applied sciences

MDPI

Article

Generative Adversarial Networks Based Heterogeneous Data Integration and Its Application for Intelligent Power Distribution and Utilization

Yuanpeng Tan [1], Wei Liu [1], Jian Su [1] and Xiaojing Bai [2,*]

[1] Beijing Key Laboratory of Distribution Transformer Energy-Saving Technology, China Electric Power Research Institute, Beijing 100192, China; tanyuanpeng@epri.sgcc.com.cn (Y.T.); liuwei75@epri.sgcc.com.cn (W.L.); sujian@epri.sgcc.com.cn (J.S.)

[2] Department of Engineering Physics, Tsinghua University, Beijing 100084, China

* Correspondence: baixiaojing@nuctech.com

Received: 31 October 2017; Accepted: 7 January 2018; Published: 11 January 2018

Featured Application: Authors are encouraged to provide a concise description of the specific application or a potential application of the work. This section is not mandatory.

Abstract: Heterogeneous characteristics of a big data system for intelligent power distribution and utilization have already become more and more prominent, which brings new challenges for the traditional data analysis technologies and restricts the comprehensive management of distribution network assets. In order to solve the problem that heterogeneous data resources of power distribution systems are difficult to be effectively utilized, a novel generative adversarial networks (GANs) based heterogeneous data integration method for intelligent power distribution and utilization is proposed. In the proposed method, GANs theory is introduced to expand the distribution of completed data samples. Then, a so-called peak clustering algorithm is proposed to realize the finite open coverage of the expanded sample space, and repair those incomplete samples to eliminate the heterogeneous characteristics. Finally, in order to realize the integration of the heterogeneous data for intelligent power distribution and utilization, the well-trained discriminator model of GANs is employed to check the restored data samples. The simulation experiments verified the validity and stability of the proposed heterogeneous data integration method, which provides a novel perspective for the further data quality management of power distribution systems.

Keywords: intelligent power distribution and utilization; heterogeneous data integration; generative adversarial networks; peak clustering; finite open coverage

1. Introduction

With the rapid development of smart grid and sensing technology, China's power user side data showed high complexity and redundancy. Since 2011, the user side data volume of power distribution system in China has been booming, from GB to TB, even to PB level, and gradually forms a big data system. Facing the era of big data, power companies have not only improved traditional MySQL, Oracle, and other relational database systems, but also produced lots of new big data systems, such as HBase, GBase, and etc. All of these database systems mentioned above have already formed a multi-source heterogeneous big data system for intelligent power distribution and utilization (IPDU) [1–4]. On the other hand, affected by local economic levels, the monitoring and testing conditions in local power companies and manufacturers for distribution network equipment are quite different, while parts of the complex monitoring and testing equipment is unreasonable and impossible to repeat purchase, leading to the further heterogeneity of IPDU big data.

The IPDU big data contains operation data and external scene information of distribution network, and supports the further decision analysis for the planning, construction, operation, maintenance, and other business sectors of power distribution system. Through the deep mining of multi-source data, it would be easy to realize the accurate analysis of current situations and future trends of distribution networks, and operation control of intelligent power distribution network [5,6]. However, the traditional data analysis and decision technologies of distribution networks require that all the data samples share the same monitoring and testing indexes, which makes the IPDU heterogeneous big data difficult to be effectively utilized, resulting in a huge waste of data resources. Because of the structure heterogeneity among the existing database systems and the limitation of data qualities from both objective and subjective factors, the IPDU big data system could not directly meet the requirements of traditional data analysis and decision technologies. At the same time, it also comes into a small sample environment for parts of samples in IPDU big data, and brings out great challenges for the data quality management of distribution networks, which makes it difficult to guarantee the accuracy of analysis and decisions [7,8]. Therefore, the study on heterogeneous data integration technology has a significant influence on the future intelligent power distribution and utilization, which can also greatly improve the accuracy and efficiency of distribution network's operation and decision-making.

So far, data integration researches in intelligent power distribution and utilization could be broadly divided into two categories, i.e., time-series data integration and index data integration. The research subjects of the former one mainly include load data, distributed power output, wind speed of wind turbine, and etc. These kinds of data integration technologies are relatively mature, and already form a relatively perfect research system, in which grey relational analysis, collaborative filtering, Markov chain, support vector machine and neural network are widely used as data analysis tools [9–12]. The research subjects of index data integration mainly include the structured and semi-structured data, in addition to time-series data. Many experts and scholars have also carried out some inspiring works in this research field, such as: Liu and et al. introduced low rank and sparsity theory for data integration to detect the false data injection in power grid [13]; Xu and et al. proposed an XLPE power cable lifetime evaluation method by employing low-rank matrix completion technology [14]; Mateos and et al. used robust nonparametric regression via sparsity control to perform data cleaning and repair tasks [15].

Although the techniques mentioned above can meet the requirements of some engineering applications in accuracy, but still not be able to satisfy the efficiency requirement. In order to deal with IPDU big data problems, some experts and scholars turned their research directions to data integration based on machine learning. Yu and et al. proposed an extreme learning machine based missing data completion method [16]. Li and Socher introduced deep learning theory to fulfill the incomplete data restoration and integration tasks, respectively [17,18]. However, these researches do not take the small sample environment of IPDU data into account, so it is difficult to be directly applied in the actual projects, and the integration of heterogeneous data in distribution network is not quite satisfying.

In order to solve the problem that heterogeneous data resources for intelligent power distribution and utilization are difficult to be effectively utilized in the small sample environment, a novel heterogeneous data integration technology that is based on generative adversarial networks (GANs-HDI) is proposed. In this proposed GANs-HDI method, the sample space expansion is realized by employing the generator of Goodfellow and et al.'s GANs [19,20], according to the targeted samples with all of the measurement indexes complete. In order to eliminate the heterogeneous characteristics, a so-called peak clustering algorithm is proposed to realize the finite open coverage of the expanded sample space, and repair those incomplete samples. Finally, the repaired samples are checked by using well-trained discriminator of GANs. By doing this, GANs learning together with clustering theory form a closed loop to improve heterogeneous data integration performance greatly. This proposed heterogeneous data integration method is helpful to realize the efficient integration of heterogeneous data, and also provides a novel perspective for the further data quality management in power companies.

