomb force vector, rendering the similarity metric interpretable. By utilizing the UCI dataset, the experimental validation demonstrates the superiority of our similarity metrics, confirming their efficacy in mitigating the interference of “distance concentration.” Compared with the existing similarity metric approaches, the AUC index of outlier detection shows an average improvement of 8.18%—and the ARI, NMI, and F_{score} indices of clustering are revamped by averages 6.56%, 6.87%, and 6.01%, respectively.

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62172293, and Scientific and Technological Innovation Team of Shanxi Province, P.R. China (201805D131007). X. Qin’s work is supported by the U.S. National Science Foundation under Grants IIS-1618669 and OAC-1642133, as well as a grant (Number 240250) from Wright Media, LLC.

Authors’ Contact Information: Jian Ying Liu, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China and School of Mathematics and Computer Science, Shanxi Normal University, Taiyuan, China; e-mail: ljysx@stu.tyust.edu.cn; Chaowei Zhang, Department of Computer Science, Yangzhou University, Yangzhou, China; e-mail: cwzhang@yzu.edu.cn; Min Zhang, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China; e-mail: minzhang1@tyust.edu.cn; Xiao Qin, Computer Science, Auburn University, Auburn, Alabama, USA; e-mail: xqin@auburn.edu; Jifu Zhang (corresponding author), School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, China; e-mail: zjf@tyust.edu.cn.



This work is licensed under a [Creative Commons Attribution International 4.0 License](https://creativecommons.org/licenses/by/4.0/).

© 2025 Copyright held by the owner/author(s).

ACM 1556-472X/2025/4-ART75

<https://doi.org/10.1145/3715963>

CCS Concepts: • **Information systems** → **Data mining**;

Additional Key Words and Phrases: Similarity Metric, Distance Concentration, Chebyshev Coulomb Force, Metric Matrix, Chebyshev Norm

Associate Editor: Nitesh Chawla

ACM Reference format:

Jian Ying Liu, Chaowei Zhang, Min Zhang, Xiao Qin, and Jifu Zhang. 2025. Similarity Metrics: Chebyshev Coulomb Force and Resultant Force for High-Dimensional Data. *ACM Trans. Knowl. Discov. Data.* 19, 3, Article 75 (April 2025), 25 pages.
<https://doi.org/10.1145/3715963>

1 Introduction

With the rapid advancement of data mining applications, a massive amount of high-dimensional data has been accumulated. Not unexpectedly, the exploration of data mining technology to unlock hidden useful information and knowledge has aroused wide concern. As one of the study hotspots in data mining and machine learning, similarity metric portrays the degree of similarity among data objects [24]. “Distance concentration” is one manifestation of the “curse of dimensionality” incurred in high-dimensional spaces; that is, with rising dimensionality, distances tend to become almost indistinguishable [39]. Consequently, with the increase in dimensions, it is arduous to differentiate between the nearest and farthest neighbors of a given data object. Due to the interference of “distance concentration” in high-dimensional data space, the distances among data objects tend to be concentrated, forming tightly clustered groups in attribute space. Hence, the existing similarity metrics—including but not limit to Euclidean distance, Manhattan distance, Minkowski distance, Mahalanobis distance, cosine distance, and Coulomb’s force—fail to accurately characterize the differences among data objects: these metrics are ineffective for high-dimensional data analysis [17]. Additionally, most similarity metrics typically deal with dimensional similarity information as a whole, using a single metric value to indicate the degree of similarity among data objects [19]. These approaches are incapable of adequately reflecting differences that may exist between dimensions, resulting in a lack of interpretability in similarity measures. In a nutshell, for high-dimensional data, the existing similarity metrics inevitably suffer from the interference of “distance concentration”—and are inappropriate for high-dimensional data mining tasks [27].

Coulomb force and resultant force have played crucial roles in the development of high-dimensional similarity metrics [38]. In particular, Coulomb force reflects dimensional differences among data objects through vector projection in each dimension—thus depicting the similarity among data objects, yet the Coulomb resultant force quantifies the deviation of the data objects from their dataset centers. With Coulomb force vectors in place, the influence of attribute dimensions with small deviations is weakened—and the role of attribute dimensions with large deviations is amplified. The higher the dimension, the richer the deviation information, thus effectively alleviating the interference of the “curse of dimensionality.” Similarity calculation anchored on Coulomb force, however, adopts a norm distance ($p = n - 1$) that is vulnerable to the interference of norm “distance concentration” [9]. Simultaneously, since the Coulomb force and the resultant force ignore the correlation among attributes, the separability and information representation capabilities are poor. In addition, some distance functions [2, 39], such as fractional norm and cosine similarity that still exhibit “distance concentration,” are relatively less affected compared to L1 norm—Manhattan distance, with L1 norm’s influence being even smaller than that of L2 norm. Overall, the degree of “distance concentration” varies among different norm similarity metrics.

In this study, we propose two novel similarity measures, Chebyshev Coulomb force and Chebyshev Coulomb resultant force, utilizing the Chebyshev p -norm and Coulomb force to effectively alleviate the “distance concentration” interference. The two metrics have a strength of enhancing the interpretability of similarity measures. The key contributions are tabulated as follows:

- A new metric matrix is constructed, which practically eliminates dependencies among high-dimensional attributes.
- We theoretically analyze and demonstrate that for high-dimensional data, the Chebyshev p -norm helps to lessen the degree of convergence of “distance concentration.”
- Two similarity metrics, Chebyshev Coulomb force and Coulomb resultant force, are introduced—and the similarity equation for Coulomb force module length is obtained using the metric matrix and Chebyshev p -norm.
- The extensive experiments are carried out for high-dimensional data mining applications such as clustering analysis and outlier detection, confirming that our similarity metrics can remarkably alleviate the interference of “distance concentration.”

The remainder is organized as follows. Section 2 discusses previous work. Section 3 describes the background knowledge on metric spaces and Coulomb forces. Section 4 elaborates on the construction of metric matrix. Section 5 proposes two novel similarity metrics which is called Chebyshev norm Coulomb force and resultant force. Section 6 articulates the experimental setup along with the comprehensive analysis of the results. Finally, Section 7 concludes this article.

2 Related Work

Similarity metrics among data objects have profound impact on data mining and machine learning. Indeed, most data mining and machine learning algorithms, such as outlier detection NHOD [38], **k-nearest neighbor (KNN)** classification [37], and k -means cluster analysis [15], rely on similarity metrics. Generally speaking, existing similarity metrics can be categorized into two groups: traditional similarity metrics and similarity metrics based on distance metric learning.

Traditional similarity metrics include distance measures and similarity coefficient calculations. In high-dimensional space, the interference of “distance concentration” makes common distance metrics (e.g., Euclidean distance, Manhattan distance, Minkowski distance, Chebyshev distance, Mahalanobis distance, Lang’s distance, and other classical distance metrics) undesirable for reflecting relative distances among data items. That is, the relative difference of the nearest and farthest distances to any objects tend to approach 0 as the dimensionality grows [14]. Therefore, high-dimensional data similarity metrics powered by Euclidean and the other classical distance measures, such as the fractional norm, inevitably encounter the “curse of dimensionality” in high-dimensional space. Commonly used similarity coefficient calculation methods include Jaccard coefficient, Pearson correlation coefficient, and Cosine metric, to name just a few. More specifically, Jaccard coefficient is a measure of the degree of overlap between two binary variables, which effectively gauge similarity of the high-dimensional data in terms of attributes, yet cannot reflect similarity in high-dimensional space. The Cosine metric and Pearson correlation coefficient are applicable for high-dimensional space, and Cosine metric is considered as a distance metric that mitigates the interference of “distance concentration” of the p -norms. Nonetheless, cosine metric suffers from the interference of “distance concentration” [23]. Zhu et al. proposed a similarity metric called Coulomb resultant force, which is different from cosine similarity metric—an approach that only considers the direction between data objects and ignores differences in magnitude. The Coulomb resultant force of a data object echos its degree and direction of deviation from a dataset center. Additionally, the Coulomb force vector contains similarity information in any attribute dimension. Unfortunately,

the similarity calculation enabled by the Coulomb force essentially adopts Minkowski's distance, making it susceptible to the interference of Minkowski's "distance concentration" [38].

Most machine learning algorithms are sensitive to a given distance or similarity metric [1, 28, 32], and it is often arduous to capture the underlying semantic space of a problem with traditional similarity metrics like Euclidean distance. In response to the deficiencies, methods derived from metric learning were proposed to project data into traditional similarity metrics [16, 18, 30]. Remarkably, Mahalanobis distance metric learning is one of the most commonly used metric learning algorithms [34–36], which is equivalent to Euclidean distance under linear transformations. Besides this, representative metric learning algorithms include Linear Discriminant Analysis [13], Mixed Metric Learning [22], Nearest Neighbor Search Metric Learning [26], and Curve Distance Metric Learning [5]. Although the metric learning algorithms are practical for data with a simple structure, the algorithms cannot learn distance for non-linear datasets with a complex structure. Afterward, several strategies have been put forward in the literature [6, 26], including kernel methods and deep metric learning. Kernel methods in metric learning show significant potential for capturing non-linear relationships in data, as explored in the prior works [20, 22, 29]. These methods, however, face challenges related to computational costs and kernel selection. Deep metric learning, including methods like Deep Local Metric Learning [7], Deep Adversarial Metric Learning [8], and Dimensionality Reduction by Learning an Invariant Mapping [11], relies on deep network structures rather than distance metrics of data in a new representation space. Despite strong feature representation capabilities, deep metric learning approaches lack interpretability—and requires high data scalability and powerful computational resources. A handful approaches have been introduced to enhance interpretability in metric learning. For instance, a proposed method [31] sheds bright light on distance calculations in high-dimensional data, providing valuable insights into the underlying principles of similarity metrics.

Growing attention has been paid toward high-dimensional similarity search from the perspectives of data management and data mining: good examples include **locality sensitive hashing (LSH)** [25] and **hierarchical navigable small world (HNSW)** indexes [21]. LSH, a technique adopted in approximate nearest neighbor search, aims to efficiently find data points that are close to a given query point in high-dimensional spaces. HNSW—a powerful index structure for vector similarity search—combines skip lists and navigable small world graphs to achieve high search speeds and excellent recall. While LSH and HNSW are centered around efficient approximate nearest neighbor search, our similarity metric approach is hinged on physics-inspired principles.

