Article

Sorting Center Value Identification of “Internet + Recycling” Based on Transfer Clustering

Cheng Cheng and Xiaoli Luan *

Key Laboratory of Advanced Process Control for Light Industry, Jiangnan University, Wuxi 214122, China
* Correspondence: xlluan@jiangnan.edu.cn

Abstract: As the core link of the “Internet + Recycling” process, the value identification of the sorting center is a great challenge due to its small and imbalanced data set. This paper utilizes transfer fuzzy c-means to improve the value assessment accuracy of the sorting center by transferring the knowledge of customers clustering. To ensure the transfer effect, an inter-class balanced data selection method is proposed to select a balanced and more qualified subset of the source domain. Furthermore, an improved RFM (Recency, Frequency, and Monetary) model, named GFMR (Gap, Frequency, Monetary, and Repeat), has been presented to attain a more reasonable attribute description for sorting centers and consumers. The application in the field of electronic waste recycling shows the effectiveness and advantages of the proposed method.

Keywords: value identification; transfer clustering; inter-class balanced data selection; GFMR model; Internet + Recycling

1. Introduction

In 2015, China’s recycling pattern shifted from “manual recycling” to “Internet + Recycling” [1]. The “Internet + Recycling” process is very important to identify the value of the sorting center [2]. Take electronic waste (e-waste) recycling as an example. The company will offer information to sorting centers and receive a commission [3]. Therefore, the “Internet + Recycling” companies will design a specific strategy based on sorting center costs [4], recycling channels [5], and value dimensions [6]. In this way, the company can improve its business competitiveness and reduce churn at the sorting center. However, the data set of sorting center is small and imbalanced that is called “Absolute Rarity”, [7] which makes it difficult to carry out the task of value assessment. The traditional oversampling methods [8] may not evaluate the accurate value for sorting centers.

Transfer learning is a branch of machine learning that has been shown to help solve problems with small data sets. It has been widely used in image classification [9], signal processing [10], and text classification [11]. An effective model for the target domain can be obtained by leveraging useful relative information from the source domain. However, there is still limited research on the transfer clustering problem. Jiang et al. proposed transfer spectral clustering (TSC), which could transfer knowledge from related clustering task [12]. Wang et al. extended three traditional Gaussian mixture model (GMM) to transfer clustering versions [13]. These methods are more suitable for the clustering problem that has definite boundaries. However, the amount of sorting centers is too limited to obtain boundaries. Fuzzy c-means (FCM) [14] is a clustering algorithm that could solve this problem. The algorithm are more applicable in many fields by changing the objective function of the FCM [15]. Transfer fuzzy c-means (TFCM) [16] is a transfer clustering version of FCM which has good performance on small data set clustering by transferring knowledge from the relative source domain’s cluster centers. There are plentiful customers in “Internet + Recycling” which contain useful and relative information. Thus, in this paper, we adopt TFCM to transfer the cluster centers of customers as knowledge to assist
cluster sorting centers. In order to achieve accurate cluster centers, a comprehensive model that can describe the characteristics of customers is necessary. The RFM model with clustering algorithm has been widely used in customer value identification. Pondel et al. compared three different clustering algorithms’ results of 56,237 customers who made at least 2 purchases in the online store [17]. Kumar et al. practiced the RFM model on 127,037 business customer [18]. Many other scholars also adjust the RFM model according to the features of customers. For example, Li et al. used an improved RFM model with added indicators to classify 4000 customers on the e-commerce platform [19]. Fahed et al. used an enhanced RFM model to classify 42,172 retail customers [20]. However, in the “Internet + Recycling” process, some high-value customers are the individual economy which is rare in customers. The original RFM model could not describe a comprehensive characterization of the high-value customers.