2. Generative Adversarial Networks Based Sample Space Expansion

Facing the big data for intelligent power distribution and utilization (IPDU), power companies have not only improved traditional MySQL, Oracle, and other relational database systems, but also produced lots of new big data systems, such like HBase, GBase, and etc. All of these database systems mentioned above provide excessive multi-source heterogeneous samples, as shown in Figure 1. These targeted samples together form a real space, in which, according to Heine-Borel theorem, a limited number of open intervals could be chosen to form a finite open coverage of this targeted sample set. In each open interval, the samples shall hold the same data characteristics, and be able to support other samples with missing indexes. However, in some small sample environments, the samples are not always enough for data completion and integration tasks in all of the open intervals. Therefore, in order to obtain satisfying data integration results, this paper introduces generative adversarial theory to enrich the sample space.

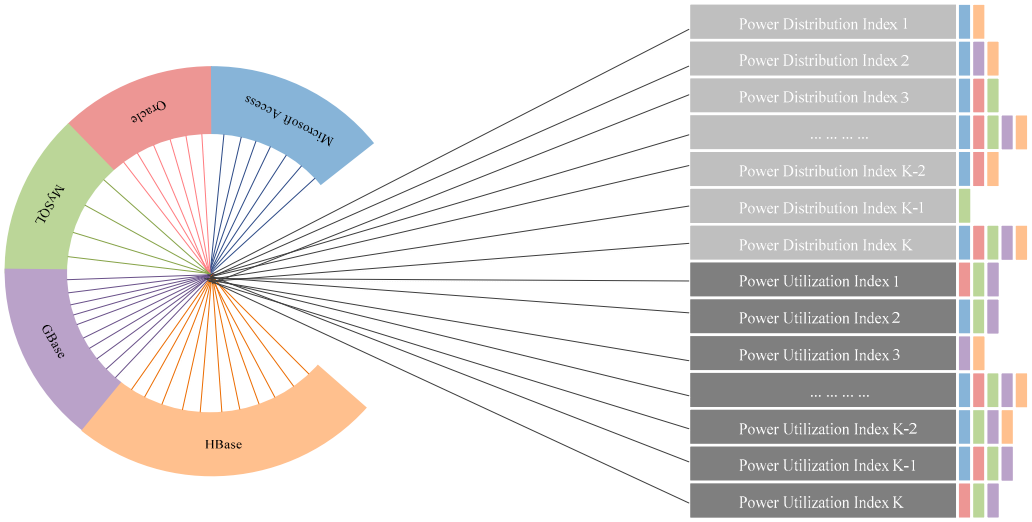


Figure 1. Multi-source heterogeneous big data system for intelligent power distribution and utilization (IPDU).

Generative adversarial networks (GANs) is a generative model derived from Nash zero-sum game, in which the generator model and discriminator model are invited to participate. The generator model is designed to learn the distribution of training data, while the discriminator is designed to estimate the probability that the targeted data sample comes from training data rather than the generator. Both of these two models could improve their performances in mutual confrontation and iterative optimization, extend the targeted sample set, improve the discrimination ability, and approach the Nash equilibrium eventually [19]. As one of the most exciting ideas in the research field of machine learning over the last decade, the theory of GANs has been widely used in image and graphic processing, natural language processing, computer virus monitoring, chess game programming, and etc.

Inspired by Goodfellow and Springenberg's works [20–22], GANs theory is employed to realize the expansion of targeted sample space in this paper. First of all, a targeted data set $D = \{d_i\}_{i=1}^{N}$ with all the measurement indexes is constructed, where N stands for the sample number of data set. GANs algorithm is used to train generator **G** and discriminator **D** in TensorFlow platform. Taking $D = \{d_i\}_{i=1}^{N}$ as inputs and zeros as outputs, the discriminator **D** could be initialized in TensorFlow as the following equation:

$$\mathbf{D}(d_i) = \sum_{j=1}^{L} \sum_{i=1}^{N} \beta_j \mathbf{g}(\omega_i^{\mathrm{T}} d_i + b_j) \tag{1}$$

where, L is the number of hidden neural nodes, $\omega_i \in R^K$ is the input weights of i-th hidden neural node, and $\beta_j \in R$ and $b_j \in R$ represent the output weights and threshold values of j-th hidden neural node respectively, $\mathbf{g}(\cdot) : R \to R$ stands for the activation function in neural networks.

Furthermore, train the generator **G** and discriminator **D** simultaneously: adjusting parameters for **G** to minimize $\log(1 - \mathbf{D}(d))$ and for **D** to minimize $\log \mathbf{D}(d)$, as if they are following the two-player min-max game with value function $v(\mathbf{G}, \mathbf{D})$ [21]:

$$\mathbf{G} = \sum_{j=1}^{L} \sum_{i=1}^{N} \beta_j \mathbf{f}((\varpi_i, \cdot) + b_j) = \underset{\mathbf{G}}{\operatorname{argmin}} \max_{\mathbf{D}} v(\mathbf{G}, \mathbf{D}) \tag{2}$$

$$s.t.\ v(\theta^{(\mathbf{G})}, \theta^{(\mathbf{D})}) = {}_{d \in \overline{D}} \log \mathbf{D}(d) + {}_{d \in \widetilde{D}} \log(1 - \mathbf{D}(d)) \tag{3}$$

where, $\widetilde{D} = \{d_i\}_{i=1}^{M}$ stands for the data set consisted of new generated samples from generator **G**; L is the number of hidden neural nodes; $\varpi_i \in R^P$ is the input weights of i-th hidden neural node, $\beta_j \in R^K$ and $b_j \in R$ represent the output weights and threshold values of j-th hidden neural node. respectively, and $\mathbf{f}(\cdot) : R \to R$ stands for the activation function in neural networks. By using the well-trained generator **G**, M new data samples could be generated with random vector set $Z = \{z_i \in R^P\}_{i=1}^{M}$ as the inputs. Take $D = \{d_i\}_{i=1}^{N}$ and $\widetilde{D} = \{d_i\}_{i=1}^{M}$ as inputs and zeros and ones as outputs respectively, train and renew the discriminator **D**.

Finally, determine whether the probability of newly generated samples falls within the interval $[0.5 - c, 0.5 + c]$ by using discriminator **D**. If this condition is satisfied, then it demonstrates that generator **G** performs well in convergence. Combine the new generated sample set $\widetilde{D}$ and original data set D, and denote the combination as $D_{GANs} = \{d_i\}_{i=1}^{N+M}$ for the future data restoration. Otherwise, the discriminative error of **D** is back propagated to retraining of generator **G**. More obviously, the calculation processing of GANs could be shown as Figure 2.

