Research Journal of Applied Sciences, Engineering and Technology 6(8): 1472-1478, 2013
DOI:10.19026/rjaset.6.3974
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 31, 2012 Accepted: January 03, 2013 Published: July 10, 2013

**Research Article**

# The Research on Curling Track Empty Value Fill Algorithm Based on Similar Forecast

Zhao Peiyu and Li Shangbin
Harbin Engineering University, Harbin, China

**Abstract**: The sparsity problem could result in a data-dependent reduction and we couldn't do rough set null value estimates, therefore, we need to deal with the problem of a sparse data set before performing the null value estimate and padded by introducing a collaborative filtering technology used the sparse data processing methods - project-based score prediction in the study. The method in the case of the object attribute data sparse, two objects can be based on their known attributes of computing the similarity between them, so a target object can be predicted based on the similarity between the size of the other objects to the N objects determine a neighbor collection of objects and then treat the predicted target unknown property by neighbors object contains attribute values to predict.

**Keywords:** Artificial Neural Network (ANN), Autonomous Hybrid Power System (AHPS), curling track, Static Var Compensator (SVC)

## INTRODUCTION

Sparsity problem is one of the priorities of the recommended techniques of data tables that contain a large number of null values due to appear in the actual recommendation system often, collaborative filtering technology to achieve firstly is to deal with the sparsity of the data table, otherwise, the e-commerce the system would not be able to type of data for processing. Similarly, information system containing a large number of null values for subsequent data processing has brought great difficulties and cannot generate accurate and effective decision-making rules. Collaborative filtering the processing object technology can be a two-dimensional table of data, the same, object handling in rough set theory is a two-dimensional table, therefore, can score prediction method using collaborative filtering technology the sparse information systems rough set data processing.

Nearest-based collaborative filtering recommendation algorithm needs to measure the similarity between different users and then select the highest number of user's similarity to the current user as the current user's nearest neighbor set, the last collection by the recommendation algorithm based on neighbor ratings recommended produce results (Yu *et al*., 2001; Krasowski, 1988; Zou *et al*., 2001; Li, 2001; Zhang *et al*., 2003). Promotion rough set field to predict the value of an object's empty, you need to measure the similarity between the object and other objects and then select the highest similarity with the object of several objects most collection of its neighbors, then by valuation algorithm according to the neighbor set of attribute values null value of the object. First introduce three conventional similarity measure calculated the similarity between the object the i and object (Li and Shi, 2006; Miao and Hu, 1999; Nejman, 1994) j. Calculation process, first need to calculate the intersection AAA specific collection of objects the I and objects j non-empty properties as a formula:

- **Cosine similarity formula (cosine):** value of the attribute to see as a vector on the two-dimensional data table, if the attribute of the object is unknown, then the attribute value is set to 0, the cosine similarity between objects through vector Folder angle amount. The designed object i and object j property values on a 2-dimensional data table for vector $\bar{\imath}, \bar{\jmath}$, the formula expressed as:

$$sim(i, j) = \cos(\bar{i}, \bar{j}) = \frac{\bar{i}, \bar{j}}{\|\bar{i}\| * \|\bar{j}\|}$$

Molecule is a vector inner product of the two attribute values and the denominator is the product of the two attribute values vector mode.

Vector correlation has similar formula: taking object *i* and objects $A_{ij}$ as an example, the similarity between objects the *i* and objects *j sim* (*i*, *j*) by the Pearson correlation coefficient metric formula *j* set of common attributes known attribute values:

$$sim(i,j) = \frac{\sum_{a \in A_{ij}} (R_{i,a} - \overline{R_i})(R_{j,a} - \overline{R_j})}{\sqrt{\sum_{a \in A_{ij}} (R_{i,a} - \overline{R_i})^2} \sqrt{\sum_{a \in A_{ij}} (R_{j,a} - \overline{R_j})^2}}$$

**Corresponding Author:** Zhao Peiyu, Harbin Engineering University, Harbin, China

The $R_{i,a}$, objects are represented $i$ on the attribute value of the attribute a $\bar{R}_i$ and $\bar{R}_j$ may denote the average property values of object the $i$ and object $j$.

