electronics

MDPI

Article

Research on the Method of Hypergraph Construction of Information Systems Based on Set Pair Distance Measurement

Jing Wang [1,2], Siwu Lan [3], Xiangyu Li [3], Meng Lu [3], Jingfeng Guo [1,*], Chunying Zhang [3,*] and Bin Liu [4]

[1] College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China; wj2018jsj@stumail.ysu.edu.cn
[2] Basic Teaching Department, Tangshan University, Tangshan 063210, China
[3] College of Science, North China University of Science and Technology, Tangshan 063210, China; lansw@stu.ncst.edu.cn (S.L.); lixiangyu@stu.ncst.edu.cn (X.L.); olimeng@stu.ncst.edu.cn (M.L.)
[4] Big Data and Social Computing Research Center, Hebei University of Science and Technology, Shijiazhuang 050018, China; liubin@hebust.edu.cn
* Correspondence: jfguo@ysu.edu.cn (J.G.); zchunying@ncst.edu.cn (C.Z.)

Abstract: As a kind of special graph of structured data, a hypergraph can intuitively describe not only the higher-order relation and complex connection mode between nodes but also the implicit relation between nodes. Aiming at the limitation of traditional distance measurement in high-dimensional data, a new method of hypergraph construction based on set pair theory is proposed in this paper. By means of dividing the relationship between data attributes, the set pair connection degree between samples is calculated, and the set pair distance between samples is obtained. Then, on the basis of set pair distance, the combination technique of k-nearest neighbor and ε radius is used to construct a hypergraph, and high-dimensional expression and hypergraph clustering are demonstrated experimentally. By performing experiments on different datasets on the Kaggle open-source dataset platform, the comparison of cluster purity, the Rand coefficient, and normalized mutual information are shown to demonstrate that this distance measurement method is more effective in high-dimensional expression and exhibits a more significant performance improvement in spectral clustering.

Keywords: high-dimensional data; set pair distance; hypergraph construction; high-dimensional representation; hypergraph spectral clustering

Citation: Wang, J.; Lan, S.; Li, X.; Lu, M.; Guo, J.; Zhang, C.; Liu, B. Research on the Method of Hypergraph Construction of Information Systems Based on Set Pair Distance Measurement. *Electronics* **2023**, *12*, 4375. https://doi.org/10.3390/electronics12204375

Academic Editor: Shinichi Yamagiwa

Received: 27 August 2023
Revised: 17 October 2023
Accepted: 18 October 2023
Published: 23 October 2023

1. Introduction

Machine learning is one of the important research fields in computer science [1]. With the development of artificial intelligence technology, machine learning has attracted wide attention with many methods having been proposed, and it has been successfully applied in various practical systems [2–5]. However, there are still many challenging problems to be solved. In the past few years, machine learning methods based on graph structure have attracted more and more attention, mainly because of their inherent advantages. Compared with vector-based machine learning methods, machine learning methods based on graph structure can effectively capture the spatial, topological, and functional relations of data and can dig out the inherent relations hidden in information systems and express them intuitively [6].

As a kind of special graph of structured data, a hypergraph can better intuitively describe not only the higher-order relation and complex connection pattern between nodes but also the implicit relation between nodes. For example, in a paper collaboration network, an ordinary graph can only express the implicit relationship between two authors who co-write an article; the implicit relationship between several authors who co-write an article cannot be expressed. In other words, it is difficult or even impossible for ordinary graphs to distinguish the interaction between samples within various structures. In a hypergraph,

a hyperedge can contain any node, and the implicit relationship between nodes can be expressed intuitively in the hypergraph. If the author is regarded as a node, and the paper co-authored by several authors is regarded as a supersede, the hypergraph can intuitively represent this cooperative relationship. Therefore, at present, hypergraphs are more widely used in data mining [7,8], social network analysis [9–11], recommendation system [12–14], and other fields.

In the application of hypergraphs, the use of an effective hypergraph construction method plays an important role in the construction of a hypergraph and has a direct impact on the structure and performance of that hypergraph. In the process of hypergraph construction, the most important consideration is the distance measurement between nodes. A suitable distance measurement algorithm can help reveal the similarity and correlation between nodes so as to effectively build the hypergraph.

Distance measurement is used to learn a metric matrix that can effectively reflect the distance between data samples by training a given sample set so that the distribution of similar samples in the new feature space is tighter and the distribution of heterogeneous samples more dispersed [15]. The traditional distance measurement methods, such as Euclidean distance and cosine distance, are used to construct hypergraphs, and good results are obtained in many cases. However, as the field of data science and machine learning evolves, we are faced with more and more complex data types, and traditional distance metrics are often poorly suited to the task [16–18]. As shown in Figure 1, with a continuous increase in data dimensions, the calculation of Euclidean distance and cosine distance is affected by the so-called "dimensional disaster" problem. The distances between samples are not particularly stable, and the distances between samples will become nearly equal [19,20], resulting in the observer's inability to effectively distinguish the differences between different samples.

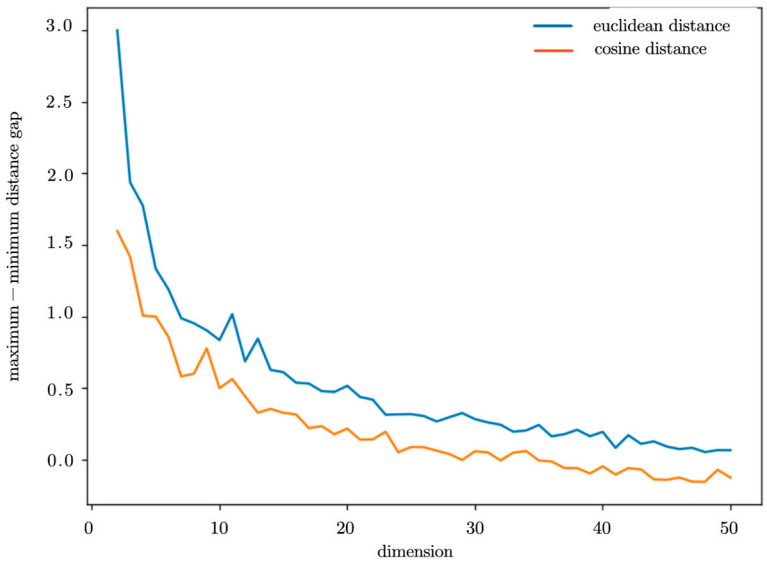

Figure 1. The high-dimensional representation of Euclidean distance and cosine distance.

Therefore, researchers began to propose a variety of new distance measurements to meet the needs of different data types and problems. Tao Yang et al. [21] proposed a distance measurement algorithm based on DTW for segmented time series data, which effectively reduced the time complexity of traditional hierarchical clustering and improved the performance of the algorithm. Considering the global structural information of the data, Guang Junye et al. [22] calculated the effective distances between the data using sparse reconstruction technology, thus replacing the traditional Euclidean distance, and applied it to the classical clustering algorithm, which significantly improved the clustering effect compared with Euclidean distance. Based on rough set theory, Liang Jiye et al. [23] proposed a new distance measurement method to measure the difference between two

attribute values under the same classification attribute and applied it to the traditional K-Modes clustering algorithm. Experiments showed that the distance measurement was more effective. Li Jipu et al. [24] proposed a new metric function considering the probability of neighboring points becoming neighboring points and applied it in LPP and *k*-nearest neighbor classifications, which can minimize the intra-class divergence and maximize the inter-class separation degree better than the traditional Euclidean distance. Existing clustering classification algorithms for uncertain data do not take into account the problem of consistency between possible worlds. Liu Han et al. [25] proposed a clustering classification framework for uncertain data based on similarity matrix consistency learning. Based on the consistency principle of possible worlds, the framework improved the performance of clustering and classification by minimizing the inconsistency of similarity matrices between different possible worlds. Although there has been a lot of research work on distance measurement, there is still a problem, in that the existing distance measurement may not be able to fully explore the potential relationships of the data where high-dimensional complex data are involved [26–29].