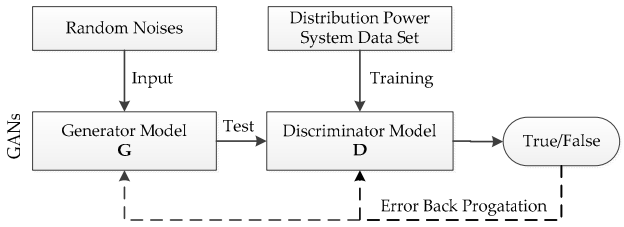
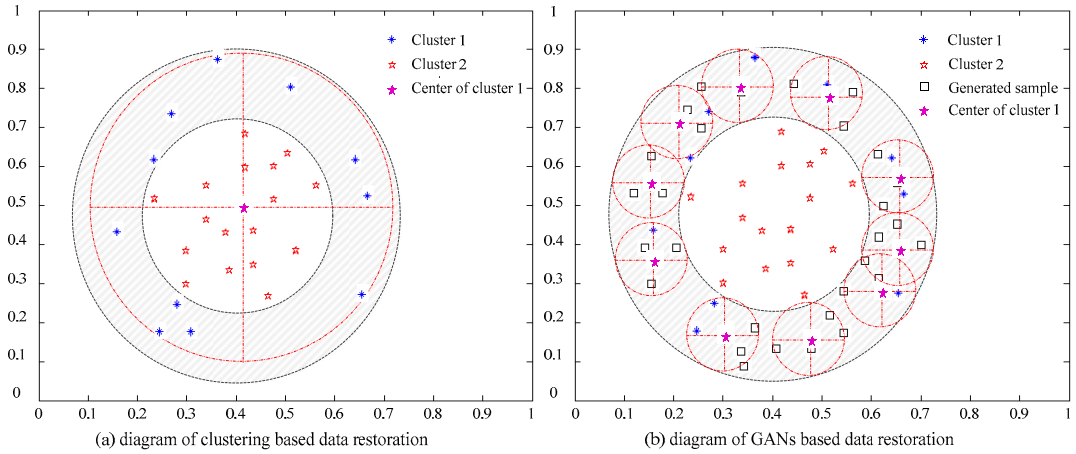


Figure 2. Diagram of generative adversarial networks (GANs) calculation processing.

3. Peak Clustering Based Data Restoration

Based on the proliferation result of data samples by introducing generator discussed above, a so-called peak clustering based incomplete data restoration method is proposed in this section. In order to overcome the restoration failures of traditional algorithms with linear inseparable data, the proposed method constructs as few as possible open intervals with a fixed neighborhood radius for all of the data samples. Then, the set of open intervals form a finite open coverage, avoiding the interference of linear inseparable data samples on clustering results, as shown in Figure 3.

Inspired by Rodriguez's work in Reference [23], peak clustering algorithm is proposed for incomplete data restoration to improve the calculation efficiency, while sustaining the restoration precision. Supposing the finite open coverage $Coverage_i(d)$ contains n_i data samples $\{d_j\}_{j=1}^{n_i}$, calculate the peak distance of density peaks (distance between data sample and density peak point) $Dist(d_j,\ Temp\_Peaks_i(d))$. Then, n_i clusters are constructed according to the phase angle with the density peak point as the center, and each cluster contains only one sample [24]. If the absolute value of the peak distance difference of the cluster with similar phase angle is smaller than or equal to the threshold value, the two classes are merged, and the distance between the peak point of the new class and the peak to peak value of the density is calculated.



(a) diagram of clustering based data restoration (b) diagram of GANs based data restoration

Figure 3. Diagram of comparative performance of data restoration.

If the absolute value of the peak distance difference of the clusters with the similar phase angle is no larger than the threshold value k, combine the two clusters and calculate new density peak points and peak distances. Repeat the operations above, until the absolute value of the peak distance difference of the clusters with the similar phase angle becomes larger than the threshold value k, or the total cluster number becomes 1. Then, end the iteration and output the clustering result of the last iteration.

Finally, after the peak clustering of targeted data samples, the weighted averages of corresponding values in complete samples could be used as the predictive values of missing data. The concept of entropy in information theory is introduced, and the weighting coefficient is determined by the similarity between data samples. Generally speaking, the process of peak clustering based data restoration could be shown in the following Table 1.

Table 1. Peak clustering based data restoration algorithm.

Inputs	Establish the combined dataset $D_{GANs} = \{d_i\}_{i=1}^{N+M}$; initialize threshold values R_1, R_2 and clustering threshold value k;
Step 1	Establish the finite open coverage of the targeted combined dataset: **Step 1.1:** if $Centre \neq \varphi$, randomly select $d \in Centre$ from the central point set; otherwise, go to Step 1.4. **Step 1.2:** calculate i-th open interval: $Coverage_i(d) = \{d' \in D_{GANs} \mid Dist(d,d') \leq R_1\}$; $Coverage = \{Coverage_i(d)\}$ and renew the temporary peak point set $Temp\_Peaks \leftarrow \{Temp\_Peaks, d\}$, $i \leftarrow i+1$. **Step 1.3:** renew the central point set $Centre \leftarrow C_{Centre}\{d' \mid Dist(d,d') \leq R_2\}$, and calculate the peak point set: $Temp\_Peaks_i(d) = \frac{\sum_{d' \in Coverage_i(d)} d'}{\lvert Coverage_i(d) \rvert}$, $Peaks = \{Temp\_Peaks_i(d)\}$ where, $\lvert\cdot\rvert$ represents the elemental number of vector. Then, return to Step 1.1. **Step 1.4:** Return the finite open coverage set *Coverage* and peak point set *Peaks*.
Step 2	Based finite open coverage set and peak point set, perform the peak clustering task: **Step 2.1:** Establishing subsets according to phase angle clockwise: $ClusterSet = \left\{ temp\_set_i \middle\vert temp\_set_i \in Coverage,\ \lvert temp\_set_i \rvert = 1,\ temp\_set_i \neq temp\_set_j, i \neq j \right\}$; **Step 2.2:** Calculate $D_i = min\left\{ Dist(d_j, Peaks) \middle\vert d_j \in temp\_set_i \right\}$; **Step 2.3:** If the condition $\lvert ClusterSet \rvert > 1$ and $max(\lvert D_{GANsi} - D_{GANsi+1} \rvert) \leq k$ are satisfied, combine $temp\_set_i$ and $temp\_set_{i+1}$, return to Step 2.2; otherwise, return *ClusterSet*.
Step 3	Based on information entropy theory, implement the incomplete data restoration task: **Step 3.1:** calculate Euclidean distance as similarity $\{s_j\}_{j=1}^{n_i}$, and normalize the similarity set: $p_j = s_j / \sum_{j=1}^{n_i} s_j$; **Step 3.2:** calculate the entropy value of each complete data sample $h_j = -p_j \ln p_j$; calculate the weight of each complete data sample $w_j = (1-h_j)/(n_i - \sum_{j=1}^{n_i} h_j)$; calculate missing attribute values $f = \sum_{j=1}^{n_i} w_j x_j$, where x_j represents the corresponding attribute values of data samples in the group.
Outputs	Restored dataset $\hat{D} = \{\hat{d}_i\}_{i=1}^{N+M}$.