In summary, the "distance concentration," driven by "curse of dimensionality," makes the traditional similarity metrics inadequate for high-dimensional data mining tasks such as clustering analysis and outlier detection. Subsequently, similarity metrics powered by distance metric learning have the advantage of autonomously learning specific distance functions for various tasks, thereby enhancing the performance of machine algorithms. When projecting the features of data into a new feature space, the dimensionality reduction lead to the loss of crucial information, making it insufficient to mirror the overall characteristics of data objects. On the other hand, non-reduced projections are still vulnerable to the interference from the "distance concentration" of traditional data similarity metrics for high-dimensional data. More importantly, similarity metrics derived from metric learning often exhibit limited interpretability.

3 Preliminaries

Given a high-dimensional dataset DS , let $DS = \{X_1, X_2, \dots, X_m\}$ be a set of m relevant attribute data objects, and $X_i = [X_{i1}, X_{i2}, \dots, X_{in}]$ be a data object of n attributes. Let L denote the metric matrix, and $DS' = [X'_1, X'_2, \dots, X'_m]$ be a set of irrelevant attributes projected by metric matrix L . Table 1 lists the symbols and notation used throughout this manuscript.

Table 1. Symbols and Notation

Symbol	Description	Symbol	Description
X	Set of metric space	d	Metric (distance function)
DS	Dataset	DS'	Irrelevant attribute dataset
X_i	The i th object in X	X'_i	The i th object in DS'
X_{ij}	The j th attribute of X_i	X'_{ij}	The j th attribute of X'_i
M	The number of objects	n	The number of attributes
L	Metric matrix	w_i	Eigenvector
E	Electric field intensity	q	Electric quantity of object
k_m	Electrostatic constant	r	Unit vector

3.1 Metric Space

A metric space characterizes the structure of distances among elements in a set. Given any two elements in metric space, the distance between the elements can be measured by a function commonly referred to as “metric” or “distance function.” In general, for set X and distance function d , the metric space can be defined as a tuple (X, d) , where d satisfies the following properties [4]:

- (1) $\forall x, y \in X, d(x, y) \geq 0$ positiveness
- (2) $\forall x \in X, d(x, x) = 0$ reflexivity
- (3) $\forall x, y \in X, d(x, y) = d(y, x)$ symmetry
- (4) $\forall x, y, z \in X, d(x, y) \leq d(x, z) + d(z, y)$ triangle inequality

According to a prior study [4], it is believed to a pseudo metric space where property 1 does not holds. In particular, property 4 can be appropriately scaled to $d(x, y) \leq \alpha d(x, z) + \beta d(z, y) + \delta$, and if property 3 holds, one can easily conclude that $\alpha = \beta$ for consistency.

3.2 P-Norms

The p-norm, a metric in vector space, quantifies the size or length of a vector. Given any n-dimensional vector $x = (x_1, x_2, \dots, x_n)$, then we have [9]:

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}. \quad (1)$$

The p-norms are typically valid in which p is a real number greater than or equal to 1. Obviously, different values of p will lead to various norms. For instance, when $p = 1$, it is sometimes called L1-norm or Manhattan metric; for the $p = 2$ case, it corresponds to L2-norm or Euclidean metric; the limit for $(p = \infty)$, it is infinity-norm or Chebyshev metric. In some cases, triangular inequality of the p-norm no longer holds (e.g., $0 < p < 1$)—and a consequence is “pseudo-norm” or “fractional norm.”

There is an extension of p-norms [9], allowing p to take values from positive real number $(\mathcal{R}^+)$, and it is demonstrated that “distance concentration” is an intrinsic property of p-norm, independent of the sample size. Most machine learning applications—including regularization, distance metrics, and feature selection—widely adopt the p-norm. Intriguingly, regarding the “distance concentration” phenomenon under the p-norms, the following theorem and lemma are theoretically confirmed in an earlier study [9].

THEOREM 1. *Let $X = (X_1, \dots, X_n)$ be a random vector with component being independent and following the distribution $\mathcal{F}$, then*

$$\lim_{n \rightarrow \infty} \frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)} = 0,$$

where $E(\|X\|_p)$ represents the expectation of the p -norm $\|X\|_p$, $\text{Var}(\|X\|_p)$ is the variance of the p -norm $\|X\|_p$.

In view of Theorem 1, the major consequence is that “concentration” is indeed inherent quality whatever the p -norms is in high-dimensional space from the perspective of variance and expectation of norm. Additionally, it is further evident that distribution of the component is supposed to taken into account to elaborate on “concentration” rather than a finite sample size or the “empty space” phenomenon. It is worth noting that this conclusion is applicable to all p -norms, which is independent of sample size.

LEMMA 1. *Let $X = (X_1, \dots, X_n)$ be a random vector with component being independent and following the distribution $\mathcal{F}$, then*

$$\lim_{n \rightarrow \infty} \frac{E(\|X\|_p)}{n^{1/p}} = \mu_{\mathcal{F},p}^{1/p},$$

where $\mu_{\mathcal{F},p} = E(|X_i|^p)$.

LEMMA 2. *Let $X = (X_1, \dots, X_n)$ be a random vector with component being independent and following the distribution $\mathcal{F}$, then*

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(\|X\|_p)}{n^{2/p-1}} = \frac{\sigma_{\mathcal{F},p}^2}{\left(p \cdot \mu_{\mathcal{F},p}^{(p-1)/p}\right)^2},$$

where $\mu_{\mathcal{F},p} = E(|X_i|^p)$, $\sigma_{\mathcal{F},p}^2 = \text{Var}(|X_i|^p)$.

3.3 Coulomb Force and Coulomb Resultant Force

Coulomb force, serving as a similarity metric, models the dimensional differences among data objects through the vector projection on each dimension, thus effectively alleviating the “distance concentration” interference caused by the “curse of dimensionality” [38]. Given data objects X_i, X_j in DS , the Coulomb force and Coulomb resultant force are defined as [38]:

$$\vec{F}(X_i, X_j) = Eq\vec{r} = k_m \frac{qq'}{\|r - r'\|^{n-1}} \vec{r}, \quad (2)$$

where $\|r - r'\|$ denotes modulus length of data objects X_i, X_j , $\vec{r}$ represents information about the direction of the vector $X_i - X_j$. $E = K_m q / \|r - r'\|^{n-1}$ indicates the electric field strength at X_j of the electrostatic field formed by source charge X_i .

In an n -dimensional electrostatic field Q consisting of m charges, the sum of the Coulomb forces exerted on data object X by the other data objects $X' \in Q \setminus X$ is referred to as the Coulomb resultant

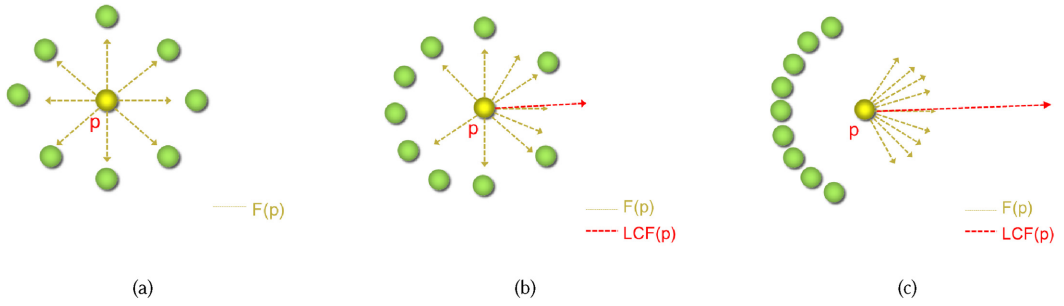


Fig. 1. Coulomb force and Coulomb resultant force.

force, which is expressed as

$$\begin{aligned}\overrightarrow{LCR}(X) &= \sum_{X' \in Q \setminus X} \overrightarrow{F}(X, X') \\ &= [f_1(X), \dots, f_d(X), \dots, f_n(X)],\end{aligned}\quad (3)$$

where $f_d(X)$ is the summation of the values of the Coulomb forces of data object X in dimension d : the sum can be calculated as $\sum_{X' \in Q \setminus X} f_d(X, X')$. In Equations (2) and (3), the Coulomb force signifies the relative positional relationship between charge p and the other charges, and the Coulomb resultant force reflects the degree to which charge p deviates from the electrostatic field. Figure 1 illustrates the Coulomb force and Coulomb resultant force of charge p in three distributions of electrostatic fields, where the solid dots represent charges, the yellow arrows represent Coulomb force, and the red arrows denote Coulomb resultant force. The green charge gradually converges around the yellow charge. Further, under the three different distributions of electrostatic fields formed by green charges, the Coulomb resultant force on charge p is distinct—and aligns with the distribution of each electrostatic field. To be specific, the closer the green charges gather around p , the greater the Coulomb resultant force. Regarding the dataset as an electrostatic field and the data objects as charges within the field, the Coulomb force indicates the similarity among data objects, whereas the resultant Coulomb force characterizes the deviation of a data object from its center of the dataset.

4 Concentration Effects

An earlier study [9] indicates that the phenomenon of “distance concentration” is strongly associated with the correlation among the attributes of data objects. That is, the stronger the correlation or dependence between high-dimensional attributes, the more pronounced the “distance concentration” is. Besides, in high-dimensional space, the p -norm distance measure, defined by the p -norm, gives rise to varying degrees of “distance concentration” for different values of p .

4.1 Metrics Matrix

Linear Metric Learning is indeed a machine learning approach to learning a linear transformation of data that helps in distinguishing samples from diverse categories. This transformation is typically represented by a matrix—referred to as metric matrix or metric learning matrix—that is learned from training data. Practically speaking, assorted tasks may require different formulations of an objective function, which in turn influence learned metrics [34].