In summary, to get the accurate value identification of sorting centers, TFCM that transferring knowledge from customers is adopted to solve the problem of the small amount of data set, and inter-class balanced data selection (IBDS) is proposed to help solve the problem of imbalanced data set. In order to obtain accurate customer cluster centers for transfer, an improved RFM model GFMR is proposed. The application in “Internet + Recycling” company proves our approach can effectively improve the accuracy of classifying sorting centers. With the accurate value identification of sorting centers, “Internet + Recycling” companies can apply their marketing strategies more precisely, which could improve business competitiveness and reduce sorting center churn.

2. Acquisition of Customer Cluster Centers

Accurate cluster centers of customers is the prerequisite for TFCM. The RFM model is a popular customer value analysis tool widely used to measure customer lifetime value as well as customer value identification and behavioral analysis. The original RFM definition is as follows:

| R | recency of the last trade |
| F | frequency of trades |
| M | monetary value of the trades |

However, the original RFM model could not identify the active customer. The R of RFM is almost the same for a new customer and an active loyal customer. In the “Internet + Recycling” company, the top 20% customers who are active in trade devote over 60% trading volume. Therefore, it is important to divide high-value customers from others.

The high-value customers of “Internet + Recycling” companies are sometimes the individual economy. They will recycle some specific goods and store them. When the price of the goods is relatively high, they will place several orders online. Therefore, their characteristics are short trade gap, high frequency, huge monetary value, and focus on particular goods. In order to strengthen the ability to identify high-value customers, this paper proposed GFMR model as follows.

| G | average trade gap time between two trade |
| F | frequency of trade |
| M | monetary value of all trade |
| R | maximum number of repeat transactions for the same goods |

Especially, G is defined as in Equation (1)

\[
G = \begin{cases} 
\frac{T_s}{T_f - T_l} & \text{if } F = 1 \\
\frac{T_s}{F} & \text{if } F \neq 1
\end{cases}
\]

where \(T_s\) denotes the statistical interval, \(T_f\) denotes the first date the consumer traded during the statistical interval, \(T_l\) denotes the last date the consumer traded during the statistical interval. If a customer has only traded one time during the \(T_s\), this paper
assumes the consumer’s average trade gap is bigger than the statistical interval set \( G = T_s \). Otherwise, calculate the true trade gap.

The GFMR model will reduce the effect of randomness because the four indicators all have small relation to the sampling date. The average gap will separate intensive trade consumers from the others. The frequency and monetary value will identify loyal consumers. The repeat recycling times could identify individual economies. Consequently, GFMR is more suitable for identifying the consumer value of “Internet + Recycling”. Based on GFMR model, k-means algorithm [21] is used to obtain the cluster centers.

Definition of variables:
\[
D_S = \{ x_j^S \} (j = 1, 2, \ldots, M): \text{domain consisting of the standardized GFMR data of } m \text{ customers}
\]
\[
K: \text{the number of clusters}
\]
\[
\tilde{v}_k: \text{the } k^{th} \text{ cluster center of customers}
\]
Steps of acquiring customer cluster centers by k-means algorithm are as followed:

Step1: Randomly generate \( \tilde{v}_k (k = 1, 2, \ldots, K) \) as initial cluster centers.

Step2: Calculate the distance of each \( x_j^S \) to \( \tilde{v}_k \) as \( \| x_j^S - \tilde{v}_k \|_2^2 \) and classify the sample into the cluster corresponding to the minimal distance.

Step3: Calculate the mean value of all samples within each cluster and update the \( \tilde{v}_k \).

Step4: Repeat step2 and step3 until the maximum number of iterations.

The final \( V = \{ \tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_K \} \) is the customer cluster centers.

3. Transfer Clustering for Sorting Centers

The data set of the introduced customer clustering study ranged from 4000 to 127,037, the amount of sorting centers is only 223 which is far from the modeling order of magnitude. In transfer learning, the domain containing a large amount of useful information is often defined as the source domain. The domain we need to learn are defined as the target domain. In this paper, the customers data set is the source domain \( D_S = \{ x_j^S \} (j = 1, 2, \ldots, M) \) and the labeled sorting centers data set is the target domain \( D_T = \{ (x_i^T, L(x_i^T)) \} (i = 1, 2, \ldots, N) \).