4. Realization of GANs Based Heterogeneous Data Integration

In order to solve the problem that the IPDU heterogeneous data resources are difficult to be effectively utilized in the small sample environment, a novel generative adversarial networks based heterogeneous data integration technology (GANs-HDI) is proposed in this paper. In the GANs-HDI method, the sample space is expanded by introducing GANs, according to the targeted samples with all of the measurement indexes complete. According to all of the complete and fixed samples, peak clustering and information entropy are employed to restore the incomplete ones. Based on the new sample set expanded by the generative model of GANs, this method constructs a peak clustering model to realize the finite open coverage of the restored sample space, and repair those incomplete samples with entropy function. Finally, all of the repaired samples would be checked by using well-trained discriminator of GANs to guarantee the heterogeneous data integration performances. Generally speaking, the process of GANs based heterogeneous data integration could be presented, as shown in Figure 4.

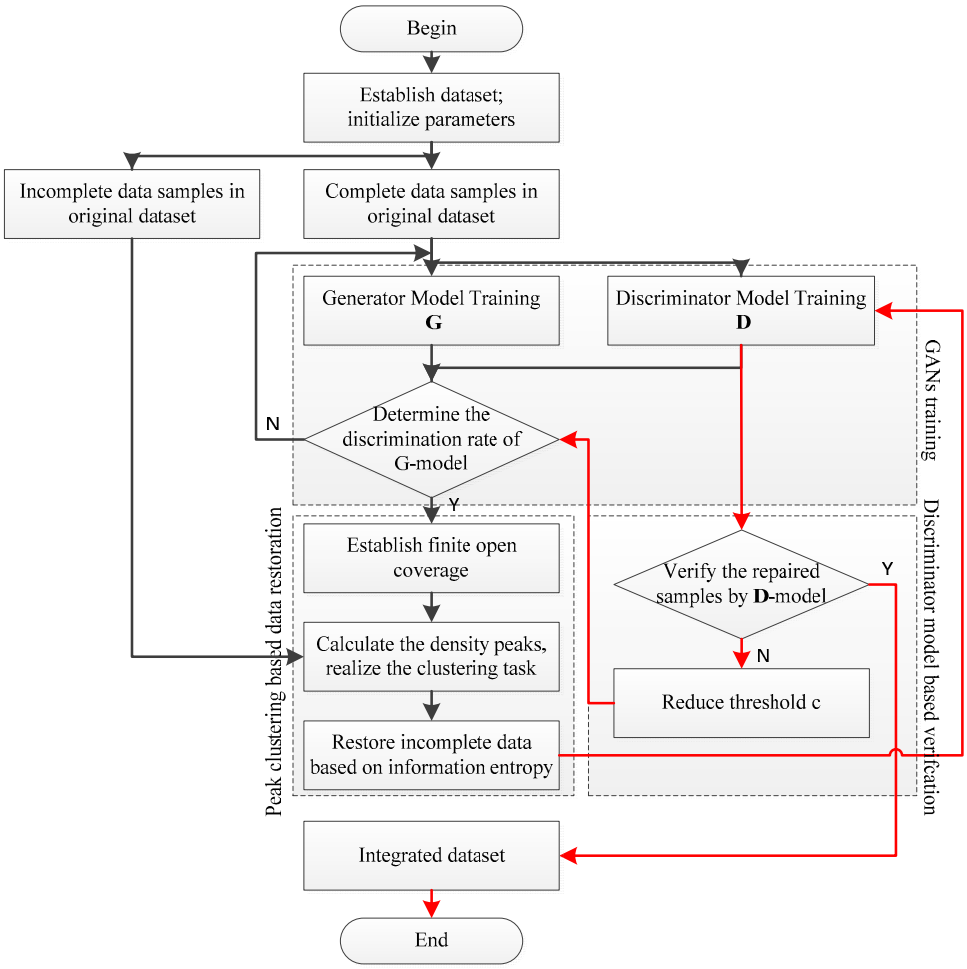


Figure 4. Diagram of GANs based heterogeneous data integration for intelligent power distribution and utilization.

In this paper, select all the data samples with all measurement indexes complete from heterogeneous database $D = \{d_i\}_{i=1}^{N}$, and denote them as a dataset $\overline{D} = \{d_i\}_{i=1}^{\overline{N}}$ for the further generator **G** and discriminator **D**'s training in GANs. Then, taking $\overline{D} = \{d_i\}_{i=1}^{\overline{N}}$ and $\widetilde{D} = \{d_i\}_{i=1}^{M}$ generated by generator **G** as inputs and zeros and ones as outputs respectively, train and renew the discriminator **D**. Finally, if the data samples generated by **G** could meet iterative termination condition,

combine the new generated sample set $\widetilde{D}$ and original dataset $\overline{D}$, and denote the combination as $D_{GANs} = \overline{D} \cup \widetilde{D} = \{d_i\}_{i=1}^{\overline{N}+M}$, perform the peak clustering based data restoration task; otherwise, the discriminant error of discriminator **D** would be back propagated to re-train the generator **G**, as shown in Table 2.

Table 2. GANs Based Heterogeneous Data Integration.