In addition, a reasonable and effective method for constructing hypergraphs has a direct impact on their structure and performance. A *k*-nearest neighbor or ε-radius neighborhood partition method is easy to implement [30–32], so these are widely used in hypergraph construction. However, when considering the problem of data density or sparsity, both methods have their limitations. In the sparse region, although nodes are far apart, the *k*-nearest neighbor method will still edge a node to its nearest neighbor node. In this case, the neighbor nodes of this node may contain different nodes. Unreasonable ε values can also result in disconnected nodes, subgraphs, or single nodes. Therefore, in order to improve the quality and reliability of hypergraph structure, it is also very important to explore a new method for constructing hypergraphs.

Data points in high-dimensional space have more freedom and variability, resulting in the distribution of data becoming sparser and more diffuse, thus increasing the uncertainty of the data. Set pair theory [33] is a system analysis method for uncertain information proposed by the Chinese scholar Zhao Keqin in 1989, which carries out dialectical analysis and mathematical treatment on the certainty and uncertainty in the system [34–36]. Precisely because it has the characteristic of better measuring the uncertainty relationship, set pair theory provides a good method for distance measurement in high dimensions. Based on this consideration, a new distance measurement method based on the set pair connection degree is proposed. The set pair connection degree of the two data nodes is calculated by studying the same, different, and inverse relationships among the attributes of the data nodes, which are then converted into the node distance. By comparing different distance algorithms and hyperedge construction methods, hypergraphs are constructed on multiple high-dimensional data sets, and spectral clustering is evaluated and analyzed experimentally.

2. Materials and Methods

In this section, the theoretical knowledge used in this paper is briefly introduced, and the hot issues solved and studied by these theories are expounded.

2.1. The Basic Theory of Set Pairwise

In 1989, Zhao Keqin creatively proposed the set pair analysis theory for the problem of uncertainty. A set pair is a pair composed of two sets with certain connection. Set pair theory is a certain uncertainty theory, which studies the relationship between two sets from three aspects: identity (same), difference (different), and opposition (negative). It treats certainty and uncertainty as a whole, and its core idea is that in a certain uncertain system, certainty and uncertainty are interrelated, influence and restrict each other, and can be transformed under certain conditions. The two sets of the set pair are analyzed, and the expression $u = a + bi + cj$ is established to describe the random fuzzy uncertainty problem.

A correlation expression is a mathematical expression. Based on the same difference opposition between two sets in the problem studied, the association expression reduces

the abstract problem of uncertainty to a mathematical expression. Under the requirement of problem W, there are two sets A and B with an uncertain relation, and A and B have N properties. The sets A and B can be constructed into set pair $H(A,B)$, and the uncertainty expression in set pair $H(A,B)$ can be expressed as

$$u(H) = \frac{S}{N} + \frac{F}{N}\, i + \frac{P}{N} j \tag{1}$$

In the equation, $u(H)$ represents the connectivity coefficient between the set pair $H(A,B)$. N refers to the number of characteristics in the set pair. S denotes the number of shared characteristics between the two sets in the set pair. P represents the number of opposing characteristics relative to the two sets in the set pair. $F = N - S - P$ characteristics are neither opposing nor identical. The coefficient i represents the uncertainty of difference, with a range of values from -1 to 1. The coefficient j represents the degree of opposition, with a fixed value of -1. Let $a = S/N$ be the degree of identity between sets A and B. Let $b = F/N$ be the degree of difference between sets A and B. Let $c = P/N$ be the degree of opposition between sets A and B. The formula that expresses the connectivity coefficient can be denoted as $u(H) = a + bi + cj$.

With continuous and systematic in-depth research on the theory of set pair analysis, set pair theory is promoted and extended in different fields and has been extensively applied in fields including mathematical analysis, physics, earth sciences, life sciences, information science, and management science. In China, Fengchao Liu [37] utilized the set pair analysis method to construct an evaluation index system for regional independent innovation capability. It has also been applied to the eight major economic regions in China to analyze the independent innovation strengths and characteristics of each region, which is of significant importance for future development. Fei Su [38] used the entropy method to determine the weights of evaluation indicators such as the sensitivity and response capability to the gradual depletion of exploitable resources. The set pair analysis method was employed to construct an assessment model for economic vulnerability, which was then applied to the oil city of Daqing. In foreign countries, Peng Zhang [39] utilized the set pair analysis method to comprehensively evaluate the performance of nano SiO_2 and PVA fiber-reinforced polymer mortar. The set pair analysis method was used to conduct a standardized and quantitative evaluation of various aspects such as mechanical properties, durability, and processability. Weichao Yu [40] proposed an assessment and prediction method for the vulnerability of natural gas supply chains based on set pair analysis and provided recommendations to address the vulnerability of China's natural gas supply chain. Rui Wang [41] developed an airport bird-strike risk assessment model based on pentuple correlation coefficients, which can accurately predict risk trends. This model is of great significance for airport personnel in carrying out bird-strike prevention work.

2.2. The Foundation Theory of Hypergraphs

A hypergraph is an extension of a graph, where each edge is not limited to connecting two vertices but can be connected to 1 to n nodes, known as hyperedges. Each hyperedge represents a set of data points, so hypergraphs can represent more complex relationships among objects.

According to the characteristics of a hypergraph, a hypergraph with N nodes and M hyperedges can be defined as follows: $H = (V, E, W)$. In a hypergraph, $V = \{v_1, v_2, ..., v_N\}$ represents the set of nodes, $E = \{e_1, e_2, ..., e_M\}$ represents the set of hyperedges, and the diagonal matrix W represents the weights of the hyperedges. In the case of a hypergraph, the adjacency matrix used for simple graphs is no longer suitable. Instead, we define an incidence matrix H as the mathematical representation of a hypergraph. When a node $v \in V$ is incident to a hyperedge $e \in E$ in a hypergraph, we represent this relationship by setting $H_{ve} = 1$. A hypergraph contains high-order information that is missing in traditional graphs, primarily reflected in the higher-order relationships between hyperedges. The degree matrix of a hypergraph can be divided into the hyperedge degree matrix and the

node degree matrix. The hyperedge degree represents the number of nodes contained within a hyperedge. The degree matrix of hyperedges is defined as $B_{ee} = \sum_{v=1}^{N} H_{ve}$. The degree of a hyperedge represents the number of nodes it contains. The degree matrix of nodes is defined as $D_{vv} = \sum_{e=1}^{M} W_{ee} H_{ve}$, where W_{ee} is the weight of the associated hyperedge. The degree of a node represents the number of hyperedges it is incident to. Both the degree matrix of hyperedges (B_{ee}) and the degree matrix of nodes (D_{vv}) are diagonal matrices, where $D \in R^{N \times N}$ and $B \in R^{M \times M}$.

To address the issue of information loss in a regular undirected graph, one can construct a hypergraph. For example, consider the problem of modeling collaboration relationships among authors in academic papers. By constructing an undirected graph where vertices represent papers and edges connect two vertices if they have at least one common author, one can optimize the graph further by assigning edge weights equal to the number of shared authors. However, in cases where an author has written three or more papers, this approach may still result in some information loss. To overcome this, one can construct a hypergraph where vertices represent papers and hyperedges represent all the papers associated with a particular author. This way, the issue of information loss in a regular graph can be addressed.

Hypergraphs, as an extension of graph theory, enable the representation of multivariate relationships and higher-order relationships, providing more powerful modeling capabilities [42–47]. Wang Shen [48] developed an online social network information propagation model by combining hypergraph-based network topology and an improved SIR model. This model better adapts to online social networks and provides a theoretical basis for studying the propagation and governance of information in such networks. Ling Tian [49] proposed a three-layer architecture for knowledge hypergraphs, aiming to better represent and extract hyper-relational features. This approach enables the efficient modeling of hyper-relational data and facilitates rapid knowledge inference. Peiyan Wang [50] designed a knowledge hypergraph link prediction model based on tensor decomposition. This model effectively models the roles of entities in different relations and positions, providing a highly effective solution to the problem of knowledge hypergraph link prediction. Cola Vincenzo Schiano di [51] modeled electronic health appointment data as a hypergraph structure and utilized machine learning algorithms to analyze and mine these data. The analysis results were used to improve the management and service optimization of electronic health systems. Xiang Gao [52] proposed a seizure detection method based on hypergraph features and machine learning. This method enables the accurate detection of epileptic seizures. This research is of great significance for improving the diagnosis and treatment of epilepsy patients. It can be seen from Refs. [53–63], hypergraphs find extensive applications in various fields such as social network analysis, knowledge graph construction, machine learning, and more.