When it comes to a high-dimensional data space, the dimensionality reduction, such as PCA, is a way of mitigating the impact of “distance concentration.” When projecting the features of data into

a new feature space, the dimensionality reduction lead to the loss of crucial information, making it insufficient to mirror the overall characteristics of data objects. An earlier study [9] shows that in view of the “distance concentration” phenomenon, the attributes of data objects should be as uncorrelated as possible. Hence, the learning objective of metric matrix L is to transform a set of n -dimensional correlated vectors into n -dimensional uncorrelated vectors. In one-dimensional space, the variance indicates the extent of dispersion of the data, while for high-dimensional data space, the covariance reflects the level of correlation between two attributes. More precisely, a covariance of 0 signifies that the two attributes are linearly uncorrelated. Ultimately, it is conducive to the goal of uncorrelated attributes that through the n unit orthogonal bases of the metric matrix, projecting original data onto that results in attributes with a covariance of 0. Simultaneously, this approach maximizes the variance of the attributes to offer a more effective data characterization. The learning process of metric matrix L is described as follows.

To achieve an irrelevant and variance-maximizing subspace, the original high-dimensional data space is subjected to linear projection. Given dataset $DS' = (X'_1, X'_2, \dots, X'_m) \in R^{m \times m}$ in n -dimensional space, the dataset in n -dimensional space is transformed after projection

$$DS' = L^T DS. \quad (4)$$

where the metric matrix L is composed of n n -dimensional vector, and $X'_i = L^T X_i$ is the representation of the data object X_i in the dataset DS' . The covariance matrix C of the dataset DS is expressed as

$$C = COV(DS) = \frac{1}{m}(DS - \overline{DS})(DS - \overline{DS})^T. \quad (5)$$

Then, we obtain covariance matrix C' of dataset DS' as

$$\begin{aligned} C' &= COV(DS') = \frac{1}{m}(DS' - \overline{DS'})(DS' - \overline{DS'})^T \\ &= \frac{1}{m}L^T(DS - \overline{DS})(DS - \overline{DS})^T L = L^T CL \end{aligned} \quad (6)$$

The objective of metric matrix L is to maximize the variance of the dataset DS' while ensuring the covariance between data attributes as 0. Thus, the optimization objective is defined as

$$\begin{aligned} \max_L & tr(L^T CL) \\ st. & L^T L = I. \end{aligned} \quad (7)$$

To solve the above objective function, Lagrange multipliers are introduced:

$$Cw_i = \lambda_i w_i. \quad (8)$$

The metric matrix L can be derived by decomposing the eigenvalues of the covariance matrix. Specifically, the n eigenvectors corresponding to the n eigenvalues are used to construct the metric matrix L . The steps for solving metric matrix L are outlined as follows:

- Step 1: Decentralize all samples. $X_i \leftarrow X_i - \sum_{i=1}^m X_i$, subtracting each row (attribute) in the DS from the mean of that row, that is, data centralization.
- Step 2: Calculate the covariance matrix C of the dataset DS according to Equation (5). Where C_{ij} depicts the correlation between attributes i and j .
- Step 3: Perform the eigenvalue decomposition on the covariance matrix C , and extract the eigenvalues from the diagonal of matrix.
- Step 4: Solve the eigenvector w_i corresponding to the eigenvalue λ_i . According to Equation (8), the eigenvector w_i under each eigenvalue is calculated, which is constitutive of the metric matrix $L = (w_1, w_2, \dots, w_n)$.

Metric matrix L serves as a transformation that reshapes the original attribute space, being converted from a correlated state to an uncorrelated one. By applying L as described in Equation (4), each data object in dataset DS is projected to create transformed dataset DS' . The covariance matrix C' of dataset DS' quantifies the correlation between attributes. When $i = j$, C'_{ii} signifies the variance of the attributes, which resembles the extent of dispersion of the attribute values. On the other hand, when $i \neq j$, C'_{ij} indicates the correlation among attributes. Referring to Equation (7), the diagonal elements of covariance matrix C' of the dataset DS' are all positive, the remaining elements are 0, and the sum of the diagonal elements is the largest possible. This outcome demonstrates that, through the mapping of dataset DS' by metric matrix L , the correlation between attributes is eliminated while simultaneously maximizing the variance of each attribute. Intriguingly, this method not only enhances the separability and information representation in high-dimensional attribute space but also provides an effective way of alleviating the effect of “distance concentration.”

4.2 Degree of Distance Concentration

The phenomenon of “distance concentration” is analyzed and discussed specially under the Minkowski norm [14], which was extended to include the fractional norm [9]. It is noteworthy that as dimensionality increases, the p-norms exhibit a pronounced “distance concentration” phenomenon, indicating that in high-dimensional spaces, regardless of the specific p-norm used, the norms tend to aggregate within a narrow range. Theorem 1 suggests that concentration effect is an inherent property of all p-norms in high-dimensional space. However, different p-values lead to various extent to “distance concentration.” To quantify the extent of “distance concentration” phenomenon with p-norms, we introduce a specific ratio to assess the degree of “distance concentration.”

$$RV_{\mathcal{F},p} = \frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)}, \quad (9)$$

where $RV_{\mathcal{F},p}$ denotes the relative variance of the norm.

$RV_{\mathcal{F},p}$ characterizes the degree of concentration effect of the p-norms by correlating the measure of diffusion—standard deviation—with the measure of location or expectation. It takes a non-negative range of values, where smaller values denote a more tightly clustered distribution of norm values, while larger values signify a broader spread of norm values. To delve further into the interplay between relative variance and the chosen p-value within a specific dimensionality, we leverage Lemmas 1 and 2 to bring forth Theorem 2, the proof of which is given in the Appendix.

THEOREM 2. *Let $X = (X_1, \dots, X_n)$ be a random vector with component being independent and following the distribution $\mathcal{F}$, then*

$$\lim_{n \rightarrow \infty} \sqrt{n} \cdot \frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)} = \frac{1}{p} \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}}$$

where $\mu_{\mathcal{F},p} = E(|X_i|^p)$, $\sigma_{\mathcal{F},p}^2 = \text{Var}(|X_i|^p)$.

Theorem 2 illustrates that in the case of large dimensionality (n), the relative variance $RV_{\mathcal{F},p}$ is associated with the distribution of $|X_i|^p$ and the chosen p-value. Indeed, the mean and variance of $|X_i|^p$ depend on the selection of the p-value as well as the distribution of $|X_i|$; therefore, the relative variance $RV_{\mathcal{F},p}$ is intricately tied to the choice of p-value in the context of the p-norm.

Theorem 2 asserts that when n is large, the product of $\sqrt{n}$ and relative variance $RV_{\mathcal{F},p}$ converges to the product of $1/p$ and $\sigma_{\mathcal{F},p}/\mu_{\mathcal{F},p}$. To investigate the factors influencing the relative variance $RV_{\mathcal{F},p}$,

we obtain $\frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)} \approx \frac{1}{\sqrt{n}} \cdot \left(\frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}} \right)$ by moving the equation of $\sqrt{n}$. Thus, when the dimension n is sufficiently large, the relative variance $RV_{\mathcal{F},p}$ is approximately expressed as

$$RV_{\mathcal{F},p} = \frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)} \approx \frac{1}{\sqrt{n}} \cdot \left(\frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}} \right). \quad (10)$$

Furthermore, when the dimension is fixed (larger), relative variance $RV_{\mathcal{F},p}$ with respect to p is defined as

$$\mathcal{K}_{\mathcal{F},p} = \frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}}. \quad (11)$$

Equation (11) suggests that for a fixed p -value, the relative variance is affected by the distribution of the $|X_i|^p$ within a constant dimension d . To address data with varying scales and dimensions, standardization is a prevalent data pre-processing operation that scales the data to the same data interval and range to alleviate the impact of differences in scale, features, and distribution on the model. In classification and clustering algorithms, Z-Score normalization and Max-Min normalization methods are widely employed, especially when a distance metric is utilized to measure similarity. Nonetheless, the Max-Min normalization method proves to be highly sensitive to outliers and noise, so Z-Score normalization is typically favored. After Z-Score normalization, the value of each attribute adheres to a standard normal distribution, where we have mean value $\mu = 0$ and SD $\sigma = 1$. In particular, if variable X_i conforms to the standard normal distribution, Corollary 1 is established, with the detailed proof included in the Appendix.

COROLLARY 1. *If the variable X_i subjects to the standard normal distribution, the relative variance $RV_{\mathcal{F},p}$ at the fixed dimension is bounded in the case of $0 < p < 1$. When it comes to $p > 2$, the relative variance tends to increase as p increases. Remarkably, when $p \rightarrow \infty$, the relative variance reaches its highest value.*

Corollary 1 asserts that for a fixed dimension, relative variance $RV_{\mathcal{F},p}$ is prone to fluctuate with the norm parameter p . In other words, the p -norm distance metric defined by the value of p results in varying distance-concentration degrees. Specifically, $0 < p < 1$ entails that relative variance $RV_{\mathcal{F},p}$ is bounded; in the case of $p > 2$, relative variance $RV_{\mathcal{F},p}$ goes up with an increasing p -value. Notably, when $p \rightarrow \infty$, the relative variance reaches its maximum value—and tends to an equilibrium state. At this point, the extent of “distance concentration” denoted by the norm distance metric is maximized, a condition referred to as the “Chebyshev norm.”

5 Similarity Metrics

The phenomenon of “distance concentration” arising from the “curse of dimensionality” is, in fact, associated with both attribute correlation and a specific choice of the p -norms. Hence, eliminating attribute correlation and opting for the Chebyshev p -norm can be an effective means of mitigating the impact of “distance concentration.”

5.1 Chebyshev Coulomb Force

In Equation (2), Coulomb force—denoted as a vector $\vec{F}(X_i, X_j)$ —is expressed as $[f_1(X_{ij}), f_2(X_{ij}), \dots, f_k(X_{ij}), \dots, f_n(X_{ij})]$, which characterizes the similarity between data objects X_i and X_j . Here, $f(X_{ij})$ represents differences between data objects X_i and X_j on each dimension. If the differences between any pair of data objects X_i and X_j are non-zero in all the dimensions (i.e., $X_{ik} \neq X_{jk}$ for all k), then the Coulomb force will also be non-zero—and it is independent of the dimensions of the data objects.