Due to the small size of the data set, there are no legible boundaries between each class. FCM could help solve this problem. The objective function of the original FCM is as follows.

\[
\min_{U,V} J_{FCM} = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik}^a \| x_i^T - \tilde{v}_k \|_2^2
\]
\[
st. u_{ik} \in [0,1], \sum_{k=1}^{K} u_{ik} = 1, 0 < \sum_{i=1}^{N} u_{ik} < N
\]

where \( K \) denotes the number of clusters, \( U = [u_{ik}]_{K \times N} \) is the fuzzy/possibilistic partition matrix whose element \( u_{ik} \) denotes the membership of the \( i^{th} \) sample belonging to the \( k^{th} \) class, \( a \) denotes the fuzzy index, \( V = [v_1, v_2, \ldots, v_K]^T \) is the matrix of \( K \) cluster centers whose element \( v_k \) denotes the \( k^{th} \) cluster center of sorting centers.

Transferring knowledge from the \( D_S \) is a must because the \( D_T \) cannot be trained to a satisfactory model on its own. There is a transfer learning version of FCM in which the objective function is defined as followed [16].

\[
\min_{U,V} J_{TFCM} = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik}^a \| x_i^T - \tilde{v}_k \|_2^2 + \lambda_1 \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik}^a \| x_i^T - \tilde{v}_k \|_2^2 + \lambda_2 \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik}^a \| \tilde{v}_k - v_k \|_2^2
\]
\[
st. u_{ik} \in [0,1], \sum_{k=1}^{K} u_{ik} = 1, 0 < \sum_{i=1}^{N} u_{ik} < N
\]

where \( \lambda_1 \) and \( \lambda_2 \) are non-negative balance parameters.

The learning rules based on Equation (3) are as follows:
where subset of source domain. It is easy to obtain:

\[ v_k = \frac{\sum_{i=1}^{N} u_{ik}^t x_i + \lambda_1 \sum_{i=1}^{N} u_{ik}^t v_h}{\sum_{i=1}^{N} u_{ik}^t + \lambda_2 \sum_{i=1}^{N} u_{ik}^t} \]  

(4)

\[ u_{ik} = \left( \frac{1}{\sum_{k=1}^{C} \left[ \|x_i^T - v_k\|^2 + \lambda_1 \|x_i^T - \tilde{v}_k\|^2 + \lambda_2 \|v_k - \tilde{v}_k\|^2 \right]^{1/2}} \right) \]  

(5)

Two distributions. The KL divergence is defined as follows.

\[ D_{KL}(p||q) = \sum_{i=1}^{m} p(x_i) \log \frac{p(x_i)}{q(x_i)} \]

(6)

\[ = \sum_{i=1}^{m} p(x_i) \log p(x_i) - \sum_{i=1}^{m} p(x_i) \log q(x_i) \]

\[ = H(p(q)) - H(p) \]

where \( H(p, q) \) denotes the cross-entropy, \( H(p) \) denotes the information entropy.

In \( D_{KL}(p||q) \), the \( H_p \) is constant. And proved by Gibbs’ inequality that \( H(p, q) \) is bigger than \( H_p \), so \( D_{KL}(p||q) \) is monotonic and \( H(p, q) \) could represent \( D_{KL}(p||q) \). The smaller \( H(p, q) \) means \( q \) is closer to \( p \).

In this paper, assume the distribution of the source domain is \( p \) which has \( m \) samples, and the distribution of the target domain is \( q \) which has \( n \) samples. In this paper, we attempt to find source domain samples that are more similar to the target domain. Turn into the math equation, the smaller \( H(q, p) \) is what we want. So order the source domain by \( H(q, p, i) = -q(x_i) \log p(x_i) \), and select the relative smaller \( s \) samples to compose the subset of source domain. It is easy to obtain:

\[ H'(p, q) = -\sum_{i=1}^{s} q(x_i) \log p(x_i) < H(p, q) \]

(7)