3. Distance Measurement Based on Set Pairwise Theory

3.1. Machine Learning Methods Based on Hypergraphs

Graph structures are crucial for information encoding, from bioinformatics to computer vision, as the prevalence of complex graph-structured data continues to grow. Data represented in graph structures contain more information compared to data represented in vector form (i.e., information system data) [64–68]. In ordinary graphs, the edges between nodes only reflect a certain relationship between two nodes. In hypergraphs, the 'hyperedges' can contain an arbitrary number of nodes and can reflect relationships that exist among multiple nodes. Machine learning methods for information systems can be classified into three categories: vector-based machine learning methods, network-based machine learning methods, and hypergraph-based machine learning methods [69–72]. The distinction among these three machine learning methods lies in the form in which the information system data are transformed, whether into vectors, graphs, or hypergraphs, followed by the application of different machine learning techniques for data training. The illustration of these three methods is shown in Figure 2.

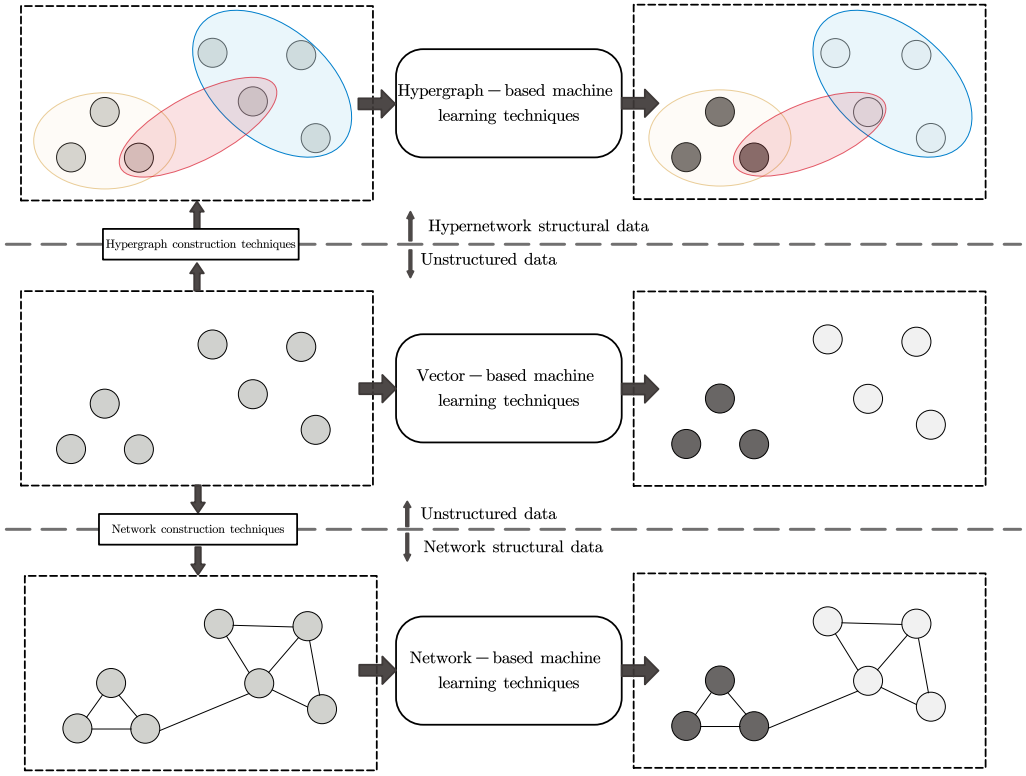

Figure 2. Three machine learning methods for information systems.

According to the general representation method of machine learning tasks, a collection of N data samples denoted as $D = \{x_1, x_2, \ldots, x_N\}$ (non-network structured data), transforming the sample set D into a hypergraph structure $H = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes in the hypergraph and $E = \{E_1, E_2, \ldots, E_m\}$ is the set of hyperedges in the hypergraph. The set $E_i = \{v_{i1}, v_{i2}, \ldots, v_{ij}\}, (i = 1, 2, \ldots, m; j = 1, 2, \ldots, N)$ is referred to as the hyperedges of the hypergraph. The transformation process establishes a mapping: $D \rightarrow H = (V, E)$. However, the key to hypergraph learning is the ability to transform a collection of data samples into a hypergraph structure, specifically, the ability to obtain the hyperedges of the hypergraph. Typically, two factors are considered to determine whether to establish a hyperedge: the similarity function and the hypergraph construction technique.

Effectively measuring the similarity between nodes and hyperedges is an important problem. Traditional similarity measurement methods often fail to capture the high-order structure and relationship of hypergraphs. Therefore, it is necessary to design new measurement methods to improve clustering accuracy and robustness. By incorporating the connectivity measurements and concepts from set pair analysis theory into hypergraph construction, we can depict the similarities, differences, and inverse relationships between nodes, thereby obtaining distance measurement methods between nodes. This paper proposes a novel distance measurement method based on set pair connectivity, which calculates the set pair connectivity between two nodes by studying the similarities, differences, and inverse relationships between their attribute values. Due to the correspondence between distance and connectivity, that is, smaller distances correspond to higher connectivities, and greater distances correspond to lower connectivities, the connectivity can be transformed into a distance measurement method by performing some simple transformations.

3.2. Node Pair Distance

Node pair distance is a variation of node pair connectivity. Each data instance in the dataset is treated as a node, with a node representing a set of attribute values across multiple dimensions. Differing from traditional similarity calculation methods, this paper introduces

the concept of node pair in node similarity calculation and provides the following definition for the connectivity between nodes:

$$\mu(v_i, v_j) = a + b * i + c * j \tag{2}$$

Among these, $\mu(v_i, v_j)$ represents the connectivity between node v_i and node v_j, where μ is within the range of $[-1,\ 1]$. A larger value of μ indicates a higher similarity, while a smaller value indicates a higher dissimilarity; a represents the degree of agreement between node v_i and node v_j, b represents the degree of uncertainty, and c represents the degree of opposition between node v_i and node v_j. It is required that $a + b + c = 1$. i is the marker for uncertainty, and j is the marker for opposition. During the calculation, both i and j participate as coefficients, with j always taking the value -1. The value of i can vary within the range of $[-1,\ 1]$ depending on the specific situation.

With regard to the degree of agreement a, uncertainty b, and opposition c in $\mu(v_i, v_j)$, this paper considers the node attributes and focuses on the node pair composed of v_i and v_j. The following descriptions are provided for the same attribute S, the different attribute F, and the opposing attribute P between v_i and v_j.

(1) The representation of $S(v_i, v_j)$ denotes the set of similar attributes between v_i and v_j, which is calculated as follows:

$$S(v_i, v_j) = \{x_k \mid \left(\left|v_{ik} - v_{jk}\right|\right) < \alpha\} \tag{3}$$

(2) The representation of $F(v_i, v_j)$ denotes the set of uncertain attributes between v_i and v_j, which is calculated as follows:

$$F(v_i, v_j) = \{x_k \mid \alpha < \left(\left|v_{ik} - v_{jk}\right|\right) < \beta\} \tag{4}$$

(3) The representation of $P(v_i, v_j)$ denotes the set of opposing attributes between v_i and v_j, which is calculated as follows:

$$P(v_i, v_j) = \{x_k \mid \left(\left|v_{ik} - v_{jk}\right|\right) > \beta\} \tag{5}$$

Then, $\mu(v_i, v_j)$ can be represented as:

$$\mu(v_i, v_j) = \frac{\left|S(v_i, v_j)\right|}{K} + \frac{\left|F(v_i, v_j)\right|}{K} * i + \frac{\left|P(v_i, v_j)\right|}{K} * j \tag{6}$$

where x_k represents the k-dimensional attribute, v_{ik} represents the k-dimensional attribute value of the i-th node, α is the similarity boundary, β is the opposition boundary, and K represents the number of attributes in the dataset. By using the above equation, the pairwise association degree between v_i and v_j can be calculated. A higher association degree indicates a greater similarity between v_i and v_j, while a lower degree indicates a lesser similarity.