Inputs	Establish the original dataset $D = \{d_i \in R^K\}_{i=1}^{N}$; initialize discrimination rate threshold c, reducing pace α, sample number $\overline{N}$, activation function **g** and **f**, hidden neural node number L, thresholds R_1, R_2, clustering threshold k.
Step 1	Initialize central point set $Centre = D$, and initialize peak point set $Temp\_Peaks = \{\}$.
Step 2	Select all data samples with all measurement indexes complete from heterogeneous database $D = \{d_i\}_{i=1}^{N}$, and denote them as a dataset $\overline{D} = \{d_i\}_{i=1}^{\overline{N}}$. Train generative adversarial networks, and obtain the generator and discriminator. Determine whether the discrimination rate of newly generated samples falls within the interval $[0.5 - c, 0.5 + c]$ by using discriminator **D**. If this condition is satisfied, combine the new generated samples and original dataset, and denote the combination as $D_{GANs} = \{d_i\}_{i=1}^{\overline{N}+M}$, and return to Step 3. Otherwise, the discriminative error of **D** is back propagated to retraining of generator **G**.
Step 3	Employ peak clustering algorithm to repair the samples with incomplete information, and assume the restoration of original dataset $\hat{D} = \{\hat{d}_i\}_{i=1}^{N}$.
Step 4	Determine whether the repaired samples can be verified by discriminator. If it fails, the threshold c would be reduced to $c - \alpha$ based on the reducing pace α, and return to Step 2.3; otherwise, return integrated dataset $\hat{D} = \{\hat{d}_i\}_{i=1}^{N}$.
Outputs	Integrated dataset $\hat{D} = \{\hat{d}_i\}_{i=1}^{N}$.

5. Simulation Experiments and Result Analysis

In this section, the simulation experiments are divided into two parts, i.e., data restoration on University of California Irvine (UCI) standard datasets and heterogeneous data integration on intelligent power distribution and utilization datasets. The former one is performed to verify the validity and stability of our proposed GANs-HDI algorithm, while the latter one is performed to test the actual effect of our proposed GANs-HDI algorithm for intelligent power distribution and utilization heterogeneous data in TensorFlow platform. All of the following simulation experiments were performed in Matlab 2012a and JetBrains PyCharm 2017.2 environment with Core-TM i3-M330@2.13GHz and NVIDIA GeForce 840M processor, respectively.

5.1. Simulation Experiments on UCI Standard Datasets

The simulation experiment introduced three UCI standard datasets, i.e., '*Abalone*', '*Heart Disease*', and '*Bank Marketing*', for performance comparison of data restoration in the Matlab 2012a environment. In this simulation experiment, the incomplete sample proportion in the total samples was set as 20%, and the information loss rate was 25%. Incomplete data sample and missing indexes were randomly selected. The detailed information of the three UCI standard datasets is as shown in Table 3. Taking '*Abalone*' dataset as an example, 60 samples were randomly selected from a total of 4177 data samples as the incomplete samples. In these 60 samples, two indexes were randomly picked out to delete their corresponding information, and formed a data sample set that to be repaired.

Table 3. Detail information of UCI standard datasets.

Datasets	Incomplete Sample Number	Total Sample Number	Missing Dimensional Number	Total Dimensional Number
Abalone	835	4177	2	8
Heart Disease	60	303	19	75
Bank Marketing	9042	45,211	4	17

In order to verify the data restoration performance of the proposed GANs-HDI algorithm on UCI standard datasets, k-nearest neighbors (k-NN), error-back propagation (BP), matrix completion [14],

Deep Learning [18], and proposed Peak Clustering in Section 2 were chosen as control groups with parts of the model parameters selected by experience. Specifically speaking, the cluster number was equal to the sample class number in k-NN algorithm. The numbers of hidden neural nodes were set to be 10/25/12 and the layer number set to be 8 for three UCI standard datasets in BP algorithm with Sigmodal function as the activation function. The layer number set to be 8, and the numbers of hidden neural nodes were set to be [15, 12, 12, 10, 10, 8, 8, 8]/[35, 24, 24, 17, 17, 15, 15, 15]/[20, 18, 18, 15, 15, 12, 12, 12] for three UCI standard datasets in Deep Learning algorithm with Sigmodal function as the activation function. In the proposed Peak Clustering and GANs-HDI algorithms, the threshold of discrimination rate was set to be $c = 0.05$, the reducing pace was set to be $\alpha = 0.0015$, the numbers of hidden neural nodes were set to be $L = 10/25/12$, initialized threshold values as $R_1 = 0.85$ and $R_2 = 0.6$, clustering thresholds as $k = 2$, select Sigmodal function and new generation proportion $\overline{N}/N = 0.5$.

Repair the incomplete data samples with k-NN, Peak Clustering, BP, Matrix Completion [14], Deep Learning [18], and GANs-HDI algorithms, respectively. Then, determine whether the categories of restored samples were correct or not by using support vector machine (SVM), and calculate the accuracy values. Repeat 10 trials independently, and calculate the averages and root mean squared error (RMSE) of the accuracy values of data restoration results, as shown in Table 4 (more details in Table A1).

Table 4. Comparison of four algorithms on UCI datasets.

Algorithm	Index	Dataset		
		Abalone	**Heart Disease**	**Bank Marketing**
k-NN	Accuracy/%	42.87	33.33	60.02
	RMSE	/	/	/
	Time-consuming/s	7.2130	2.1621	50.1315
Peak Clustering	Accuracy/%	58.56	43.33	67.92
	RMSE	0.013379	0.022361	0.018203
	Time-consuming/s	9.1655	2.3097	76.1617
Matrix Completion [14]	Accuracy/%	53.27	40.21	61.23
	RMSE	0.075152	0.065465	0.116122
	Time-consuming/s	17.5660	10.6516	226.6646
BP	Accuracy/%	62.46	50.17	64.26
	RMSE	0.039062	0.135082	0.037608
	Time-consuming/s	17.2132	8.1261	180.2661
Deep Learning [18]	Accuracy/%	71.26	66.02	69.95
	RMSE	0.036261	0.0461610	0.091664
	Time-consuming/s	140.5594	82.6167	362.9500
GANs-HDI	Accuracy/%	94.24	69.17	89.72
	RMSE	0.014235	0.018634	0.001142
	Time-consuming/s	154.1288	89.2661	1374.1626

According to the data shown in Table 4, it is obvious that the time-consuming of both k-NN, matrix completion, Peak Clustering algorithms held the almost same quantity level, while Peak Clustering performed much better than the traditional k-NN algorithm and matrix completion algorithm on the accuracy of data restoration, especially held a more prominent repair effect for linear inseparable data samples. The data restoration performances of BP and Deep Learning [18] algorithm succeeded to beat Peak Clustering algorithm on UCI datasets. However, its RMSE was far from requirement, so the BP algorithm is not stable enough to carry out the engineering application directly. It is worth noting that the GANs-HDI algorithm is far superior to the other control groups in both of the accuracy and RMSE with 20–35 percentage points ahead. However, the algorithm takes a longer time to run, and needs to rely on regularization constraints and distributed computing technologies to improve its convergence efficiency.