Algorithm 1: TFCM

1. Initialize the number of iterations as \( t = 0 \) and the \( u_{ik} \) randomly; Set the maximum number of iterations \( t_{max} \) and threshold \( \epsilon \); Set the balance parameters \( \lambda_1 \) and \( \lambda_2 \);

2. Update the \( v_k(t) \) using Equation (4)

3. Set \( t = t + 1 \);

4. Update the \( u_{ik}(t) \) using Equation (5)

5. If all \( \|u_{ik}(t) - u_{ik}(t - 1)\| < \epsilon \) or \( t = t_{max} \), then terminate; else go to 2.

There TFCM Algorithm 1 is described below.

The computational complexities of TFCM is \( O(tNK + tC) \) that is the same as FCM.

This method transfers knowledge from the source domain to the target domain through source domain cluster centers. Changing \( \lambda_1 \) and \( \lambda_2 \) could adjust the level of learning from the source domain. As is proved in the article [16], if the source domain has bad knowledge, it will have a negative influence on the clustering performance in the target domain, which is called negative transfer. The original article tried and failed to adopt appropriate parameter values for reducing the effect of the bad source domain. Therefore, in this paper we choose data selection methods rather than adopting parameter values that have been proven to be effective.

In data selection, the key is to find a measurement between the source and target domain. Kullback–Leibler (KL) divergence is often used to measure the difference between two distributions. The KL divergence is defined as follows.

The computational complexities of TFCM is \( O(tNK + tC) \) that is the same as FCM.
which means we can measure source domain samples by $H(q, p, i)$, the smaller the sample is closer to the target domain.

However, due to the imbalanced number of different categories, if not separate the target domain or the source domain and calculate the distribution individually, a class with a larger amount will overwrite the features of a class with a smaller amount. As is shown in Figure 1, when fitting the distribution of G of consumers. The five best-fitting distributions all neglect small samples.

![Fitted Distributions of G of Customers](image)

**Figure 1.** Fitted Distributions of G of Customers

This paper proposes an inter-class balanced data selection method (IBDS). The steps of IBDS are given in Algorithm 2.

**Algorithm 2: IBDS**

**Input:** labeled target domain $D_T$, unlabeled source domain $D_S$

1. Separate $D_T$ by $L(x^T_i)$ for $\{D^k_T\}$ ($k = 1, 2, ..., K$)
2. Calculate the geometric center of $\{D^k_T\}$ for $C_T = \{c^k_T\}$
3. Calculate the distance between $x^S_j$ and $C_T$ for $Dis_j = \{dis^k_j\}$
4. Classify $x^S_j$ by minimum $Dis_j$ to $\{D^k_S\}$
5. Fit distribution of $\{D^k_T\}$ for $\{p^k_T\}$
6. Fit distribution of $\{D^k_S\}$ for $\{Q^k_S\}$
7. In k class calculate the $H(q, p, i)$
8. Order $\{D^k_S\}$ by $H(q, p, i)$ the smaller the better
9. Taking the smallest sample size in each category as $s$
10. Combine the first $s$ samples of each category for $D'_S$

**Output:** $D'_S$

The computational complexities of IBDS is $O(K^2M + bN)$, where $b$ is the number of features of $D_T$. The proposed algorithm will increase the complexity of the algorithm to some extent, but it can significantly improve the overall accuracy of the algorithm.

This paper ordered the source domain samples by calculating $H(q, p, i)$ for each sample in a different category. Then, use the best $s$ samples of each category to get a balanced and more similar subset that provides better cluster center TFCM. The method has proved to be effective.
4. Experimental Results

In this section, the proposed algorithm is evaluated on real-world data set. This paper collected 754,904 e-waste recycling order records of consumers and 19,703 e-waste transporting records of sorting centers from January 2021 to December 2021 of China’s “Internet + Recycling” company. 308,059 consumers and 223 sorting centers were detected. The company has developed four marketing programs targeting high-value, potential-value, stable-value and low-value sorting centers. The goal of this paper is to accurately identify 4 types of sorting center values.