By performing the calculations, we can obtain the matrix VM that represents the pairwise association degree between nodes as follows:

$$VM = \begin{bmatrix} - & \mu_{v_1v_2} & \mu_{v_1v_3} & \cdots & \mu_{v_1v_n} \\ \mu_{v_2v_1} & - & \mu_{v_2v_3} & \cdots & \mu_{v_2v_n} \\ \mu_{v_3v_1} & \mu_{v_3v_2} & - & \cdots & \mu_{v_3v_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{v_nv_1} & \mu_{v_nv_2} & \mu_{v_nv_3} & \cdots & - \end{bmatrix} \tag{7}$$

VM is a symmetric matrix. After performing maximum–minimum normalization on VM, we obtain VM' as follows:

$$VM' = \begin{bmatrix} - & \mu'_{v_1v_2} & \mu'_{v_1v_3} & \cdots & \mu'_{v_1v_n} \\ \mu'_{v_2v_1} & - & \mu'_{v_2v_3} & \cdots & \mu'_{v_2v_n} \\ \mu'_{v_3v_1} & \mu'_{v_3v_2} & - & \cdots & \mu'_{v_3v_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu'_{v_nv_1} & \mu'_{v_n}2 & \mu'_{v_nv_3} & \cdots & - \end{bmatrix} \tag{8}$$

By subtracting VM' from a homogeneous matrix J consisting of all ones, we obtain the set pairwise distance matrix VD between nodes as follows:

$$VD = J - VM' = \begin{bmatrix} - & d_{v_1v_2} & d_{v_1v_3} & \cdots & d_{v_1v_n} \\ d_{v_2v_1} & - & d_{v_2v_3} & \cdots & d_{v_2v_n} \\ d_{v_3v_1} & d_{v_3v_2} & - & \cdots & d_{v_3v_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{v_nv_1} & d_{v_nv_2} & d_{v_nv_3} & \cdots & - \end{bmatrix} \tag{9}$$

3.3. Distance Evaluation

The evaluation of distance measurement methods depends on specific application scenarios and requirements. Typically, evaluations can be conducted based on criteria such as rationality, non-negativity, symmetry, and applicability.

(1) In terms of rationality analysis, a good distance measurement method should align with common sense and intuition, effectively quantifying the differences or similarities between objects in a reasonable manner. In the case of pairwise distances derived from the pairwise association degree between nodes, it can adequately measure the degree of dissimilarity between nodes. Therefore, it is reasonable to consider it as a distance measurement method.

(2) In terms of non-negativity analysis, the range of values for the pairwise association degree is $[-1, 1]$. However, during the conversion to pairwise distance, normalization is performed to map the resulting distances to the interval [0, 1]. Therefore, the pairwise association degree satisfies the non-negativity criterion.

(3) In terms of symmetry analysis, the pairwise association degree of two nodes involved in the calculation is equal. Therefore, it satisfies the symmetry criterion.

(4) In terms of applicability analysis, this paper conducts experimental evaluations on the pairwise distance from the following two aspects:

- Distance algorithm high-dimensional representation

In high-dimensional spaces, traditional distance calculations become difficult and unreliable, especially for Euclidean distance and cosine similarity. Specifically, in high-dimensional spaces, the farthest and nearest distances between any points tend to converge to be nearly equal. Therefore, it is necessary to design a metric that measures the difference between the farthest and nearest distances, as shown in the following formula:

$$diff = lg\left(\frac{max(dist) - min(dist)}{min(dist)}\right) \tag{10}$$

To assess the set pairwise distance proposed in this paper, an experiment was conducted where 500 randomly generated data points were evaluated. The dimensions of the data ranged from 2 to 50. The differences between the maximum and minimum distances were calculated for Euclidean distance, cosine distance, and pairwise distance. The results are illustrated in Figure 3.

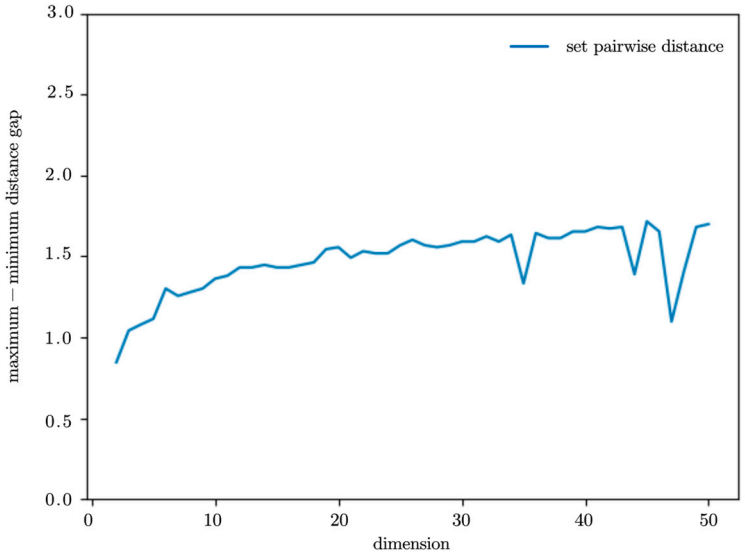

Figure 3. High-dimensional representation of pairwise distance.

By studying Figure 3, it can be observed that, with the increase in dimensions, the set pairwise distance no longer exhibits the phenomenon of gradually decreasing difference between the maximum and minimum distances as seen in Euclidean distance and cosine similarity. Instead, it consistently demonstrates an overall stable trend with fluctuations, indicating a significant difference. This suggests that the set pairwise distance remains effective in high-dimensional environments.

- Cluster effectiveness evaluation

The evaluation of a distance metric also requires consideration of its performance in specific applications. This paper focuses on the construction and application of hypergraphs. Therefore, different distances can be used to construct hypergraphs and perform hypergraph spectral clustering. Ultimately, the performance of different distance metrics can be compared based on cluster purity.

4. Hypergraph Construction of Information Systems Based on Set Pairwise Distance

4.1. Hyperedge Construction Based on knn and ε-Radius

The traditional process of constructing a hypergraph is relatively simple and usually involves using either the knn or ε-radius to determine the hyperedges for each node. In the ε-radius method, all nodes within the ε-radius of a research node are grouped into a single hyperedge. On the other hand, Using k-nearest neighbor method to construct superedge is to classify the research node and its k nearest neighbor nodes into one superedge.

However, both of these methods have limitations when it comes to handling dense or sparse data. In sparse regions, even if the nodes are far apart, the knn method still connects a node to their k closest neighbors. In such cases, the neighbors of a node may include dissimilar nodes. Similarly, using an inappropriate ε value can result in unconnected nodes, subgraphs, or isolated nodes.

To combine the advantages of these two methods, a combination technique is employed to construct hyperedges. If we denote the neighborhood of v_i as $N(v_i)$, then:

$$N(v_i) = \begin{cases} \varepsilon\_radius(v_i), |\varepsilon\_radius(v_i)| > k \\ k\_NN(v_i), |\varepsilon\_radius(v_i)| \leq k \end{cases} \tag{11}$$

where $\varepsilon\_radius(v_i)$ returns the set of nodes in the ε-neighborhood of node v_i and $k\_NN(v_i)$ represents the set of k-nearest neighboring nodes of v_i. The ε-radius method is used for dense regions, while the knn method is used for sparse regions. The construction process can be described as follows: compute the number of nodes within the ε-radius of node v_i. If

it is greater than k, then all the nodes within the ε-radius of v_i form a hyperedge. Otherwise, v_i is connected to its k-nearest neighbors to form a hyperedge, thereby determining the hyperedge to which each node belongs.

As shown in Figure 4, the ε-neighborhood of node a is $\{b, c\}$, and its k-nearest neighbors are $\{b, c, d\}$. Therefore, the hyperedge to which node a belongs is $\{a, b, c, d\}$. Similarly, the ε-neighborhood of node f is $\{c, d, e, g\}$, and its k nearest neighbors are $\{c, d, g\}$. Thus, the hyperedge to which node f belongs is $\{c, d, e, f, g\}$.