The metric matrix modeled in Equation (8) transforms the high-dimensional correlated attribute space DS into the uncorrelated attribute space DS' , eliminating the dependencies among attributes in the high-dimensional data space—and mitigating the interference of “distance concentration.” To ensure equal weight of each attribute in distance measurement while mitigating the influence of different attribute units, we utilize standardization to achieve dimensional consistency. In the standardized uncorrelated attribute space DS' , attribute values are scaled to a standard range, with a mean of 0 and a SD of 1, resulting in attribute values following a standard normal distribution. Upon projecting with the metric matrix L and standardizing, data object can be labeled as $L^T X_i$. Furthermore, Corollary 1 (see Section 4.2) stipulates that when attribute variables follow a standard normal distribution, the Chebyshev norm can effectively alleviate the interference of “distance concentration.” Consequently, to effectively measure the similarity between high-dimensional data objects X_i and X_j , the electric field intensity E in Equation (2) is redefined as using the metric matrix and Chebyshev norm of Corollary 1:

$$\begin{aligned} E = \|X_i - X_j\| &= \lim_{p \rightarrow \infty} \left(\sum_{k=1}^n |L_k^T X_i - L_k^T X_j|^p \right)^{1/p} \\ &= \max_{k=1}^n \left(|L_k^T X_i - L_k^T X_j| \right) = \max_{k=1}^n \left(|L_k^T (X_i - X_j)| \right), \end{aligned} \quad (12)$$

where L_k^T is the transpose of the k th column vector of the metric matrix L .

For data objects X_i and X_j , the attributes are projected to irrelevant attributes through the vector L_k^T of the metric matrix. In the sequel, the Chebyshev norm is employed to reflect the electric field intensity E at the location of X_j . Importantly, the larger the value of E , the higher the similarity between X_i and X_j , and vice versa.

Moreover, the parameter q in Equation (2) represents the electric quantity of a data object. Given that the importance of data objects in dataset DS is in common, we treat different data objects as charges with the same electric quantity q , regardless of their positions. Thus, q is set to a constant value, which has little impact on similarity measures. For simplicity in calculation, q is set to 1 as referenced in [38]. According to Equations (2) and (12), the Chebyshev Coulomb force between any two data objects X_i and X_j is denoted as $\vec{CF}(X_i, X_j)$. Thus, we have

$$\begin{aligned} \vec{CF}(X_i, X_j) &= Eq\vec{r} = \max_{k=1}^n \left(|L_k^T (X_i - X_j)| \right) \vec{r} \\ &= [f_1(X_{ij}), f_2(X_{ij}), \dots, f_k(X_{ij}), \dots, f_n(X_{ij})]. \end{aligned} \quad (13)$$

Equation (13) suggests that as the similarity between data objects X_i and X_j increases, the magnitude of Chebyshev Coulomb force $\vec{CF}(X_i, X_j)$ drops. Conversely, the greater discrepancy between the data objects X_i and X_j , the larger of Chebyshev Coulomb force $\vec{CF}(X_i, X_j)$. From a vector perspective, as the difference between the data objects X_i and X_j in dimension intensifies, the Chebyshev Coulomb force $\vec{CF}(X_i, X_j)$ also increases, and the opposite is true.

Compared with the traditional distance metrics, we advocate for a full consideration of the impact of the correlation among data attributes on the “distance concentration.” Through the metric matrix L , dependencies among original data attributes are eliminated, thus leading to a new data representation $L_k^T X_i, L_k^T X_j$. Subsequently, for the above data object $L_k^T X_i$ and $L_k^T X_j$, the similarity is calculated using Chebyshev’s norm, which effectively mitigates the interference of “distance concentration.” What is noteworthy is that the component force $f_k(X_{ij})$ of Chebyshev Coulomb force $\vec{CF}(X_i, X_j)$ captures the discrepancy between data objects X_i and X_j in each dimension: this method provides a granular and informative representation of the overall similarity.

Accordingly, the magnitude of the Chebyshev Coulomb force based on Equation (13) is expressed as

$$\left\| \overrightarrow{CF}(X_i, X_j) \right\| = \max_{k=1}^n \left(\left| L_k^T (X_i - X_j) \right| \right). \quad (14)$$

The magnitude $\left\| \overrightarrow{CF}(X_i, X_j) \right\|$ in Equation (14) captures the overall similarity of the data objects X_i and X_j : the greater the similarity, the smaller the Coulomb force. On the contrary, a stronger discrepancy between data objects X_i and X_j results in a larger Coulomb force. Additionally, as a distance metric, $\left\| \overrightarrow{CF}(X_i, X_j) \right\|$ possesses the following properties, the proof of which is available in the Appendix.

PROPERTY 1. *Given any two data objects X_i and X_j , in a high-dimensional metric space, the following distance metric properties is satisfied by the magnitude $\left\| \overrightarrow{CF}(X_i, X_j) \right\|$:*

- (1) $\left\| \overrightarrow{CF}(X_i, X_j) \right\| \geq 0$ *positiveness*
- (2) $\left\| \overrightarrow{CF}(X_i, X_j) \right\| = 0 \Leftrightarrow X_i = X_j$ *reflexivity*
- (3) $\left\| \overrightarrow{CF}(X_i, X_j) \right\| = \left\| \overrightarrow{CF}(X_j, X_i) \right\|$ *symmetry*
- (4) $\forall X_t, \left\| \overrightarrow{CF}(X_i, X_j) \right\| \leq \left\| \overrightarrow{CF}(X_i, X_t) \right\| + \left\| \overrightarrow{CF}(X_t, X_j) \right\|$ *Triangular inequality*

As is shown in property 1, similarity metric of the Chebyshev Coulomb force's magnitude satisfies the mathematical properties of distance metrics, including positiveness, reflexivity, symmetry, and triangular inequality.

5.2 Chebyshev Coulomb Resultant Force

Assessing the similarity between a data object and a local dataset has been a subject of intense investigation in the realm of data mining such as outlier detection. Coulomb resultant force, as a similarity metric, quantifies the deviation information of data object in each dimension concerning the local dataset and characterizes its similarity to the dataset center. In the electrostatic field constructed by dataset LDS , in accordance with Equations (3) and (13), the Chebyshev Coulomb resultant force of data object X_i receiving Chebyshev Coulomb force from other data object $X_j \in LDS \setminus X_i$ is denoted as $\overrightarrow{CF_{LDS}}(X_i)$, which can be expressed as

$$\begin{aligned} \overrightarrow{CF_{LDS}}(X_i) &= \sum_{X_j \in LDS \setminus X_i} \overrightarrow{CF}(X_i, X_j) \\ &= \sum_{X_j \in LDS \setminus X_i} \max_{k=1}^n \left(\left| L_k^T (X_i - X_j) \right| \right) \vec{r} \\ &= [f_1(X_i), f_2(X_i), \dots, f_k(X_i), \dots, f_n(X_i)]. \end{aligned} \quad (15)$$

where $f_k(X_i)$ is the Coulomb component force corresponding to the resultant force $\overrightarrow{CF_{LDS}}(X_i)$ in dimension k , which can be calculated as $\sum_{X_j \in LDS \setminus X_i} f_k(X_i, X_j)$, $k = 1, 2, \dots, n$.

The Chebyshev Coulomb resultant force $\overrightarrow{CF_{LDS}}(X_i)$ reflects the overall deviation of the data object from the center of the local dataset. Figure 2 reveals Chebyshev Coulomb resultant force under different dataset distributions, where yellow dot denotes data objects in high-density regions, blue dot represents data objects in low-density regions, purple arrow is the Chebyshev Coulomb resultant force of data objects. Moreover, it demonstrates that data points in high-density regions are

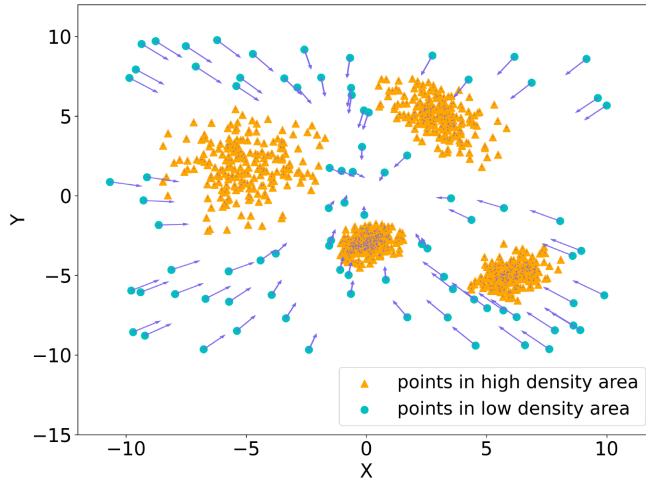


Fig. 2. Chebyshev Coulomb resultant force.

uniformly distributed, resulting in weak and divergent Coulomb forces among nearby neighbors. These opposing forces effectively cancel each other out, leading to a smaller Chebyshev Coulomb resultant force. In contrast, data objects in the low-density data, such as the boundary points or outliers, have unevenly distributed neighborhoods, among which there are strong and directionally concentrated Chebyshev Coulomb force that results in a large the Coulomb resultant force.

A more significant aspect is that Chebyshev Coulomb force $\overrightarrow{CF_{LDS}}(X_i)$ effectively alleviates the interference of “distance concentration.” First of all, it is the cumulative sum of Chebyshev Coulomb force $\overrightarrow{CF}(X_i, X_j)$, which lowers the degree of “distance concentration” through the metric matrix L and the Chebyshev norm. Second, in high-dimensional datasets, a significant number of irrelevant dimensions exist, leading to challenges for traditional distance measures due to the phenomenon of “distance concentration.” However, Chebyshev Coulomb resultant force is a vector sum of Coulomb forces. When data objects are in proximity to the dataset center along a certain dimension, its Coulomb component is relatively small. Conversely, in dimensions that provide distinguishing information, data objects have higher Coulomb components. High dimensions yield abundant deviation information, providing extensive discriminative capabilities. The Chebyshev Coulomb resultant force diminishes ineffective dimensions while amplifying effective ones, characterizing the deviations of data objects from the dataset center in each dimension. In contrast to local dimensionality reduction methods including subspace-based methods that achieve reduction by sacrificing partial information, the Chebyshev Coulomb resultant force autonomously filters out ineffective dimensions and weights in effective ones, thereby preserving global data information. That is, it meticulously considers subtle deviations on each dimension, enabling the Chebyshev Coulomb resultant force to alleviate “distance concentration” and enhance the accuracy of similarity measurement.