4.1. Data Processing

The consumer is modeled by GFMR. Due to the trade of sorting centers, it is always counted by cars without the specific trade category. So, the model used is the GFM model. The data is normalized by

\[ x' = \frac{x_{\text{max}} - x}{x_{\text{max}} - x_{\text{min}}} \]  

\[ x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]

The smaller \( G \) is better. Thus \( G \) is normalized by Equation (8). \( F, M, \) and \( R \) are the bigger the batter, so they are normalized by Equation (9). \( G', F', M', R' \) are the standardized variables.

4.2. Customer Value Identification

Based on the RFM and GFMR model, results of the k-means cluster algorithm to identify the value of customer are as follows.

As is shown in Table 1, the difference of \( R \) between four clusters is not obvious. Thus, the algorithm may incorrectly classify some high-value users into potential-value user groups, which leads to the high \( M \) of potential-value. Between stable-value and low-value, the difference between these two clusters is mainly the \( R \) which is caused by the random time the new customer begins to use the service. As is shown in Table 2, the difference of \( G \) between four clusters is very obvious, and the added \( R \) vivid segments high-value customers from the others. Low-value customers who trade only once are clearly separated from stable-value users who trade more actively. Thus, the result of the GFMR model is better than the RFM model. The cluster centers are more informative for sorting centers clustering.

| Cluster    | Count | \( R \)    | \( F \)    | \( M \)    | \( G \)    |
|------------|-------|------------|------------|------------|------------|
| High-value | 165   | 119.3455   | 304.3000   | 16,427.7000| 0.9715     |
| Potential-value | 1146 | 124.1798   | 90.7749    | 5467.7400  | 5.5829     |
| Stable-value | 153,094 | 95.2453   | 1.7242     | 96.5492    | 76.1596    |
| Low-value  | 153,654 | 286.7248  | 1.4013     | 76.7217    | 1.0000     |

4.3. Sorting Center Value Identification

However, there is still a large gap between clustering centers of the source domain and target domain, as shown in Tables 3 and 4.
Table 3. True centers of target domain.

| Cluster       | $G'$ | $F'$ | $M'$ |
|----------------|------|------|------|
| High-value     | 0.9967 | 0.7258 | 0.7035 |
| Potential-value| 0.9718 | 0.1771 | 0.3338 |
| Stable-value   | 0.5945 | 0.0073 | 0.0107 |
| Low-value      | 0.1865 | 0.0036 | 0.0059 |

Table 4. Cluster centers of $D_s$.  

| Cluster       | $G'$ | $F'$ | $M'$ | Distance |
|----------------|------|------|------|----------|
| High-value     | 0.9922 | 0.3373 | 0.2806 | 0.5742   |
| Potential-value| 0.9847 | 0.0070 | 0.0067 | 0.3689   |
| Stable-value   | 0.7912 | 0.0030 | 0.0041 | 0.1969   |
| Low-value      | 0.0000 | 0.0000 | 0.0013 | 0.1866   |

This means that there is a great difference between the source domain and the target domain. Therefore, IBDS is practiced to find a more appropriate source domain. The cluster centers of $D'_S$ are shown in Table 5.

Table 5. Cluster centers of $D'_S$.  

| Cluster       | $G'$ | $F'$ | $M'$ | Distance |
|----------------|------|------|------|----------|
| High-value     | 0.9987 | 0.6216 | 0.4156 | 0.3062   |
| Potential-value| 0.9779 | 0.1483 | 0.2847 | 0.0572   |
| Stable-value   | 0.6937 | 0.0021 | 0.0031 | 0.0997   |
| Low-value      | 0.1392 | 0.0016 | 0.0035 | 0.0473   |

As is shown in Table 5, the four clustering centers are all closer to the target domain. IBDS orders samples by $H(q, p, i)$, the smaller, the more similar to the target domain. The variation of distance between the four clustering centers and the true target domain centers with different top ratios of $D'_S$ is shown in Figure 2.

Figure 2. Variation of distance between true centers and clustering centers.