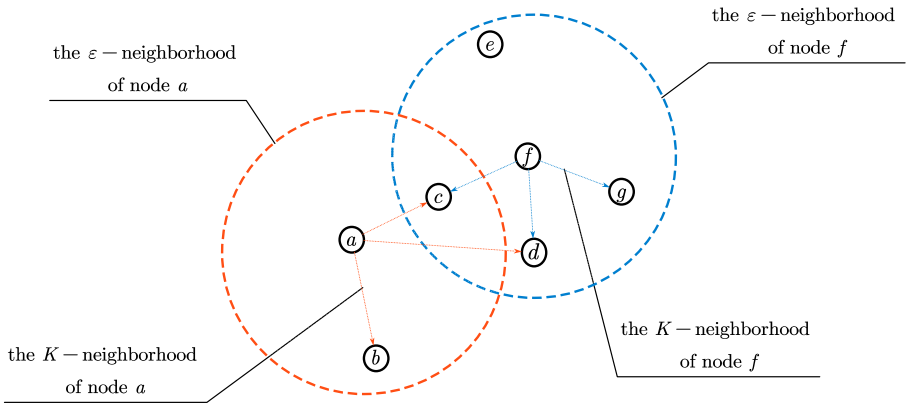

Figure 4. Neighborhood partitioning using knn and ε-radius.

4.2. Hypergraph Construction Based on Node Set Pairwise Distance

The process of constructing a hypergraph based on set pair distance is shown in Figure 5. Starting from the node set V, we iterate through each node v_i and use the set pair distance matrix to obtain its ε_set of ε-neighborhood and k_set of knn. Using the combination technique of k-nearest neighbors and ε-radius, we determine the hyperedge e_i to which v_i belongs. We then check if e_i is already included in the current set of hyperedges E. If it is, we move on to the next node; otherwise, we add e_i to E. We continue this process until all nodes have been traversed, resulting in a complete set of hyperedges E. We then assign a weight ω to each hyperedge, resulting in the weight matrix W. Finally, we obtain the hypergraph $HG(V, E, W)$.

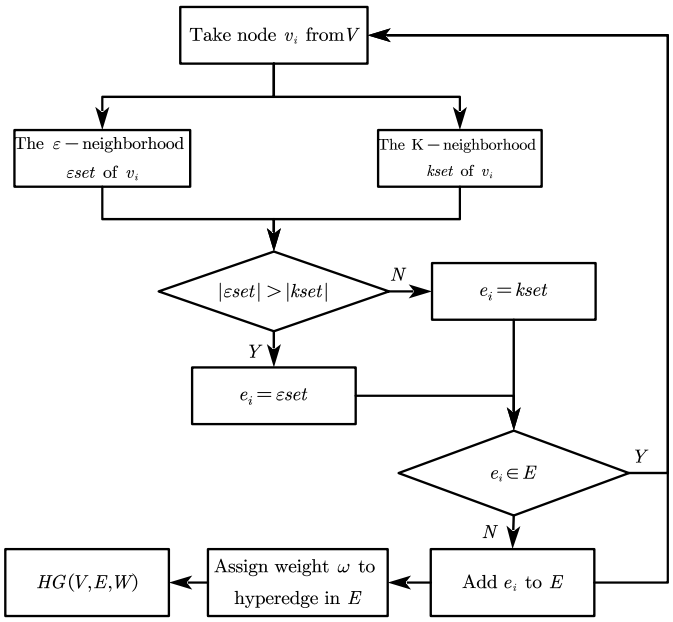

Figure 5. Hypergraph construction process based on set pair distance.

From the algorithm flow in Figure 5, it can be seen that for the construction process of a hypergraph $HG(V, E, W)$, the time complexity is mainly reflected in the traversal of nodes and the allocation of hyperedge weights. Therefore, the time complexity reaches $O(|V|+|E|)$, and the spatial complexity is mainly reflected in the storage of the distance matrix. Therefore, the spatial complexity is $O\left(|V|^2\right)$, which is generally within an acceptable range of performance.

4.3. Experimental Analysis of Hypergraph Spectral Clustering

4.3.1. Dataset Introduction

To demonstrate the performance of the algorithm on various datasets, this paper conducted experiments using datasets of different scales, all of which were obtained from the Kaggle open dataset platform. Table 1 provides an overview of the datasets used in this study.

Table 1. The details of data sets.

Data Sets	Samples	Attribute	Class
conversion_predictors	273	16	2
breast_cancer	569	30	2
dermatology	358	34	6
icr_processed	541	60	2
cortex_nuclear	552	80	8

4.3.2. Experimental Comparison

In hypergraph spectral clustering, the construction of the hypergraph often directly influences the clustering results. This paper conducted experimental comparisons regarding two aspects: different hypergraph construction methods and distance metrics between nodes.

Comparison of Clustering Experiments Using Different Distance Metrics

When constructing the hypergraph, distance is commonly used as a measure. In order to further explore the advantages of pairwise distance, this study conducted spectral clustering based on hypergraphs constructed using Euclidean distance (euc_dist), cosine distance (cos_dist), and pairwise distance (spc_dist). By comparing the clustering purity, Rand coefficient, and normalized mutual information under the best parameter settings, the effectiveness of pairwise distance was demonstrated.

From Tables 2–4 below, it can be observed that regardless of whether low-dimensional or high-dimensional data are involved, the hypergraphs based on set pairwise distance exhibit better performance in spectral clustering. In particular, for the cortex_nuclear dataset, the clustering effect using set pairwise distance is significantly superior to that achieved using other distances. Although some results on other datasets are not as good as those of other distance-based spectral clustering, the difference is not significant, indicating that spectral clustering based on set pairwise distance is also effective.

Table 2. Clustering purity under different distances for spectral clustering.

Data Sets	euc_dist	cos_dist	spc_dist
conversion_predictors	0.623	0.630	**0.729**
dermatology	0.860	**0.905**	0.866
breast_cancer	**0.944**	0.856	0.902
icr_processed	0.880	**0.887**	0.880
cortex_nuclear	0.875	0.810	**0.920**

Table 3. Rand index under different distances for spectral clustering.

Data Sets	euc_dist	cos_dist	spc_dist
conversion_predictors	0.055	0.063	**0.207**
dermatology	0.851	0.868	**0.869**
breast_cancer	**0.786**	0.505	0.640
icr_processed	**0.395**	0.393	0.331
cortex_nuclear	0.838	0.754	**0.900**

Table 4. Normalized mutual information under different distances for spectral clustering.

Data Sets	euc_dist	cos_dist	spc_dist
conversion_predictors	0.065	0.076	**0.203**
dermatology	0.911	**0.929**	0.842
breast_cancer	**0.720**	0.417	0.597
icr_processed	0.241	**0.252**	**0.252**
cortex_nuclear	0.901	0.887	**0.945**

Comparison of Clustering Experiments Using Different Hypergraph Construction Methods

In this study, the construction of hypergraphs involves two hyperparameters, namely, the number of neighbors k and the radius ε. To explore their influence on hypergraph construction, experiments were conducted on five datasets to adjust these hyperparameters. The experimental comparison and analysis were performed on different hypergraph construction methods, including the ones based on knn, ε-radius, and a combination of *k*-nearest neighbor and ε-radius techniques. The spectral clustering results were analyzed, and the clustering results are shown in Figures 6–8.

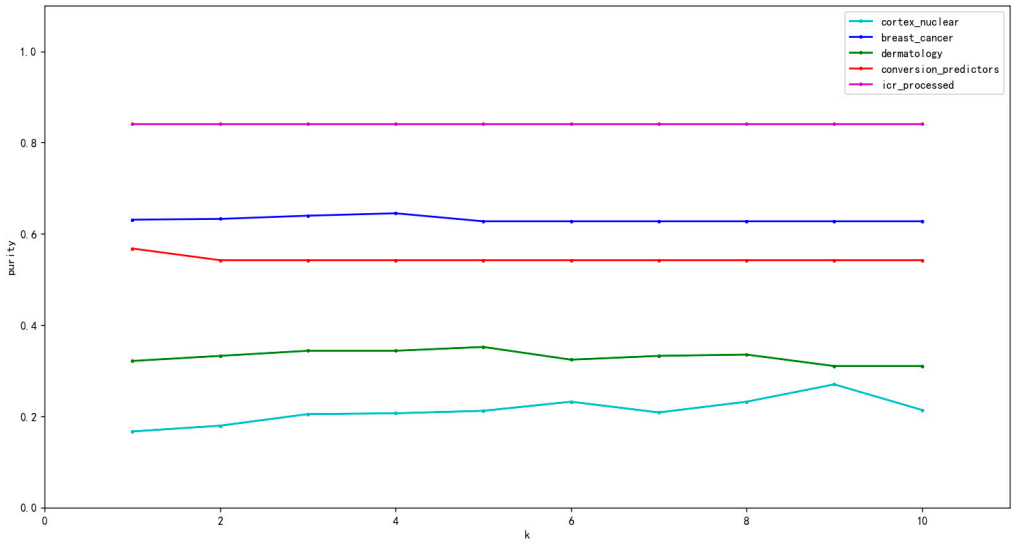

Figure 6. Hypergraph spectral clustering results based on knn.