5.3 Interpretability

The traditional distance metric only describes the overall similarity among data objects, ignoring the nuanced variations on each dimension. Even though the similarity among objects is in common, similarity information carried by attributes may be diverse. The Chebyshev Coulomb force not only unveils the overall similarity among data objects but also captures the differences on each dimension, thereby providing an explanation for the similarity among data objects.

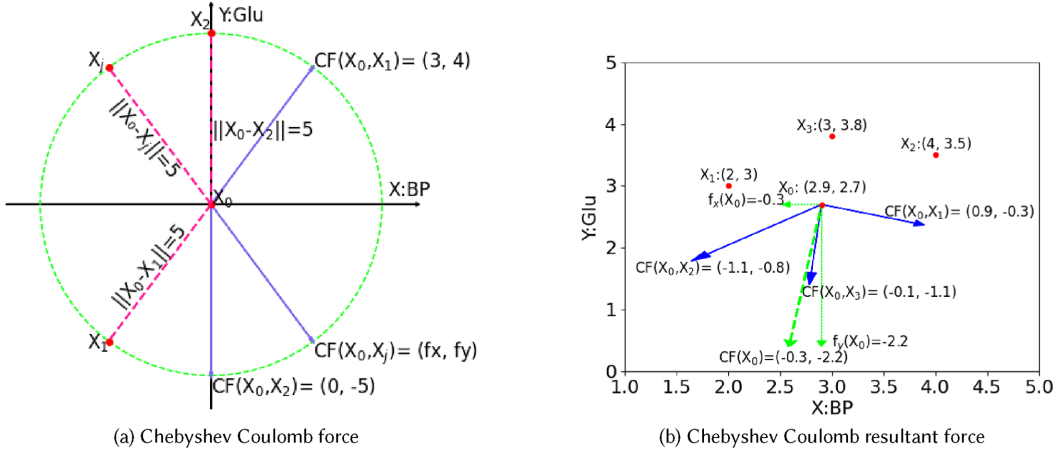


Fig. 3. Chebyshev Coulomb force and resultant force.

Figure 3(a) reveals the Chebyshev coulomb force among data objects, where the red solid point denotes the data object, the purple solid arrow represents the Chebyshev coulomb force, dimension X is **blood pressure (BP)**, and dimension Y denotes **blood glucose (Glu)**; the data object X_0 is located in the center of the circle, and data object X_j is located on the green circle centered on the data object X_0 with radius 5.

It is evident from Figure 3(a) that the discrepancy between any data objects X_j and X_0 in Glu and BP on the circle maybe entirely different. However, the similarity between X_j and X_0 on the circle is identical under the traditional distance metric, overlooking these differences in Glu and BP. In contrast, the differences on each dimension are characterized by Chebyshev Coulomb force, which provides a further explanation of the similarity between X_j and X_0 . For instance, X_0 exhibits the similar characteristics with data objects X_1 and X_2 in terms of the magnitude of the Chebyshev Coulomb force; that is, $\|\vec{CF}(X_0, X_1)\| = \|\vec{CF}(X_0, X_2)\| = 5$. Nonetheless, the Chebyshev coulomb force of X_0 to X_1 is $\vec{CF}(X_0, X_1) = (3, 4)$, indicating that the BP and Glu of data object X_1 are higher than that of X_0 . Analogously, there is $\vec{CF}(X_0, X_2) = (0, -5)$, which means that data objects X_0 and X_2 have the same BP, and Glu of X_0 is lower than that of X_1 by 5. In conclusion, we have Chebyshev coulomb force $\vec{CF}(X_0, X_j) = (f_{BP}, f_{Glu})$ for any data object X_0 and X_j on the circle, where the value denotes the discrepancy in BP, Glu, respectively, offering an explanation for the similarity between the data objects X_0 and X_j .

The component force $f_k(X_i)$ in Equation (15) of the Chebyshev Coulomb resultant force in corresponding dimension is the sum of the component force of the Coulomb force data object X_i received from dataset *LDS*, unfolding the deviation of data object from the local dataset center along the k th dimension. Besides, the overall degree of a data object deviating from the local dataset center is symbolized by using the magnitude of the Chebyshev Coulomb resultant force. Indeed, a local dataset typically features a common local distribution pattern. Therefore, a larger magnitude of the Chebyshev Coulomb resultant force indicates the data object deviates more significantly from the local dataset center, signifying a stronger inconsistency with the distribution pattern of the local dataset.

Figure 3(b) depicts the Chebyshev Coulomb resultant force between the data objects and the dataset center. A red solid point denotes a data object. Blue solid arrow represents the Chebyshev

coulomb force $\vec{CF}(X_i, X_j)$, green dashed arrow is Chebyshev Coulomb resultant force, and green dotted arrow denotes the component force on each dimension of Coulomb resultant force. Dimension X represents BP, while dimension Y is Glu. Data objects X_1, X_2, X_3 comprise the local dataset LDS .

As shown in Figure 3(b), the data object X_0 lies near the center of dataset LDS in dimension X . Consequently, the directions of the Coulomb force are scattered and cancel each other out, so component force of the Chebyshev Coulomb resultant force $f_x(X_0) = -0.3$ is smaller in dimension X . On the contrary, in dimension Y , X_0 is away from the center of dataset LDS , and the directions of the Coulomb forces' components in dimension Y are concentrated and superimposed on each other. Hence, the Chebyshev Coulomb resultant force displays a larger component in dimension Y compared to dimension X , which is $f_y(X_0) = 2.2$. Overall, the magnitude of the Chebyshev Coulomb resultant force $\|\vec{CF}_{DS}(X_0)\| \approx 2.22$ suggests that the data object manifests a deviation from the LDS center, thereby revealing a lack of conformity with the distribution pattern observed in the LDS . Since $|f_y(X_0)| \gg |f_x(X_0)|$, the dimension Y is regarded as a key factor in the disparity between the data object and the local dataset's distribution pattern. Compared with the traditional similarity metric, $f_k(X_i)$ unravels deviation of the data object from the data center of the local dataset in corresponding dimension, which provides a basis for the similarity analysis of the data object and the dataset—and revamps the interpretability of the similarity measures.

5.4 Time Complexity Analysis

The Chebyshev-Coulomb force in Equation (13) is centered around two main components: metric matrix learning and Coulomb force calculation. In metric matrix learning, the time complexity of centering all samples—zero-mean normalization—is $O(mn)$. Calculating the covariance matrix has a time complexity of $O(mn^2)$, while performing eigenvalue decomposition to compute eigenvalues and eigenvectors has a time complexity of $O(n^3)$. Utilizing the metric matrix to calculate the Coulomb force involves a complexity of $O(n^2)$. Therefore, the overall time complexity is dominated by the eigenvalue decomposition and eigenvector calculation, which simplifies to $O(n^3)$. In comparison, Mahalanobis distance is primarily dependent of the calculation of the metric matrix inversion, resulting in an overall time complexity of $O(n^3)$ —an equivalence with the time complexity of Chebyshev Coulomb force.

The Chebyshev Coulomb resultant force in Equation (15) measures the similarity between data object and local dataset. Its time complexity, beyond that of the Chebyshev-Coulomb force computation, also depends on the size of the local dataset: the overall time complexity is $O(n^3 + m'n^2)$, where m' is the size of the local dataset.

Like Mahalanobis distance metrics, the computational time of Chebyshev Coulomb force and resultant force primarily depends on dimension n . A high dimension gives rise to a long computation times. Although Mahalanobis distance and Chebyshev Coulomb force theoretically share the same time complexity, matrix inversion in Mahalanobis distance can be more computationally intensive in practice due to numerical stability issues and the nature of direct methods [10]. In contrast, Chebyshev Coulomb force leverages efficient algorithms, making it generally less time-consuming than Mahalanobis distance.

6 Experimental Analysis

We gauge the efficacy of the proposed similarity metric techniques through the applications of outlier detection and clustering approaches. While the empirical studies primarily focus on classic data mining applications for demonstration purposes, our similarity metrics are also applicable to a wide range of deep learning tasks.

Table 2. Outlier Datasets

ID	Dataset	Attributes	Instances	Outlies
1	KDDCup99	42	60,839	246
2	Musk	166	3,062	97
3	Arrhythmia	274	452	66
4	HAPT	561	10,929	448
5	PInternetAds	1,555	3,264	454
6	20Newsgroups	13,542	19,997	997
7	Reuters-21578	19,920	12,968	324

The experimental platform is equipped with an Intel Core i9-12900H CPU at 2.50 GHz CPU, 16 GB RAM, Windows 10 OS, and PyCharm supporting the software development using the Python programming language. The experiments, driven by the UCI datasets, are focused on systematic performance evaluation of two similarity metrics—Chebyshev Coulomb force and resultant force—from the perspectives of outlier detection and clustering tasks.

UCI Datasets for Outlier Detection. We verify and test the outlier detection accuracy of the proposed similarity metrics on seven UCI datasets. The KDDCup99 dataset captures different types of network intrusions or attacks, with U2R as the outlier. The raw Musk dataset contains multiple Musk and non-Musk samples, with the non-Musk samples merged to construct the normal in-points data and the Musk samples labeled as the outliers. The Arrhythmia dataset is composed of patient records that were categorized as normal or having some type of arrhythmia (arrhythmia types were considered outlier data). The PInternetAds dataset is constituted by images from Web pages that are categorized as advertisements and non-advertisements, with advertisements designated as outliers. The 20Newsgroups dataset contains 19,997 news articles covering 20 topics, where category 19 is viewed as outliers. Finally, the Reuters-21578 dataset involves 50 topics corresponding to many short news machines, among which sample counts fewer than 40 are considered as outliers. The meta-information of these datasets is summarized in Table 2.