As is shown in Figure 2, when taking 50% of $D'_S$ the distance is minimum. If the ratio is too large, the subset will contain some samples that are not very similar to the target domain. If the ratio is too small, the randomness of the samples will also affect the similarity between the subset and the target domain.

To prove the effect of IBDS, this paper also presents the accuracy of transferring the knowledge of $D_S$ and 50% of $D'_S$ with $l_{max} = 100$ in Tables 6 and 7. A greener background
color in Tables 6 and 7 means higher accuracy, and a redder background color means lower accuracy.

Table 6. Accuracy of transferring top 50% of $D_S'$ by TFCM.

| $\lambda_1$ | 0     | 0.005 | 0.1   | 0.5   | 0.7   | 1     | 1.5   | 10    | 50    | 100   | Average | Max   |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|-------|
| 0           | 30.04%| 31.84%| 25.56%| 83.41%| 77.13%| 90.13%| 3.59% | 2.24% | 2.24% | 2.24% | 34.84%  | 90.13% |
| 0.005       | 29.15%| 24.66%| 25.56%| 83.41%| 77.13%| 89.24%| 3.59% | 2.24% | 2.24% | 2.24% | 32.35%  | 89.24% |
| 0.1         | 69.51%| 69.00%| 2.24% | 78.92%| 75.78%| 2.24% | 3.59% | 2.24% | 2.24% | 2.24% | 30.81%  | 78.92% |
| 0.5         | 82.96%| 82.51%| 94.89%| 52.47%| 2.24% | 2.24% | 3.59% | 2.24% | 2.24% | 2.24% | 35.87%  | 89.69% |
| 0.7         | 86.10%| 86.10%| 91.93%| 28.25%| 48.43%| 2.24% | 2.24% | 2.24% | 2.24% | 2.24% | 36.91%  | 91.93% |
| 1           | 88.79%| 88.79%| 92.83%| 16.14%| 25.56%| 2.24% | 2.24% | 2.24% | 2.24% | 2.24% | 34.84%  | 92.83% |

Table 7. Accuracy of transferring $D_S$ by TFCM.

| $\lambda_1$ | 0     | 0.005 | 0.1   | 0.5   | 0.7   | 1     | 1.5   | 10    | 50    | 100   | Average | Max   |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|-------|
| 0           | 30.04%| 31.84%| 25.56%| 83.41%| 77.13%| 90.13%| 3.59% | 2.24% | 2.24% | 2.24% | 34.84%  | 90.13% |
| 0.005       | 29.15%| 24.66%| 25.56%| 83.41%| 77.13%| 89.24%| 3.59% | 2.24% | 2.24% | 2.24% | 32.35%  | 89.24% |
| 0.1         | 69.51%| 69.00%| 2.24% | 78.92%| 75.78%| 2.24% | 3.59% | 2.24% | 2.24% | 2.24% | 30.81%  | 78.92% |
| 0.5         | 82.96%| 82.51%| 94.89%| 52.47%| 2.24% | 2.24% | 3.59% | 2.24% | 2.24% | 2.24% | 35.87%  | 89.69% |
| 0.7         | 86.10%| 86.10%| 91.93%| 28.25%| 48.43%| 2.24% | 2.24% | 2.24% | 2.24% | 2.24% | 36.91%  | 91.93% |
| 1           | 88.79%| 88.79%| 92.83%| 16.14%| 25.56%| 2.24% | 2.24% | 2.24% | 2.24% | 2.24% | 34.84%  | 92.83% |
| 1.5         | 91.48%| 91.48%| 92.83%| 90.58%| 13.45%| 2.24% | 2.24% | 2.24% | 2.24% | 2.24% | 33.41%  | 92.83% |
| 10          | 96.41%| 96.41%| 96.41%| 95.07%| 95.07%| 94.17%| 2.24% | 2.24% | 2.24% | 2.24% | 34.84%  | 94.17% |
| 50          | 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 2.24% | 2.24% | 2.24% | 2.24% | 33.41%  | 96.41% |
| 100         | 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 96.41%| 33.41%  | 96.41% |
| Average     | 76.73%| 76.37%| 70.94%| 69.19%| 62.51%| 55.92%| 30.90%| 21.08%| 11.75%| 2.24% | 47.76%  | 96.41% |