Based on the observed change curves in Figure 6, the conversion_predictors dataset exhibits the highest clustering purity at k = 1, with a value of 0.568. In the case of the breast_cancer dataset, the highest spectral clustering purity is achieved at k = 4, with a value of 0.645. For the dermatology dataset, the highest spectral clustering purity is observed at k = 5, with a value of 0.352. The icr_processed dataset shows consistent spectral clustering purity regardless of the k value, with a value of 0.841. Finally, for the cortex_nuclear dataset, the highest spectral clustering purity is observed at k = 9, with a value of 0.270.

Based on the observed change curves in Figure 7, the conversion_predictors dataset exhibits the highest spectral clustering purity at ε = 0.63, with a value of 0.722. For the

breast_cancer dataset, the highest spectral clustering purity is achieved at ε = 0.59, with a value of 0.791. In the case of the dermatology dataset, the highest spectral clustering purity is observed at ε = 0.12, with a value of 0.838. The icr_processed dataset shows the highest spectral clustering purity at ε = 0.2, with a value of 0.861. Finally, for the cortex_nuclear dataset, the highest spectral clustering purity is observed at ε = 0.12, with a value of 0.632.

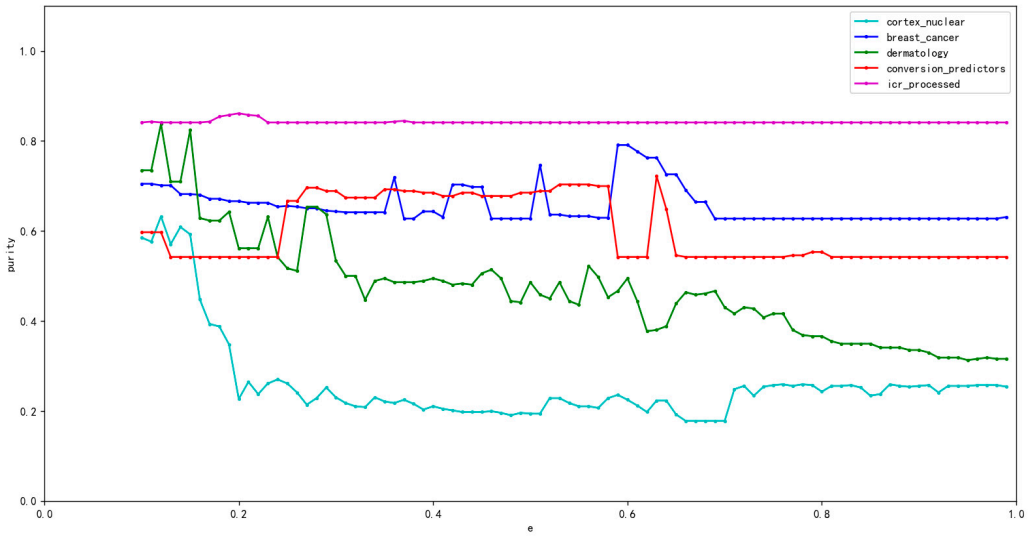

Figure 7. Hypergraph spectral clustering results based on ε-radius.

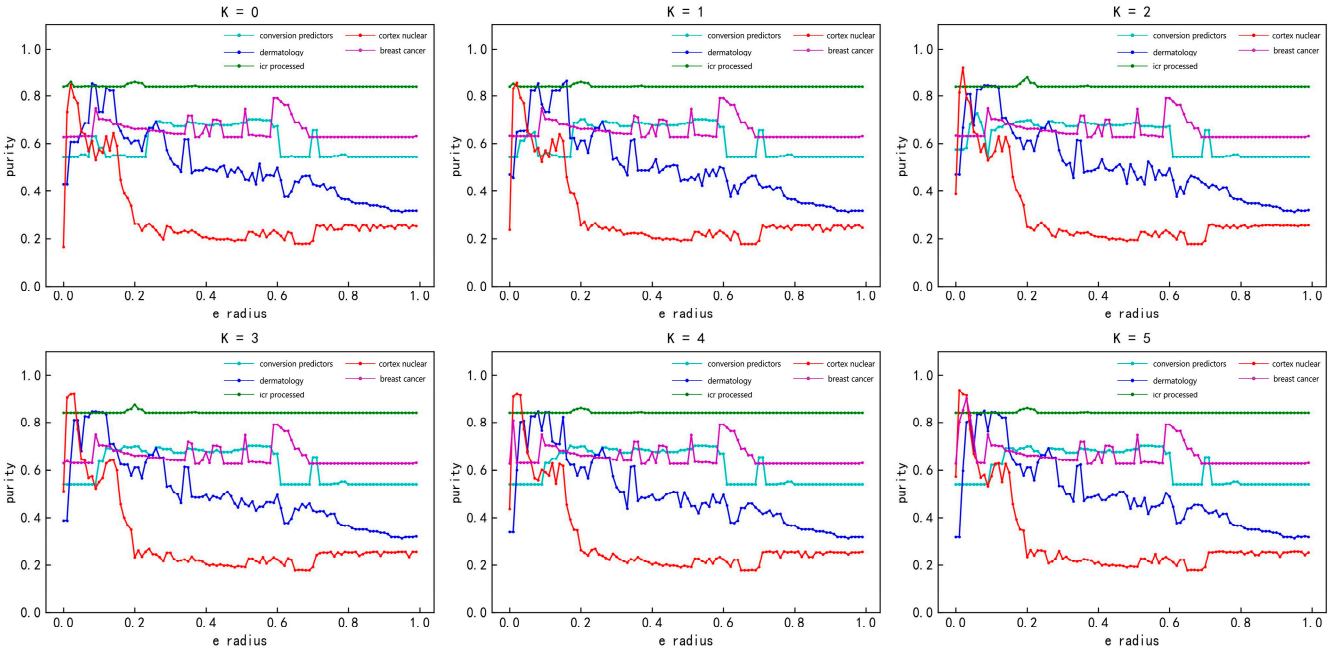

Figure 8. Hypergraph spectral clustering based on knn and ε-radius combination technique.

Based on the observed change curves in Figure 8, the conversion_predictors dataset exhibits the highest spectral clustering purity at k = 2 and ε = 0.06, with a value of 0.729. For the breast_cancer dataset, the highest spectral clustering purity is achieved at k = 5 and ε = 0.03, with a value of 0.902. In the case of the dermatology dataset, the highest spectral clustering purity is observed at k = 1 and ε = 0.16, with a value of 0.866. The icr_processed dataset shows the highest spectral clustering purity at k = 2 and ε = 0.2, with a value of 0.880. Finally, for the cortex_nuclear dataset, the highest spectral clustering purity is observed at k = 2 and ε = 0.02, with a value of 0.920. It is of note that as the dimension increases, the advantages of this construction method in clustering purity are more obvious.

5. Conclusions

In this paper, a method of constructing an information system hypergraph based on set pair distance is given by combining set pair theory. The set-to-set distance is evaluated theoretically through rationality, non-negativity, symmetry, and application performance. Then, the combined technique of k-nearest neighbor and ε-radius is used to construct the hypergraph. By comparing the maximum and minimum distances of different distance measurement methods, the effectiveness of set pair distance in high-dimensional data is demonstrated. In addition, spectral clustering experiments are carried out on the basis of hypergraph construction, and the validity of the set-to-distance measurement method is demonstrated by a comparison of cluster purity, Rand coefficient, and normalized mutual information. Although there are many high-dimensional data clustering algorithms, there is no algorithm that can be generally applied to all fields, and the current high-dimensional data clustering algorithms need to be improved. Therefore, high-dimensional data clustering remains an important direction for research.