In summary, the data restoration performances of GANs-HDI on UCI standard datasets were outstanding when compared with k-NN, BP, matrix completion, Peak Clustering, and deep learning

algorithms. The validity and stability of our proposed GANs-HDI algorithm are verified through the simulation comparison experiments. Furthermore, the experimental results from UCI standard data also showed that the sample number might have a great influence on the final performances of the GANs-HDI algorithm.

5.2. Simulation Experiments on Intelligent Power Distribution and Utilization Dataset (I)

In this section, the simulation experiment took power cable test data of sixty 22 kV XLPE power cable samples for performance comparison of heterogeneous data integration in the JetBrains PyCharm 2017.2 environment. The power cable tests include the accelerated thermal aging tensile fracture test, accelerated thermal extension test, differential scanning calorimetry test, breakdown test, and DC leakage current test. In this simulation experiment, the incomplete sample proportion in the total samples was set to be 20%, and information loss rates were set to be 15%. Incomplete data sample and missing indexes were randomly selected. According to the incomplete sample proportion, 12 samples were randomly selected from total 60 data samples as the incomplete samples. Then, in these chosen samples, two indexes were randomly picked out to delete their corresponding information, and formed a set of data samples to be repaired, according to the information loss rate. Insulating state test indicators of 22 kV XLPE power cable are as shown in Table 5. After the data restoration, this section employed support vector machine (SVM) to predict targeted power cable samples' relative aging times, where 15 samples were treated as the test group, and the other 45 samples were the training ones.

Table 5. Insulating state test indicators of 22 kV XLPE power cable [14].

Index	Sample No.			Index	Sample No.		
	1	2	3		1	2	3
Real operation time/year	4	7	11	Breakdown test pressure level diff. 1	17	17	17
Relative aging time	0.15	0.27	0.52	Breakdown test pressure level diff. 2	14	11	7
Elongation at break (%)	240	225	190	Breakdown test	0.13	0.21	0.38
Load elongation (%)	23.7	47.6	97.1	Insulation resistance per unit length/GΩ	41	29	19
Permanent elongation (%)	2.1	3.50	6.4	Operating ambient temperature/°C	90	90	90
DSC peak temperature/	263	258	243	Operating ambient temperature/°C	90	90	90

In order to verify the data integration performance of the proposed GANs-HDI algorithm, k-nearest neighbors (k-NN), and error-back propagation (BP) were chosen as control groups with parts of the model parameters selected by experience. The cluster number was equal to the sample class number in k-NN algorithm. The number of hidden neural nodes was set to be 10 in BP algorithm with Sigmodal function as the activation function. In the proposed GANs-HDI algorithm, the threshold of discrimination rate was set to be $c = 0.05$, the reducing pace was set to be $\alpha = 0.0005$, the number of hidden neural nodes was set to be $L = 10$, initialized threshold values as $R_1 = 0.85$ and $R_2 = 0.6$, clustering threshold values as $k = 2$, Sigmodal function was chosen as the activation function, and new generation proportion $\overline{N}/N$ was set to be 0.5.

Repair the incomplete data samples with k-NN, BP, and GANs-HDI algorithms, respectively, and calculated the deviation rate with the real values. After the data restoration, employ SVM to perform the relative aging time prediction tasks. Repeat 10 trials independently, and calculate the averages and RMSE of the accuracy values of data restoration results, as shown in Table 6.

Table 6. Performance comparison on heterogeneous datasets for intelligent power distribution and utilization.

Algorithm	Restoration Deviation Rate/%	Life Prediction Deviation Rate/%	RMSE
SVM	/	24.61	/
k-NN + SVM	40.26	58.65	/
BP + SVM	28.58	65.61	0.6216
GANs-HDI + SVM	22.70	86.62	0.2646

According to the data shown in Table 6, life prediction results could be greatly improve by the data restoration of missing information in this case. It also demonstrated that the newly proposed GANs-HDI algorithm can effectively deal with small sample sized life prediction problems, which cannot be handled by the combinations of traditional algorithms, as caused by the disunity on cable test categories of different manufacturers.

5.3. Simulation Experiments on Intelligent Power Distribution and Utilization Dataset (II)

In this section, the simulation experiment took the medium voltage basic data of 171 towns in power quality on-line monitoring system, from 2015 to 2016, for performance comparison of heterogeneous data integration in the JetBrains PyCharm 2017.2 environment. In this simulation experiment, the incomplete sample proportion in the total samples was set to be 20%, and information loss rates were set to be 5%, 15%, 30%, respectively. Incomplete data sample and missing indexes were randomly selected. According to the incomplete sample proportion, parts of samples were randomly selected from total 342 data samples as the incomplete samples. Then, in these chosen samples, parts of indexes were randomly picked out to delete their corresponding information, and formed a set of data samples to be repaired, according to the information loss rate. Normalized data of typical samples is as shown in Table 7 (original data in Table A2).

Table 7. Normalized data of typical samples.