As is shown in Table 6, compared to 100% of $D_S'$, the accuracy of top 50% of $D_S'$ with the same $\lambda_1 = 10$ and $\lambda_2 = 1$ is 95.07% which is higher than 91.03%. Compared to the original $D_S$, the results of $D_S'$ have higher accuracy in most parameters. $D_S'$ exceed $D_S$ 7.13% of all kinds of parameters. When $\lambda_1$ is bigger than $\lambda_2$, it usually has good accuracy because $v_k$ is affected by the randomness of small data set. This impact can be reduced by enhancing the learning of high quality data in the source domain. Thus, IBDS combined with TFCM effectively clusters small and imbalanced data sets.

To highlight the advantages of our approach, we also compared it with FCM and CSS (clustering with stratified sampling technique) [8]. The detailed results are shown in Figure 3.

Figure 3. Comparison of different clustering approaches.

Obviously, TFCM combined with IBDS could get the result that is most close to real situation. The data set of the sorting center is small and imbalanced. Thus, the accuracy of
FCM is only 60.09%. FCM classifies some high-value sorting centers into the potential-value sorting centers and classifies some potential-value sorting centers into the stable-value sorting center because setting the cluster centers in the dense samples could achieve a lower score of the Equation (2). CSS is an imbalanced data classification algorithm. The accuracy of CSS is 83.41%. CSS is more accurate in identifying low-value sorting centers. But TFCM combined with IBDS outperforms CSS in identifying high-value sorting centers. Because the small data set is easily overfitted by the way of generating samples through oversampling. Transfer learning can effectively improve the accuracy of sorting center value identification and avoid overfitting at the same time.

5. Conclusions

Considering the fact that the data set of sorting centers is small and imbalanced, TFCM combined with IBDS has been proposed to solve the value identification problem. TFCM could transfer knowledge from clustering centers of customers. The IBDS could find a subset that are more similar and balanced than the target domain. In further research, different ratios of the subset exhibits different disparities from the target domain which is caused by randomness and redundant samples. A suitable ratio that balances sample diversity and validity will have better performance. In most of the parameters, IBDS elevates the accuracy which proved the validity of the method. Compared with FCM, the value assessment accuracy of the sorting center elevated from 60.09% to 91.03%. The method in this paper is also less likely to be overfitted compared to the oversampling method.

Further research will focus on automatic adjustment of the ratio to balance sample diversity and validity. In this paper, customers and sorting centers share similar characteristics. However, transferring knowledge from data sets without similar features is still a challenge.