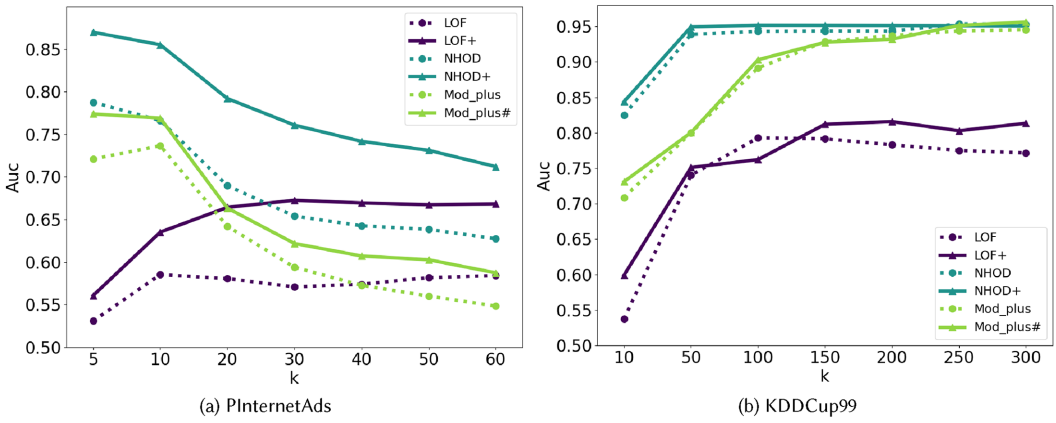
UCI Datasets for Clustering. The clustering performance of the proposed similarity metrics is verified on seven UCI datasets. More specifically, the USPS dataset is a set of handwritten digit images from the U.S. Postal Service covering the digits 0 through 9. The Mnist dataset is a large dataset containing handwritten digit images used to test the performance of machine learning algorithms. The COIL-20 dataset consists of photographs of 20 objects, with 72 images of each object, for the purpose of object recognition. The Gisette dataset is formed from 7,000 samples of high-dimensional data processed by various signal processing techniques for text categorization tasks. The Pixraw10P dataset is an image categorization dataset containing images of 10 categories. The UMIST face database is composed of 564 images of 20 individuals, which cover a range of poses from profile to frontal views. The TDT2 is a corpus dataset embracing 9,394 on-topic documents classified into 30 semantic categories. The summary of the datasets is tabulated in Table 3.

Validity Index. ROC_AUC are used for evaluating the outlier detection accuracy; **adjusted rand index (ARI)**, **normalized mutual information (NMI)**, and F_score are employed to evaluate the clustering accuracy.

Apart from the Chebyshev Coulomb force and resultant force, the other tested similarity metrics include four traditional similarity metrics: Euclidean distance (Euclidean), Manhattan distance (Manhattan), cosine similarity (Cosine), and Chebyshev distance (Chebyshev). Additionally, two non-traditional metrics anchored on learning—Mahalanobis distance [12] and Coulomb force [38]—are evaluated for the comparison purpose.

Table 3. Clustering Datasets

ID	Dataset	Attributes	Instances	Clusters
1	USPS	256	9,298	10
2	Mnist	784	10,000	10
3	COIL-20	1,024	1,440	20
4	Gisette	5,000	7,000	2
5	Pixraw10P	10,000	100	10
6	UMIST	10,304	564	20
7	TDT2	36,771	9,394	30

Fig. 4. Nearest neighbor parameter k .

6.1 Parameter k

The KNN is one of the core techniques in a diverse range of data mining and machine learning algorithms, and parameter k has a significant impact on KNN performance. To verify the influence of parameter k on the Chebyshev Coulomb force and resultant force, we assess the three outlier detection algorithms—LOF [3], NHOD [38], and Mod_plus [33]—driven by the PInternetAds and KDDCup99 datasets. Additionally, Chebyshev Coulomb force and resultant force respectively substitute for the Manhattan distance, Coulomb resultant force, and Euclidean distance: we denote these extensions as LOF+, NHOD+, and Mod_plus#.

Figure 4 reveals that the effect of parameter k to the similarity metric on the AUC is essentially consistent. Thus, when it comes to the outlier detection algorithms relying on Chebyshev Coulomb force and resultant force, parameter k can be properly set by leveraging the primitive reference.

6.2 Outlier Detection on the UCI Datasets

To compare against the six-similarity metrics, we are positioned to validate the superiority of Chebyshev Coulomb force and resultant force with regard to outlier detection handling the five UCI dataset listed in Table 1. The outlier detection algorithms evaluated include the kNN-based outlier detection algorithm or NHOD [38] as well as a non-kNN-based algorithm MCD [12]. Specially, the NHOD algorithm sets $k = \sqrt{m}$ [38], while MCD utilizes the default parameter settings from an earlier study [12]. The experimental and comparison results are plotted in Table 4.

Table 4. AUC of Outlier Detection for UCI Datasets

		KDDCup99	Musk	Arrhythmia	HAPT	PInternetAds	20Newsgroups	Reuters-21578	
Similarity		K=247	K=55	K=21	K=105	K=7	K=141	K=114	Avg.
NHOD	Euclidean	0.9227	0.9901	0.7809	0.8684	0.7040	0.7356	0.5369	0.7912
	Manhattan	0.9014	0.9774	0.7489	0.8382	0.6839	0.6984	0.5741	0.7746
	Cosine	0.8125	0.8712	0.7402	0.8495	0.6741	0.7096	0.5128	0.7386
	Chebyshev	0.9406	0.9878	0.7987	0.8879	0.7251	0.7525	0.5369	0.8042
	Mahalanobis	0.9454	0.9889	0.8202	0.8917	0.7333	0.7634	0.5636	0.8152
	Non-tradition	Coulomb force	0.9543	1.0000	0.8365	0.9205	0.7710	0.5975	0.8414
	Ours	0.9512	1.0000	0.8758	0.9814	0.8690	0.8596	0.6485	0.8836
									↑0.0553
MCD	Euclidean	0.8711	0.9912	0.7774	0.7939	0.6821	0.7291	0.5164	0.7658
	Manhattan	0.8409	0.9598	0.7478	0.7540	0.6517	0.6842	0.4863	0.7321
	Cosine	0.8472	0.9298	0.6977	0.7429	0.6309	0.6691	0.4621	0.7114
	Chebyshev	0.9084	0.9956	0.7512	0.8042	0.6918	0.7291	0.5269	0.7725
	Mahalanobis	0.9345	1.0000	0.8289	0.8252	0.7217	0.7698	0.5621	0.8060
	Non-tradition	Coulomb force	0.9209	0.9898	0.8073	0.8039	0.7016	0.5764	0.7937
	Ours	0.9342	1.0000	0.8676	0.8743	0.7788	0.7951	0.6138	0.8377
									↑0.0378

The two best AUC are marked in red and blue, respectively. ↑ represents the difference between the best and the baseline, which are marked in green.

Table 4 unveils that the Chebyshev Coulomb force and resultant force—exhibiting a clear edge over the other competitive similarity metrics—offer an average improvement of 8.18% in the AUC index, entailing that the two metrics are beneficial to the high-dimensional outlier detection applications. Compared with the traditional four metrics, our new approaches boost the AUC indexes of NHOD and MCD by an average of 10.64% and 9.22%, respectively. When it comes to two non-traditional metrics, NHOD and MCD witness an average AUC improvement of 5.53% and 3.78% offered by the new metrics. The key factor driving the observed improvements is listed below.

- The traditional similarity, such as Euclidean distance ($p = 2$), Manhattan distance ($p = 1$), Chebyshev distance ($p = \infty$), and cosine similarity, directly gauge distances among raw data objects without effectively capturing relative distances among data objects. Indeed, the relative discrepancies between the nearest and farthest distances from other data objects tend to approach 0 with the dimensionality expands, making the measures highly susceptible to the “distance concentration” phenomenon.
- Coulomb force, satisfying the $p = n - 1$ norm, overlooks the correlation between attributes—and results in weaker separability and information representation in high-dimensional attribute space. Conversely, Mahalanobis distance leverages a metric matrix to eliminate correlations among attributes, but it still employs Euclidean distance ($p = 2$) to characterize the similarity among data objects in irrelevant attribute spaces. Consequently, Mahalanobis distance is subjected to the moderate interference of “distance concentration.”
- Chebyshev Coulomb force and resultant force mitigate correlation among attributes through the metric matrix—and employ the Chebyshev Norm ($p = \infty$) to quantify the similarity among data objects. Therefore, compared to the other metrics, the new approaches are less affected by the interference of “distance concentration.”

Table 5. Performance on the UCI Datasets for Clustering Tasks

			USPS	Mnist	COIL-20	Gisette	Pixraw10P	UMIST	TDT2	Avg.
ARI	Tradition	Euclidean	0.5359	0.3814	0.6231	0.7388	0.8639	0.7705	0.6428	0.6509
		Chebyshev	0.5583	0.3976	0.6510	0.7515	0.8804	0.7218	0.6355	0.6566
		Cosine	0.5364	0.3822	0.5827	0.7388	0.8285	0.7458	0.6364	0.6358
	Non-tradition	Coulomb force	0.5368	0.3824	0.6404	0.7371	0.7830	0.7495	0.6136	0.6347
		Mahalanobis	0.5668	0.4024	0.6614	0.7571	0.8030	-	-	0.6381
		Ours	0.6115	0.4557	0.6605	0.8281	0.9063	0.8013	0.7011	0.7092
↑0.0656										
NMI	Tradition	Euclidean	0.6132	0.4997	0.7855	0.8241	0.9187	0.8775	0.7954	0.7572
		Chebyshev	0.6536	0.5427	0.8198	0.8709	0.9347	0.8457	0.7964	0.7805
		Cosine	0.6137	0.5015	0.7684	0.8241	0.9207	0.8556	0.7065	0.7415
	Non-tradition	Coulomb force	0.6139	0.5011	0.8076	0.8227	0.9095	0.8413	0.7842	0.7543
		Mahalanobis	0.6736	0.5627	0.8398	0.8908	0.9287	-	-	0.7791
		Ours	0.6762	0.5981	0.8523	0.9496	0.9890	0.9074	0.8421	0.8307
↑0.0687										
F_score	Tradition	Euclidean	0.6884	0.5490	0.7303	0.8103	0.8950	0.6981	0.7649	0.7337
		Chebyshev	0.7009	0.5495	0.7829	0.7769	0.9351	0.6171	0.7811	0.7348
		Cosine	0.6667	0.5118	0.6948	0.7769	0.9052	0.6924	0.7894	0.7196
	Non-tradition	Coulomb force	0.6672	0.5499	0.6983	0.7757	0.8032	0.6719	0.7292	0.6993
		Mahalanobis	0.7209	0.5595	0.7929	0.7969	0.9359	-	-	0.7612
		Ours	0.7545	0.6010	0.8003	0.8493	0.9847	0.7246	0.8011	0.7879
↑0.0601										

The two best contenders are marked in red and blue, respectively. ↑ represents the difference between the best and the baseline, which are marked in green.