Index	Sample No.				Index	Sample No.			
	1	2	3	4		1	2	3	4
User number of public users	0.0000	0.0002	0.2496	0.5849	Transformer number of public users	0.0000	0.0002	0.2499	0.5876
Transformer capacity of public users	0.0255	0.0182	0.3376	0.5819	User number of specialized users	0.0390	0.1093	0.2618	0.7129
Transformer number of specialized users	0.0340	0.0765	0.1789	0.4801	Transformer capacity of specialized users	0.3074	0.4216	0.5614	0.8277
Total number of transformers	0.0144	0.0305	0.2686	0.6518	Total capacity of transformers	0.1946	0.2515	0.6253	1.0000
Total number of electricity users	0.0106	0.0270	0.2616	0.6335	Total capacity of electricity users	0.1931	0.2495	0.6212	1.0000
Length of power cable line	0.4209	0.3221	0.7628	0.3740	Total length of power line	0.0287	0.0501	0.7030	0.9300
Number of switching equipment	0.0188	0.0258	0.0500	0.5296	Average segment number	0.1187	0.1115	0.4622	0.3431

In order to verify the data integration performance of the proposed GANs-HDI algorithm on IPDU heterogeneous dataset, k-nearest neighbors (k-NN), and error-back propagation (BP) were chosen as control groups with parts of the model parameters selected by experience. The cluster number was equal to the sample class number in k-NN algorithm. The number of hidden neural nodes was set to be 10 in BP algorithm with Sigmodal function as the activation function. In the proposed GANs-HDI algorithm, the threshold of discrimination rate was set to be $c = 0.05$, the reducing pace was set to be $\alpha = 0.0005$, the number of hidden neural nodes was set to be $L = 10$, initialized threshold values as $R_1 = 0.85$ and $R_2 = 0.6$, clustering threshold values as $k = 2$, and Sigmodal function was chosen as the activation function.

Repair the incomplete data samples with k-NN, BP, and GANs-HDI algorithms, respectively, and calculated the deviation rate with the real values. Repeat 10 trials independently, and calculate the averages and RMSE of the accuracy values of data restoration results, as shown in Table 8.

According to the data shown in Table 6, information loss rate and deviation rate shown a significant proportional relationship. Since there is no strong causal link between the indexes in IPDU heterogeneous datasets, the performance of traditional BP algorithm was not satisfactory in the experiments. On the other side, the performance of GANs-HDI algorithm was much better than k-NN and BP on deviation rate with 15 percentage points ahead. Moreover, when the information loss rate took 30%, the deviation rate of k-NN algorithm zoomed up, and the integration results were far away

from the real sample space. However, the new proposed GANs-HDI algorithm holds good resistance to the changes of information loss rates, and showed its outstanding stability.

Table 8. Performance comparison on heterogeneous datasets for intelligent power distribution and utilization.

Algorithm	Information Loss Rate/%	Deviation Rate/%	RMSE
k-NN	5	29.26	/
	15	38.72	/
	30	102.16	/
BP	5	45.56	0.7980
	15	63.22	0.8384
	30	88.36	1.0916
GANs-HDI	5	14.59	0.1460
	15	19.73	0.1975
	30	26.01	0.2603

When considering the influence of sample number on the algorithm performances shown in Section 4, it would be necessary to study on the relationship between data integration performance and parameters in GANs-HDI. In order to further proof the influences of incomplete sample proportion and information loss rate on the heterogeneous data integration performance of GANs-HDI, deviation rates were calculated with different incomplete sample proportions and information loss rates on IPDU heterogeneous datasets, as shown in Figure 5.

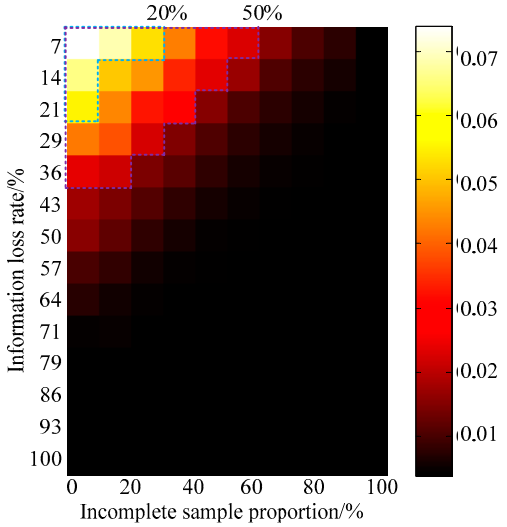


Figure 5. Deviation rate of generative adversarial networks based heterogeneous data integration (GANs-HDI) algorithm on heterogeneous datasets for intelligent power distribution and utilization.

In Figure 5, the color of each color block indicates the reciprocal of the mean of deviation rates from 10 independent repeated experiments with same incomplete sample proportion and information loss rate. The brighter the color, the better the algorithm works. With the decreases of incomplete sample proportion and information loss rate, data integration performance of GANs-HDI algorithm gradually improves. In Figure 4, the boundaries of deviation rate 20% and 50% were marked. It is obvious that, when the incomplete sample proportion is less than 30%, and the information loss rate is less than 20%, the confidence of IPDU heterogeneous data integration is considerably higher. Generally speaking, the larger volume of dataset is, the higher accuracy of data integration would be, and the

confidence level of the results of heterogeneous data integration of distribution network will also show an overall upward trend.

6. Summary

Aiming at the low utilization efficiency problem of heterogeneous data resources for intelligent power distribution and utilization in the small sample environment, this paper proposed a so-called GANs based heterogeneous data integration technology. In this proposed method, the sample space is expanded by introducing GANs theory, according to the targeted samples with all of the measurement indexes complete. Then, a novel peak clustering model is constructed to realize the finite open coverage of the expanded sample space, and repair those incomplete samples. At last, the repaired samples are checked by using well-trained discriminator of GANs. Generally speaking, according to creative establishment the finite open coverage of targeted sample space, this paper succeeded in combining of GANs learning and clustering theory, and provided a novel heterogeneous data integration, which cannot be realized by any individual theory alone.

It is worth noting that, as an important part of this work, generative adversarial network models' convergence has not been perfectly proved in theory by any experts and scholars yet, and its convergence rate still needs further improvement. In the next stage of our team's works, we would like to study on the improved convergence schemes of GANs for vector data samples, and the distributed learning schemes of GANs with heterogeneous hardware.

Author Contributions: Yuanpeng Tan and Xiaojing Bai developed the theory and carried out the experiment. Yuanpeng Tan wrote the manuscript with support from Wei Liu and Jian Su.

Conflicts of Interest: There are no conflicts of interest.

Appendix A

Table A1. Comparison detail information of four algorithms on UCI datasets.