Author Contributions: Conceptualization, J.W.; methodology, J.W.; software, S.L.; validation, S.L. and X.L.; writing—original draft preparation, S.L., X.L. and M.L.; writing—review and editing, J.W. and C.Z.; visualization, S.L.; supervision, J.G. and C.Z.; project administration, J.G., C.Z. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the S&T Program of Hebei (No.20310802D) and the National Cultural and Tourism Science and Technology Innovation Project (2020).

Data Availability Statement: The article does not create data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suo, Q.; Guo, J.L. Hypernetworks: Structure and Evolution Mechanisms Based on Hypergraphs. *Syst. Eng. Theory Pract.* **2017**, *37*, 720–734.
2. Kajino, H. Molecular Hypergraph Grammar with Its Application to Molecular Optimization. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 3183–3191.
3. Xia, X.Y.; Liu, Y.S.; Ding, Y.; Hong, Y. Granular Ball Computing Classifiers for Efficient, Scalable and Robust Learning. *Inf. Sci.* **2019**, *483*, 136–152. [CrossRef]
4. Xia, X.Y.; Zhang, H.; Li, W.H.; Wang, G.Y. A Novel Rough Set Algorithm for Fast Adaptive Attribute Reduction in Classification. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 1231–1242. [CrossRef]
5. Liu, H.; Latecki, L.J.; Yan, S. Dense subgraph partition of positive hypergraphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 541–554. [CrossRef]
6. Li, Y. Research on Three-way Decision Community Detection Based on Variable Granularity. Master's Thesis, Anhui University, Hefei, China, 2020. [CrossRef]
7. Cui, Y.; Yang, B.R. Several Applications of Hypergraphs in the Field of Data Mining. *Comput. Sci.* **2010**, *37*, 220–222.
8. Tie, J.; Chen, W.; Sun, C.; Mao, T.; Xing, G. The Application of Agglomerative Hierarchical Spatial Clustering Algorithm in Tea blending. *Clust. Comput.* **2019**, *22*, 6059–6068. [CrossRef]
9. Xiao, Y.Z.; Zhao, H.X. User Behavior Analysis in Online Social Networks Based on Hypergraph Theory. *J. Comput. Appl. Softw.* **2014**, *31*, 50–54.
10. Gupta, S.; Kumar, P. An Overlapping Community Detection Algorithm Based on Rough Clustering of Links. *Data Knowl. Eng.* **2020**, *125*, 101777. [CrossRef]
11. Fuentes, I.; Pina, A.; Nápoles, G.; Rosete, A. Rough Net Approach for Community Detection Analysis in Complex Network. In Proceedings of the International Joint Conference on Rough Set, Bratislava, Slovakia, 19–24 September 2020; Springer: Cham, Switzerland, 2020; pp. 401–415.
12. Ma, H.F.; Zhang, D.; Zhao, W.Z.; Shi, Z. Weibo Recommendation Method Based on Hypergraph Random Walk Label Expansion. *J. Softw.* **2019**, *30*, 3397–3412. [CrossRef]
13. Kejani, M.T.; Dornaika, F.; Talebi, H. Graph convolution networks with manifold regularization for semi-supervised learning. *Neural Netw.* **2020**, *127*, 160–167. [CrossRef] [PubMed]
14. Dang, N.M.; Anh, T.L. Textual Manifold-based Defense Against Natural Language Adversarial Examples. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6612–6625.
15. Shen, Y.Y.; Yan, Y.; Wang, H.Z. Research Progress on Supervised Distance Metric Learning Algorithms. *Acta Autom. Sin.* **2014**, *40*, 2673–2686.

16. Zou, P.C.; Wang, J.D.; Yang, G.Q. Time Series Distance Metric Learning with Auxiliary Information Generation. *J. Softw.* **2013**, *24*, 2642–2655. [CrossRef]
17. Zhang, C.Y.; Gao, R.Y.; Liu, F.C.; Wang, J. Set Pair K-means Clustering Algorithm for Incomplete Information Systems. *Data Acquis. Process* **2020**, *35*, 613–629. [CrossRef]
18. Zhang, C.Y.; Gao, R.Y.; Wang, J.H.; Chen, S.; Liu, F.C.; Ren, J.; Feng, X.Z. MD-SPKM: A set pair k-modes clustering algorithm for incomplete categorical matrix data. *Intell. Data Anal.* **2021**, *25*, 1507–1524. [CrossRef]
19. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is "nearest neighbor" meaningful? In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 217–235.
20. Tang, J.; Liu, J.; Zhang, M.; Mei, Q. Visualizing Large-scale and High-dimensional Data. In Proceedings of the 25th International Conference on World Wide Web, Geneva, Switzerland, 11–15 May 2016; pp. 287–297.
21. Tao, Y.; Deng, X.; Yang, F.Y. Hierarchical Clustering Algorithm Based on DTW Distance Metric. *Comput. Eng. Des.* **2019**, *40*, 116–121. [CrossRef]
22. Guang, J.Y.; Liu, M.X.; Zhang, D.Q. Application of Effective Distance in Clustering Algorithms. *J. Comput. Sci. Explor.* **2017**, *11*, 406–413.
23. Liang, J.Y.; Bai, L.; Cao, F.Y. K-Modes Clustering Algorithm Based on New Distance Measure. *J. Comput. Res. Dev.* **2010**, *47*, 1749–1755.
24. Li, J.P.; Zhao, R.Z. Application of Nearest Probability Distance in Classification of Rotary Machinery Fault Sets. *J. Vib. Shock* **2018**, *37*, 48–54. [CrossRef]
25. Han, L.; Zhang, X.C.; Zhang, X.T.; Cui, Y. Self-adapted Mixture Distance Measure for Clustering Uncertain Data. *Knowl.-Based Syst.* **2017**, *126*, 33–47.
26. Zhang, C.Y.; Gao, R.Y.; Fan, Y.X.; Wang, L.; Pei, T. Set Pair Granular Hierarchical Clustering Algorithm for Incomplete Data. *J. Mini-Micro Syst.* **2021**, *42*, 522–530.
27. Cheng, D.; Zhu, Q.; Huang, J.; Wu, Q.; Yang, L. A Hierarchical Clustering Algorithm Based on Noise Removal. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 1591–1602. [CrossRef]
28. Brown, D.; Japa, A.; Shi, Y. An Attempt at Improving Density-based Clustering Algorithms. In Proceedings of the 2019 ACM Southeast Conference (ACM SE'19), Kennesaw, GA, USA, 18–20 April 2019; pp. 172–175.
29. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
30. Rezaei, M. Improving a Centroid-based Clustering by Using Suitable Centroids from Another Clustering. *J. Classif.* **2020**, *37*, 352–365. [CrossRef]
31. Ma, J.; Jiang, X.; Gong, M. Two-phase Clustering Algorithm with Density Exploring Distance Measure. *CAAI Trans. Intell. Technol.* **2018**, *3*, 59–64. [CrossRef]
32. Kowalskip, A.; Lukasik, S.; Charythanowicz, M.; Kulczycki, P. Nature Inspired Clustering–use Cases of Krill Herd Algorithm and Flower Pollination Algorithm. In *Interactions Between Computational Intelligence and Mathematics Part 2*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 83–98.
33. Zhao, K.Q.; Xuan, A.L. Set Pair Theory: A New Method and Application of Uncertainty Theory. *Syst. Eng.* **1996**, *1*, 18–23+72.
34. Zhang, C.Y.; Guo, J.F. α-Relation Community and Dynamic Mining Algorithm of Set Pair Social Network. *Chin. J. Comput.* **2013**, *36*, 1682–1692. [CrossRef]
35. Zhang, C.Y.; Ren, J.; Liu, L.; Liu, S.; Li, X. Set Pair Three-Way Overlapping Community Discovery Algorithm for Weighted Social Internet of Thing. *Digit. Commun. Netw.* **2022**, *9*, 3–13. [CrossRef]
36. Guo, J.F.; Dong, H.; Zhang, T.W.; Chen, X. Network Embedding of Topic-Attention Network Based on Set Pair Analysis. *Int. J. Innov. Comput. Inf. Control* **2020**, *16*, 1371–1384.
37. Liu, F.C.; Pan, X.F.; Shi, D.G. Research on Evaluation of Regional Autonomous Innovation Capability Based on Set Pair Analysis. *Method China Soft Sci.* **2005**, *11*, 83–91+106.
38. Su, F.; Zhang, P.Y. Vulnerability Assessment of Economic Systems in Daqing City Based on Set Pair Analysis. *Acta Geogr. Sin.* **2010**, *65*, 454–464.
39. Zhang, P.; Zhang, X.M.; Yuan, P.; Hu, S.W. Performance optimization of geopolymer mortar blending in nano-SiO_2 and PVA fiber based on set pair analysis. *E-Polymers* **2023**, *23*, 20230015. [CrossRef]
40. Yu, W.C.; Zheng, X.B.; Wen, F.; Li, L.; Yue, Y.Z.; Shi, F.; Yang, H.; Liu, Y.; Liu, X.B. A Methodology to Evaluate the Vulnerability of the Natural Gas Supply Chain Based on Set Pair Analysis and Markov Chain. *J. Pipeline Syst. Eng. Pract.* **2023**, *14*, 04023015. [CrossRef]
41. Wang, R.; Zhao, Q.; Sun, H.; Zhang, X.D.; Wang, Y.Y. Risk Assessment Model Based on Set Pair Analysis Applied to Airport Bird Strikes. *Sustainability.* **2022**, *14*, 04023015. [CrossRef]
42. Whang, J.J.; Du, R.; Jung, S.; Lee, G.; Drake, B. MEGA: Multi-view semi-supervised clustering of hypergraphs. *Proc. VLDB Endow.* **2020**, *13*, 698–711. [CrossRef]
43. Purkait, P.; Chin, T.J.; Sadri, A.; Suter, D. Clustering with Hypergraphs: The Case for Large Hyperedges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1697–1711. [CrossRef]
44. Wang, T.; Lu, Y.; Han, Y. Clustering of High Dimensional Handwritten Data by an Improved Hypergraph Partition Method. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2017.