—Since the KDDCup99 and Musk datasets have relatively low dimensionality, measuring the relative variance to characterize the extent of “distance concentration” only yields noticeable effects when dealing with high-dimensional attributes.

6.3 Clustering on UCI Datasets

To verify the strengths of Chebyshev coulomb force in the arena of cluster analysis, we compare the six similarity metrics using the UCI datasets lined up in Table 2 coupled with the k -means clustering algorithm from the sklearn library. In particular, the k -means algorithm is configured with the parameters that include `n_clusters = clusters`, `init = “k-means++”`, the `algorithm = “elkan.”` Table 5 outlines the comparative experimental results in terms of ARI, NMI, and F_score.

Table 5 paints a clear picture of remarkable performance delivered by Chebyshev Coulomb force compared against all the alternative similarity metrics. For example, Chebyshev Coulomb force proves to be applicable to high-dimensional clustering tasks, demonstrating average improvements of 6.56%, 6.87%, and 6.01% with regard to ARI, NMI, and F_score. More importantly, the primary reason behind this superiority is twofold. First, Euclidean distance, cosine similarity, and Chebyshev distance are seriously interfered by “distance concentration,” leading to suboptimal clustering results. Second, Coulomb force ($p = n - 1$) insufficiently takes into account the correlation among attributes—and Mahalanobis distance disregards the choice of p -value for the norm, thereby being susceptible to the “distance concentration” effect. In a distinctively different way, the Chebyshev Coulomb force takes a comprehensive approach by adequately considering

attribute correlations and norm distance metrics, effectively mitigating the “distance concentration.” The enhanced performance is accomplished by constructing a metric matrix L to eliminate inter-attribute correlations—and employing the Chebyshev norm to measure the similarity among data objects. Thus, the proposed approaches are adroit at mitigating the severity of the “distance concentration” problem.

7 Conclusion

In this study, we developed two similarity metrics—Chebyshev Coulomb force and resultant force—to address the crucial and intriguing challenges of “distance concentration” in high-dimensional data. The novel similarity metrics are conducive to alleviating the interference of “distance concentration” while improving the interpretability of similarity measures. Our approach not only accurately characterizes the similarity among data objects but also resembles the dimensional discrepancies by the virtue of vector projection on various dimensions. This strategy offers a multi-faceted analysis, encompassing both overall similarity and the dimension-specific variations among data objects. Additionally, the dependencies across attributes are eliminated by means of metric matrix, and it is evident that Chebyshev norm is capable of slowing down the convergence of “distance concentration.” Furthermore, Chebyshev Coulomb force unveils differences on various dimensions, whereas Coulomb resultant force reflects deviation of data objects from the dataset center in corresponding dimensions. Thus, we conclude that the proposed similarity metrics are of strong interpretability. For the next research endeavor, we will embark on and explore the application of the Chebyshev Coulomb force and resultant force in high-dimensional data mining, machine learning, and pattern recognition.

References

- [1] G. Andresini, A. Appice, and D. Malerba. 2021. Autoencoder-based deep metric learning for network intrusion detection. *Information Sciences* 569 (2021), 706–727. DOI: <https://doi.org/10.1016/j.ins.2021.05.016>
- [2] Fabrizio Angiulli. 2017. On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness. *Journal of Machine Learning Research* 18, 1 (2017), 6209–6268.
- [3] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104.
- [4] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. 2001. Searching in metric spaces. *ACM Computing Surveys* 33, 3 (2001), 273–321. DOI: <https://doi.org/10.1145/502807.502808>
- [5] Shuo Chen, Lei Luo, Jian Yang, Chen Gong, Jun Li, and Heng Huang. 2019. Curvilinear distance metric learning. In *Advances in Neural Information Processing Systems*, Vol. 32, 4225–4234.
- [6] Xiang Deng and Zhongfei Zhang. 2022. Deep causal metric learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4993–5006.
- [7] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2017. Deep localized metric learning. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2017), 2644–2656. DOI: <https://doi.org/10.1109/TCSVT.2017.2711015>
- [8] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. 2018. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2780–2789.
- [9] Damien François, Vincent Wertz, and Michel Verleysen. 2007. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering* 19, 7 (2007), 873–886. DOI: <https://doi.org/10.1109/TKDE.2007.1037>
- [10] Gene H. Golub and Charles F. Van Loan. 2012. *Matrix Computations* (4th. ed.). Johns Hopkins University Press.
- [11] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, Vol. 2. IEEE, 1735–1742.
- [12] Johanna Hardin and David M. Rocke. 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis* 44, 4 (2004), 625–638. DOI: [https://doi.org/10.1016/S0167-9473\(02\)00280-3](https://doi.org/10.1016/S0167-9473(02)00280-3)
- [13] Peter E. Hart, David G. Stork, and Richard O. Duda. 2000. *Pattern Classification*. Wiley Hoboken.
- [14] Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. 2000. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th International Conference on Very Large Databases*, 506–515.

- [15] Haize Hu, Jianxun Liu, Xiangping Zhang, and Mengge Fang. 2023. An effective and adaptable K-means algorithm for big data cluster analysis. *Pattern Recognition* 139 (2023), 109404.
- [16] Zhanxuan Hu, Danyang Wu, Feiping Nie, and Rong Wang. 2021. Generalization bottleneck in deep metric learning. *Information Sciences* 581 (2021), 249–261. DOI: <https://doi.org/10.1016/j.ins.2021.09.023>
- [17] Wenjun Ke, Jianguo Wei, Naixue Xiong, and Qingzhi Hou. 2023. GSS: A group similarity system based on unsupervised outlier detection for big data computing. *Information Sciences* 620 (2023), 1–15. DOI: <https://doi.org/10.1016/j.ins.2022.11.078>
- [18] M. Kemertas and T. Aumentado-Armstrong. 2021. Towards robust bisimulation metric learning. In *Advances in Neural Information Processing Systems*, Vol. 34, 4764–4777.
- [19] Murat Kirişci. 2023. New cosine similarity and distance measures for Fermatean fuzzy sets and TOPSIS approach. *Knowledge and Information Systems* 65, 2 (2023), 855–868. DOI: <https://doi.org/10.1007/s10115-022-01776-4>
- [20] Xin Li, Yanqin Bai, Yaxin Peng, Shaoyi Du, and Shihui Ying. 2018. Nonlinear semi-supervised metric learning via multiple kernels and local topology. *International Journal of Neural Systems* 28, 2 (2018), 1750040.
- [21] Muthurasu Nallappan and Rajasekar Velswamy. 2024. Exploring deep learning-based content-based video retrieval with hierarchical navigable small world index and ResNet-50 features for anomaly detection. *Expert Systems with Applications* 247 (2024), 123197.
- [22] Bac Nguyen, Francesc J. Ferri, Carlos Morell, and Bernard De Baets. 2019. An efficient method for clustered multi-metric learning. *Information Sciences* 471 (2019), 149–163. DOI: <https://doi.org/10.1016/j.ins.2018.08.055>
- [23] Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 186–193. DOI: <https://doi.org/10.1145/1835449.1835482>
- [24] Noelia Rico, Pedro Huidobro, Agustina Bouchet, and Irene Díaz. 2022. Similarity measures for interval-valued fuzzy sets based on average embeddings and its application to hierarchical clustering. *Information Sciences* 615 (2022), 794–812. DOI: <https://doi.org/10.1016/j.ins.2022.10.028>
- [25] Indradyumna Roy, Rishi Agarwal, Soumen Chakrabarti, Anirban Dasgupta, and Abir De. 2023. Locality sensitive hashing in fourier frequency domain for soft set containment search. In *Advances in Neural Information Processing Systems*, Vol. 36, 56352–56383.
- [26] Yibang Ruan, Yanshan Xiao, Zhifeng Hao, and Bo Liu. 2021. A nearest-neighbor search model for distance metric learning. *Information Sciences* 552 (2021), 261–277. DOI: <https://doi.org/10.1016/j.ins.2020.11.054>
- [27] Imen Souiden, Mohamed Nazih Omri, and Zaki Brahmi. 2022. A survey of outlier detection in high dimensional data streams. *Computer Science Review* 44 (2022), 100463.
- [28] Qi Wang, Weidong Min, Qing Han, Ziyuan Yang, Xin Xiong, Meng Zhu, and Haoyu Zhao. 2021. Viewpoint adaptation learning with cross-view distance metric for robust vehicle re-identification. *Information Sciences* 564 (2021), 71–84. DOI: <https://doi.org/10.1016/j.ins.2021.02.013>
- [29] Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, 2 (2009), 207–244.
- [30] Danyang Wu, Han Wang, Zhanxuan Hu, and Feiping Nie. 2022. Improved deep metric learning with local neighborhood component analysis. *Information Sciences* 617 (2022), 165–176. DOI: <https://doi.org/10.1016/j.ins.2022.10.090>
- [31] Eric Xing, Michael Jordan, Stuart J. Russell, and Andrew Y. Ng. 2002. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, Vol. 15, 505–512.
- [32] Wenzhu Yan, Quansen Sun, Huaijiang Sun, and Yanmeng Li. 2020. Joint dimensionality reduction and metric learning for image set classification. *Information Sciences* 516 (2020), 109–124. DOI: <https://doi.org/10.1016/j.ins.2019.12.041>
- [33] Jiawei Yang, Susanto Rahardja, and Pasi Fränti. 2018. Mean-shift outlier detection. In *Proceedings of the International Conference on Fuzzy Systems and Data Mining*, IOS Press, 208–215.
- [34] H. J. Ye, D. C. Zhan, Y. Jiang, and Z. H. Zhou. 2018. What makes objects similar: A unified multi-metric learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 5 (2018), 1257–1270. DOI: <https://doi.org/10.1109/TPAMI.2018.2829192>
- [35] Han-Jia Ye, De-Chuan Zhan, Xue-Min Si, and Yuan Jiang. 2016. Learning feature aware metric. In *Proceedings of the Asian Conference on Machine Learning*. PMLR, 286–301.
- [36] Shihui Ying, Zhijie Wen, Jun Shi, Yaxin Peng, Jigen Peng, and Hong Qiao. 2017. Manifold preserving: An intrinsic approach for semisupervised distance metric learning. *IEEE Transactions on Neural Networks and Learning Systems* 29, 7 (2017), 2731–2742. DOI: <https://doi.org/10.1109/TNNLS.2017.2691005>
- [37] Shichao Zhang, Jiaye Li, and Yangding Li. 2023. Reachable distance function for KNN classification. *IEEE Transactions on Knowledge and Data Engineering* 35, 7 (2023), 7382–7396. DOI: <https://doi.org/10.1109/TKDE.2022.3185149>
- [38] P. Zhu, C. Zhang, X. Li, J. Zhang, and X. Qin. 2023. A high-dimensional outlier detection approach based on local coulomb force. *IEEE Transactions on Knowledge and Data Engineering* 35, 6 (2023), 5506–5520.