Algorithm	Dataset	Accuracy/%									
		1	2	3	4	5	6	7	8	9	10
k-NN	Abalone	42.87	42.87	42.87	42.87	42.87	42.87	42.87	42.87	42.87	42.87
	Heart Disease	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
	Bank Marketing	60.02	60.02	60.02	60.02	60.02	60.02	60.02	60.02	60.02	60.02
Peak Clustering	Abalone	59.28	60.12	61.08	58.32	61.08	58.20	56.53	58.20	57.60	59.28
	Heart Disease	43.33	39.93	43.33	48.33	43.33	41.67	41.67	41.67	44.88	44.88
	Bank Marketing	69.54	67.63	69.48	70.02	69.48	63.64	67.96	63.64	67.41	69.59
BP	Abalone	60.72	60.00	67.78	63.83	67.78	59.40	63.95	59.40	55.33	68.86
	Heart Disease	28.33	34.98	36.67	65.02	36.67	60.07	63.33	60.07	43.33	65.02
	Bank Marketing	66.47	59.32	67.72	67.49	67.72	60.01	61.81	60.01	67.74	59.80
GANs-HDI	Abalone	92.22	91.26	94.01	95.21	94.01	96.05	95.21	96.05	93.65	94.49
	Heart Disease	68.33	66.67	66.67	71.67	66.67	71.67	68.33	71.67	69.97	68.33
	Bank Marketing	89.69	89.66	89.79	89.77	89.79	89.96	89.59	89.96	89.71	90.65

Appendix B

Table A2. Original data of typical samples.

Index	Unit	Sample No.				Index	Unit	Sample No.			
		1	2	3	4			1	2	3	4
User number of public users	/	66	74	8053	18,782	Transformer number of public users	/	66	74	8062	18869
Transformer capacity of public users	kVA	63,626	45,600	843,967	1,454,859	User number of specialized users	/	359	1006	2409	6559
Transformer number of specialized users	/	510	1147	2683	7202	Transformer capacity of specialized users	kVA	461,078	632,421	842,045	1,241,475
Total number of transformers	/	576	1221	10,745	26,071	Total capacity of transformers	kVA	524,704	678,021	1,686,012	2,696,334
Total number of electricity users	/	425	1080	10,462	25,341	Total capacity of electricity users	kVA	524,704	678,021	1,688,102	2,717,469
Length of power cable line	km	117.01	89.53	212.06	103.98	Total length of power line	km	270.83	473.35	6644.55	8789.97
Number of switching equipment	/	150	206	400	4237	Average segment number	km/per segment	2.48	2.33	9.66	7.17

References

1. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mol. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
2. Song, Y.; Zhou, G.; Zhu, Y. Present status and challenges of big data processing in smart grid. *Power Syst. Technol.* **2013**, *37*, 927–935.
3. Diamantoulakis, P.D.; Kapinas, V.M.; Karagiannidis, G.K. Big data analytics for dynamic energy management in smart grids. *Big Data Res.* **2015**, *2*, 94–101. [CrossRef]
4. Kezunovic, M.; Xie, L.; Grijalva, S. The role of big data in improving power system operation and protection. In Proceedings of the 2013 IREP Symposium Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid (IREP), Rethymno, Greece, 25–30 August 2013; pp. 1–9.
5. Gungor, V.C.; Sahin, D.; Kocak, T.; Ergut, S.; Buccella, C.; Cecati, C.; Hancke, G.P. Smart grid technologies: Communication technologies and standards. *IEEE Trans. Ind. Informs.* **2011**, *7*, 529–539. [CrossRef]
6. Kim, Y.J.; Thottan, M.; Kolesnikov, V.; Lee, W. A secure decentralized data-centric information infrastructure for smart grid. *Commun. Mag.* **2010**, *48*, 58–65. [CrossRef]
7. Fluhr, J.; Ahlert, K.H.; Weinhardt, C. A stochastic model for simulating the availability of electric vehicles for services to the power grid. In Proceedings of the 2010 43rd Hawaii IEEE International Conference on System Sciences (HICSS), Honolulu, HI, USA, 5–8 January 2010; pp. 1–10.
8. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440–442. [CrossRef] [PubMed]
9. Ma, C.; Chen, Y.; Wilkins, D. Ranking of cancer genes in Markov chain model through integration of heterogeneous sources of data. In Proceedings of the 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Belfast, UK, 2–5 November 2014; pp. 248–253.
10. Chen, X.Y.; Kang, C.Q.; Tong, X.; Xia, Q.; Yang, J. Improving the accuracy of bus load forecasting by a two-stage bad data identification method. *IEEE Trans. Power Syst.* **2014**, *29*, 1634–1641. [CrossRef]
11. Dong, X.L.; Srivastava, D. Big data integration. In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE), Brisbane, Australia, 8–12 April 2013; pp. 1245–1248.
12. Lenzerini, M. Data integration: A theoretical perspective. In Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Madison, Wisconsin, 3–5 June 2002; pp. 233–246.
13. Liu, L.; Esmalifalak, M.; Han, Z. Detection of false data injection in power grid exploiting low rank and sparsity. In Proceedings of the 2013 IEEE International Conference on Communications (ICC), Budapest, Hungary, 9–13 June 2013; pp. 4461–4465.
14. Xu, G.; Tan, Y.; Huang, L. Low-rank matrix completion based lifetime evaluation of XLPE power cable. *Trans. China Electrotech. Soc.* **2014**, *29*, 268–276.
15. Mateos, G.; Giannakis, G.B. Robust nonparametric regression via sparsity control with application to load curve data cleansing. *IEEE Trans. Signal Process.* **2012**, *60*, 1571–1584. [CrossRef]
16. Yu, Q.; Miche, Y.; Eirola, E.; Van Heeswijk, M.; Séverin, E.; Lendasse, A. Regularized extreme learning machine for regression with missing data. *Neurocomputing* **2013**, *102*, 45–51. [CrossRef]
17. Li, R.; Zhang, W.; Suk, H.I.; Wang, L.; Li, J.; Shen, D.; Ji, S. Deep learning based imaging data completion for improved brain disease diagnosis. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Nagoya, Japan, 22–26 September 2014; pp. 305–312.
18. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning with neural tensor networks for knowledge base completion. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 926–934.
19. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
21. Springenberg, J.T. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv* **2015**, arXiv:1511.06390.
22. Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *arXiv* **2016**, arXiv:1701.00160.

23. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef] [PubMed]
24. Du, M.; Ding, S.; Xu, X.; Xue, Y. Density peaks clustering using geodesic distances. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 1–15. [CrossRef]