45. Kumar, T.; Vaidyanathan, S.; Ananthapadmanabhan, H.; Parthasarathy, S. Hypergraph Clustering: A Modularity Maximization Approach. *arXiv* **2018**, arXiv:1812.10869.
46. Amburg, I.; Veldt, N.; Benson, A.R. Clustering in graphs and hypergraphs with categorical edge labels. *arXiv* **2019**, arXiv:1910.09943.
47. Hayashi, K.; Aksoy, S.G.; Park, C.H. Hypergraph Random Walks, Laplacians, and Clustering. *arXiv* **2020**, arXiv:2006.16377.
48. Shen, W.; Shi, Q.R.; Wang, J.Y.; Li, H. Research on Information Diffusion Model of Online Social Networks Based on Hypergraph. *J. China Soc. Sci. Tech. Inf.* **2023**, *42*, 354–364.
49. Tian, L.; Zhang, J.C.; Zhang, J.H.; Zhou, X. Overview of Knowledge Graphs: Representation, Construction, Reasoning, and Knowledge Hypergraph Theory. *J. Comput. Appl.* **2021**, *41*, 2161–2186.
50. Wang, P.Y.; Duan, L.; Guo, Z.S.; Zhou, X. Knowledge Hypergraph Link Prediction Model Based on Tensor Decomposition. *J. Comput. Res. Dev.* **2021**, *58*, 1599–1611.
51. Cola Vincenzo, S.d.; Chiaro, D.; Prezioso, E.; Izzo, S.; Giampaolo, F. Insight extraction from e-Health bookings by means of Hypergraph and Machine Learning. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4649–4659. [CrossRef] [PubMed]
52. Gao, X.; Zhu, Y.; Yang, Y.; Zhang, F. A seizure detection method based on hypergraph features and machine learning. *Biomed. Signal Process Control* **2022**, *77*, 103769. [CrossRef]
53. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
54. Mai, X.; Cheng, J.; Wang, S. Research on semi supervised k-means clustering algorithm in data mining. *Clust. Comput.* **2019**, *22*, 3513–3520. [CrossRef]
55. Zhang, Z.; Lin, H.; Gao, Y. Dynamic Hypergraph Structure Learning. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Switzerland, 13–19 July 2018; pp. 3162–3169.
56. Ji, Z.; Fu, Z.Q.; Zhang, S.T. Fault Diagnosis of Diesel Generator Set Based on Optimized NRS and Complex Network. *J. Vib. Shock* **2020**, *39*, 246–251+260. [CrossRef]
57. Wu, Z.; Pan, S.; Chen, F.; Long, G. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef]
58. Zhang, Z.; Cui, P.; Zhu, W. Deep Learning on Graphs: A survey. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 249–270. [CrossRef]
59. Zhang, W.B.; Zhu, G.Y. A Multi Objective Optimization of PCB Prototyping Assembly with OFA Based on the Similarity of Intuitionistic Fuzzy Sets. *IEEE Trans. Fuzzy Syst.* **2020**, *29*, 2054–2061. [CrossRef]
60. Xia, S.Y.; Wu, S.L.; Chen, X.X.; Wang, G.; Gao, X.; Zhang, Q. Accurate and Efficient Neighborhood Rough Set for Feature Selection. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 9281–9294. [CrossRef]
61. Kang, Z.; Wen, L.; Chen, W.; Xu, Z. Low-rank Kernel Learning for Graph-based Clustering. *Knowl.-Based Syst.* **2019**, *163*, 510–517. [CrossRef]
62. Odili, J.B.; Noraziah, A.; Ambar, R.; Abd Wahab, M.H. A Critical Review of Major Nature-inspired Optimization Algorithms. *Eurasia Proc. Sci. Technol. Eng. Math.* **2018**, *2*, 376–394.
63. Wang, R.; Nguyen, T.T.; Li, C.; Yang, Z. Optimising Discrete Dynamic Berth Allocations in Seaports Using a Levy Flight Based Meta-heuristic. *Swarm Evol. Comput.* **2019**, *44*, 1003–1017. [CrossRef]
64. Kostopoulos, G.; Karlos, S.; Kotsiantis, S.; Ragos, O. Semi-supervised regression: A recent review. *J. Intell. Fuzzy Syst.* **2018**, *35*, 1483–1500. [CrossRef]
65. Veldt, N.; Benson, A.R.; Kleinberg, J. Minimizing Localized Ratio Cut Objectives in Hypergraphs. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 23–27 August 2020; pp. 1708–1718.
66. Thai, L.; Park, N.; Lee, D.Y. Shield: Defending Textual Neural Networks Against Multiple Black-box Adversarial Attacks with Stochastic Multi-expert Patcher. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1, pp. 6661–6674.
67. Yu, W.C.; Zheng, C.; Cheng, W.; Song, D.; Zong, B. Learning Deep Network Representations with Adversarially Regularized Autoencoders. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London: Association for Computing Machinery, London, UK, 19–23 August 2018; pp. 2663–2671.
68. Zhang, Q. Research on Overlapping Community Detection Algorithm Based on Rough Set. Master's Thesis, Southwest Jiaotong University, Chengdu, China, 2020. [CrossRef]
69. Zhang, L.Y.; Guo, J.F.; Wang, J.Z.; Zhang, C. Hypergraph and Uncertain Hypergraph Representation Learning Theory and Methods. *Mathematics* **2022**, *10*, 1921. [CrossRef]
70. Peng, J.; Zhang, B.; Sugeng, K.A. Uncertain Hypergraphs: A Conceptual Framework and Some Topological Characteristics Indexes. *Symmetry* **2022**, *14*, 330. [CrossRef]
71. Wu, D.; Nie, F.; Lu, J.; Wang, R. Balanced Graph Cut with Exponential Inter-cluster Compactness. *IEEE Trans. Artif. Intell.* **2022**, *3*, 498–505. [CrossRef]
72. Van, G.W.; Vandenhende, E.S.; Georgoulis, S.; Proesmans, M. Scan: Learning to Classify Images without Labels. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 268–285.