Author Contributions: Conceptualization, C.C. and X.L.; methodology, C.C. and X.L.; software, C.C.; validation, C.C. and X.L.; formal analysis, C.C. and X.L.; investigation, C.C. and X.L.; resources, X.L.; data curation, C.C.; writing—original draft preparation, C.C.; writing—review and editing, X.L.; visualization, C.C.; supervision, X.L.; project administration, X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Key Research and Development Project under Grant 2018YFC1900802.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References
1. Wang, H.; Han, H.; Liu, T.; Tian, X.; Xu, M.; Wu, Y.; Gu, Y.; Liu, Y.; Zuo, T. “Internet+” recyclable resources: A new recycling mode in China. Resour. Conserv. Recycl. 2018, 134, 44–47. [CrossRef]
2. He, K.; Li, L.; Ding, W. Research on recovery logistics network of waste electronic and electrical equipment in China. In Proceedings of the 2008 3rd IEEE Conference on Industrial Electronics and Application, Singapore, 3–5 June 2008; pp. 1797–1802.
3. Liu, T.; Zhang, Q.; Zheng, Z.; Wu, S.; Weng, Z. Stakeholder Analysis of the Waste Electrical and Electronic Equipment Internet Recycling Industry. Int. J. Environ. Res. Public Health 2022, 19, 10003. [CrossRef]
4. Jian, H.; Xu, M.; Zhou, L. Collaborative collection effort strategies based on the “Internet+ recycling” business model. J. Clean. Prod. 2019, 241, 118120. [CrossRef]
5. Qu, Y.; Zhang, Y.; Guo, L.; Cao, Y.; Zhu, P. Decision Strategies for the WEEE Reverse Supply Chain under the “Internet+ Recycling” Model. Comput. Ind. Eng. 2022, 172, 108532. [CrossRef]
6. Cui, Y.; Cao, Y.; Ji, Y.; Chang, I.S.; Wu, J. Determinant factors and business strategy in a sustainable business model: An explorative analysis for the promotion of solid waste recycling technologies. Bus. Strategy Environ. 2022, 31, 2533–2545. [CrossRef]
7. Al-Stouhi, S.; Reddy, C.K. Transfer learning for class imbalance problems with inadequate data. Knowl. Inf. Syst. 2016, 48, 201–228. [CrossRef] [PubMed]
8. Cao, L.; Shen, H. CSS: Handling imbalanced data by improved clustering with stratified sampling. Concurr. Comput. Pract. Exp. 2022, 34, e6071. [CrossRef]
9. Ju, J.; Zheng, H.; Xu, X.; Guo, Z.; Zheng, Z.; Lin, M. Classification of jujube defects in small data sets based on transfer learning. Neural Comput. Appl. 2022, 34, 3385–3398. [CrossRef]
10. George, D.; Shen, H.; Huerta, E. Classification and unsupervised clustering of LIGO data with Deep Transfer Learning. Phys. Rev. D 2018, 97, 101501. [CrossRef]
11. Liu, Z.-G.; Li, X.-Y.; Qiao, L.-M.; Durrani, D.K. A cross-region transfer learning method for classification of community service cases with small datasets. Knowl.-Based Syst. 2020, 193, 105390. [CrossRef]
12. Jiang, W.; Liu, W.; Chung, F.l. Knowledge transfer for spectral clustering. Pattern Recognit. 2018, 81, 484–496. [CrossRef]
13. Wang, R.; Zhou, J.; Jiang, H.; Han, S.; Wang, L.; Wang, D.; Chen, Y. A general transfer learning-based Gaussian mixture model for clustering. Int. J. Fuzzy Syst. 2021, 23, 776–793. [CrossRef]
14. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 1984, 10, 191–203. [CrossRef]
15. Pehlivan, N.Y.; Turksen, I.B. A novel multiplicative fuzzy regression function with a multiplicative fuzzy clustering algorithm. Rom. J. Inf. Sci. Technol. 2021, 24, 79–98.
16. Deng, Z.; Jiang, Y.; Chung, F.L.; Ishibuchi, H.; Choi, K.S.; Wang, S. Transfer prototype-based fuzzy clustering. IEEE Trans. Fuzzy Syst. 2015, 24, 1210–1232. [CrossRef]
17. Pondel, M.; Korczak, J. Collective clustering of marketing data-recommendation system Upsaily. In Proceedings of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Poznan, Poland, 9–12 September 2018; pp. 801–810.
18. Kumar, S.J.; Philip, A.O. Achieving Market Segmentation from B2B Insurance Client Data Using RFM & K-Means Algorithm. In Proceedings of the 2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Thiruvananthapuram, India, 10–12 March 2022; pp. 463–469.
19. Li, X.; Li, C. The research on customer classification of B2C platform based on k-means algorithm. In Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 October 2018; pp. 1871–1874.
20. Yoseph, F.; Heikkila, M. Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. In Proceedings of the 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 3–7 December 2018; pp. 108–116.
21. Borlea, I.D.; Precup, R.E.; Borlea, A.B. Improvement of K-means Cluster Quality by Post Processing Resulted Clusters. Procedia Comput. Sci. 2022, 199, 63–70. [CrossRef]