- [39] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5, 5 (2012), 363–387.

Appendices

A Proof for Theorem 2

For the first term in Theorem 2, one can derive that

$$\sqrt{n} \cdot \frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)} = \frac{\sqrt{\text{Var}(\|X\|_p)}}{n^{1/p-1/2}} \Bigg/ \frac{E(\|X\|_p)}{n^{1/p}} \quad (\text{A1})$$

To proceed, taking the limit of both sides of Equation (A1) and with the aid of Lemmas 1 and 2, we conclude that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sqrt{n} \cdot \frac{\sqrt{\text{Var}(\|X\|_p)}}{E(\|X\|_p)} \\ &= \lim_{n \rightarrow \infty} \frac{\sqrt{\text{Var}(\|X\|_p)}}{n^{1/p-1/2}} \Bigg/ \lim_{n \rightarrow \infty} \frac{E(\|X\|_p)}{n^{1/p}} \\ &= \frac{\sigma_{\mathcal{F},p}}{p \cdot \mu_{\mathcal{F},p}^{(p-1)/p}} \Bigg/ \mu_{\mathcal{F},p}^{1/p} = \frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}} \end{aligned} \quad (\text{A2})$$

Hence, the statement of Theorem 2 is proved.

B Proof for Corollary 1

Employing the assumptions of Corollary 1, if X_i obeys the standard normal distribution, it follows that

$$\begin{aligned} \mu_{\mathcal{F},p} &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} |x|^p e^{-\frac{x^2}{2}} dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x^p e^{-\frac{x^2}{2}} dx \\ &= \frac{2^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right), \end{aligned}$$

where $\Gamma(z)$ is a gamma function, $\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$. Because $\sigma_{\mathcal{F},p}^2 = \mu_{\mathcal{F},2p} - \mu_{\mathcal{F},p}^2$, we obtain

$$\begin{aligned} \sigma_{\mathcal{F},p}^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} |x|^{2p} e^{-\frac{x^2}{2}} dx \\ &\quad - \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} |x|^{2p} e^{-\frac{x^2}{2}} dx \right)^2 \\ &= \frac{2^p}{\sqrt{\pi}} \Gamma\left(\frac{2p+1}{2}\right) - \left(\frac{2^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \right)^2. \end{aligned}$$

Next, combining the aforementioned elements into Equation (11), we deduce that

$$\begin{aligned} \mathcal{K}_{\mathcal{F},p} &= \frac{1}{p} \cdot \frac{\sigma_{\mathcal{F},p}}{\mu_{\mathcal{F},p}} \\ &= \frac{1}{p} \cdot \sqrt{\frac{2^p}{\sqrt{\pi}} \Gamma\left(\frac{2p+1}{2}\right) - \left(\frac{2^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right)\right)^2} \\ &\quad \left/ \frac{2^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \right. \end{aligned} \quad (\text{A3})$$

By applying the logarithms on both sides of Equation (A3), we have

$$\begin{aligned} \ln \mathcal{K}_{\mathcal{F},p}(p) &= \frac{1}{2} \ln \left(\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2 \right) \\ &\quad - \ln \Gamma\left(\frac{p+1}{2}\right) - \ln p \end{aligned} \quad (\text{A4})$$

when $p = 2$, the inequality $\frac{1}{2} \ln \left(\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2 \right) > \ln \Gamma\left(\frac{p+1}{2}\right) > \ln p$ holds. Utilizing the derivative of the gamma function, $\frac{d\Gamma(z)}{dz} = (z-1) \Gamma(z)$, we know that the gamma function grows at a faster rate. This knowledge allows us to approximate as follows: $\frac{1}{2} \ln \left(\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2 \right) \gg \ln \Gamma\left(\frac{p+1}{2}\right) \gg \ln p$. Therefore, we conclude that

$$\ln \mathcal{K}_{\mathcal{F},p}(p) \approx \frac{1}{2} \ln \left(\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2 \right). \quad (\text{A5})$$

Continuing the analysis, in view of the term $\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2$, we compute its first-order partial derivatives. The derivative is given by

$$\begin{aligned} &\left(\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2 \right)' \\ &= \sqrt{\pi} \left(\frac{2p-1}{2} \right) \Gamma\left(\frac{2p-1}{2}\right) \\ &\quad - \left(\frac{p-1}{2} \right) \Gamma\left(\frac{p-1}{2}\right) \Gamma\left(\frac{p+1}{2}\right) \\ &= \sqrt{\pi} \left(\frac{2p-1}{2} \right) \int_0^\infty t^{\frac{2p-3}{2}} e^{-t} dt \\ &\quad - \left(\frac{p-1}{2} \right) \int_0^\infty t^{\frac{p-3}{2}} e^{-t} dt \int_0^\infty t^{\frac{p-1}{2}} e^{-t} dt. \end{aligned}$$

This expression represents the first-order partial derivatives of the given term. When $p > 2$, $2p - 3 > p - 3$, $p - 1$, and $2p - 3 > p - 3 + p - 1$, then we have $\int_0^\infty t^{\frac{2p-3}{2}} e^{-t} dt > \int_0^\infty t^{\frac{p-3}{2}} e^{-t} dt \int_0^\infty t^{\frac{p-1}{2}} e^{-t} dt$. Thus, we can easily obtain

$$\left(\sqrt{\pi} \Gamma\left(\frac{2p+1}{2}\right) - \Gamma\left(\frac{p+1}{2}\right)^2 \right)' > 0. \quad (\text{A6})$$

As the value of p increases beyond 2 ($p > 2$), the relative variance demonstrates a consistent upward trend, suggesting a monotonically increasing behavior. In particular, within the range

$0 < p \leq 1$, $\mathcal{K}_{\mathcal{F},p}(1) \leq \mathcal{K}_{\mathcal{F},p}(p) < \lim_{p \rightarrow 0} \mathcal{K}_{\mathcal{F},p}(p) \approx 1.11$; $\mathcal{K}_{\mathcal{F},p}(1) \approx 0.76$; $\mathcal{K}_{\mathcal{F},p}(2) \approx 0.71$; $\mathcal{K}_{\mathcal{F},p}(6) \approx 1.12$. Therefore, when it comes to $p \rightarrow \infty$, the relative variance tends to converge toward its maximum value.

Thus, Corollary 1 is validated.

C Proof for Property 1

(1) Positiveness. By applying Equation (14), we have

$$\left\| \overline{CF}(X_i, X_j) \right\| = \max_{k=1}^n (|L_k^T(X_i - X_j)|).$$

It is evident that for any data objects X_i and X_j , the expression $|L_k^T(X_i - X_j)| \geq 0$ holds true. Thus, we establish

$$\max_{k=1}^n (|L_k^T(X_i - X_j)|) \geq 0. \quad (A7)$$

This confirms that the condition of the positiveness is satisfied.

(2) Reflexivity. In the case where $\left\| \overline{CF}(X_i, X_j) \right\| = 0$, taking Equation (14) into account, we have

$$\max_{k=1}^n (|L_k^T(X_i - X_j)|) = 0. \quad (A8)$$

for any data objects X_i, X_j , there is $|L_k^T(X_i - X_j)| \geq 0$, so if Equation (A8) holds, then for $\forall k \in [1, n]$, $|L_k^T(X_i - X_j)| = 0$, we have to guarantee that

$$X_i = X_j$$

This establishes that the condition of the reflexivity is satisfied.

(3) Symmetry. According to Equation (14), we find that

$$\begin{aligned} \left\| \overline{CF}(X_i, X_j) \right\| &= \max_{k=1}^n (|L_k^T(X_i - X_j)|) \\ \left\| \overline{CF}(X_j, X_i) \right\| &= \max_{k=1}^n (|L_k^T(X_j - X_i)|) \end{aligned}$$

for $\forall k \in [1, n]$, $|L_k^T(X_i - X_j)| = |L_k^T(X_j - X_i)|$, thus

$$\max_{k=1}^n (|L_k^T(X_i - X_j)|) = \max_{k=1}^n (|L_k^T(X_j - X_i)|). \quad (A9)$$

This establishes the condition of the symmetry.

(4) Triangular inequality. Thanks to Equation (14), we can derive that

$$\begin{aligned} \left\| \overline{CF}(X_i, X_j) \right\| &= \max_{k=1}^n (|L_k^T(X_i - X_j)|) \\ \left\| \overline{CF}(X_i, X_t) \right\| &= \max_{k=1}^n (|L_k^T(X_i - X_t)|) \\ \left\| \overline{CF}(X_t, X_j) \right\| &= \max_{k=1}^n (|L_k^T(X_t - X_j)|). \end{aligned}$$

Applying the triangle inequality, we deduce that

$$\begin{aligned}
 & \max_{k=1}^n (|L_k^T(X_i - X_j)|) \\
 &= \max_{k=1}^n (|L_k^T(X_i - X_t + X_t - X_j)|) \\
 &\leq \max_{k=1}^n (|L_k^T(X_i - X_t)| + |L_k^T(X_t - X_j)|)
 \end{aligned} \tag{A10}$$

Combining this with Equation (14), we arrive at

$$\begin{aligned}
 \|\vec{CF}(X_i, X_j)\| &\leq \|\vec{CF}(X_i, X_t)\| \\
 &\quad + \|\vec{CF}(X_t, X_j)\|
 \end{aligned} \tag{A11}$$

Thus, we confirm that the condition of the triangle inequality is satisfied.

Received 19 March 2024; revised 23 August 2024; accepted 25 January 2025