Correction vector correlation similar formula: the cosine similarity measure without considering the different property values of the object-scale problems, modified cosine similarity measure attributes average property value by subtracting the object of improving the above defects, Object $i$ and objects $j$ the similarity between the $sim\ (i, j)$ formulas is expressed as:

$$sim(i,j) = \frac{\sum_{a\in A_{ij}}(R_{i,a}-\overline{R_i})(R_{j,a}-\overline{R_j})}{\sqrt{\sum_{a\in A_i}(R_{i,a}-\overline{R_i})^2}\sqrt{\sum_{a\in A_j}(R_{j,a}-\overline{R_j})^2}}$$

Usually, the object only if there are more non-empty properties similar to the property value to determine the similarity between objects. When the information system is a sparse data set, the collection of non-empty attributes of the two objects shared $A_{ij}$ also smaller, leading unable to determine the degree of similarity of the two objects, the same time, due to the traditional similarity calculation method is only two objects calculated measure lost a lot of useful information on the known properties of the intersection, the set of neighbors of the target object inaccurate. In summary, the three traditional similarity calculation method known attribute distribution does not apply in the case of sparse data collection, it will result in the collection of the neighbors of the target object is inaccurate, resulting in the lowering of the quality of the entire algorithm. Therefore, in the case of address data sparseness, we calculate the similarity between objects, can be set $x$ object known non-empty set of attributes with $A_x$ said, $A_{ij}$ said:

$$A_{ij} = A_i \cup A_j \ \ (i \neq j)$$

Collection based on the properties the $A_{ij}$, objects $i$ and objects $j$ the $A_{ij}$ collection of properties empty value through a collection of objects similar ratings predicted then calculated the object $i$ and objects $j$ similar collection of $A_{ij}$. This method not only can effectively resolve the relevant the less similarity measure objects known attribute data and can effectively solve the same problem all unknown attribute value in the cosine similarity measurement method and the modified cosine similarity measure out the attribute value, so the calculation is more accurate, so as to effectively improve the quality of the null value estimate (Liu *et al*., 2001; Zong *et al*., 1999; Tzung-Peihong *et al*., 2010). How to estimate the object in the collection of properties $A_{ij}$ unknown attribute value is the key. Set object III is a collection of the properties in the property values of the attributes $A_{ij}$ unknown:

$$Ni = Aij - Ai$$

Any property of $a \in N_i$, use the following steps to estimate the object $i$ the $a$ property values:

**Step 1:** Calculate the properties of $a$ similarity between other properties and constitute a collection of attributes compatible

**Step 2:** Attributes compatible set the highest similarity to certain attributes as a set of neighbors of the property, that is, the set of neighbors of the property selected collection of $M_a = \{I_1, I_2, \ldots, I_v\}$ makes $a \notin M_a, M_a \subseteq A_{ij}$ and a attributes $I_1$ with attributes similar $sim\ (a, I_1)$ highest attributes The $sim\ (a, I_2)$ followed $I_2$ and attribute p similarity and so forth.

The third step: get $M_a$, a attributes predictive value estimate object $i$ formula:

$$a = \overline{R_a} + \frac{\sum_{j\in M_a} sim(a,j)*(R_{i,j} - \overline{R_a})}{\sum_{j\in M_a}(|sim(a,j)|)}$$

$\bar{R}_a$ said the average value of properties in the attributes compatible class, $R_{i,j}$ objects $i$ property values on properties $j$.

Collection $A_{ij}$ attributes are not empty; you can use the object similarity formula to get the similarity between objects. Similarly you can get the score predicted value of the formula for:

$$\Pr e_{i,p} = \overline{R_p} + \frac{\sum_{j\in M_p} sim(i,j)*(R_{j,p} - \overline{R_p})}{\sum_{j\in M_p}(|sim(i,j)|)}$$

## FEATURE WEIGHTING BASED ON ENTROPY AND MUTUAL INFORMATION

The prediction of the null value is similar through the generation and the target object for a neighbor set, the use of the collection on the target air-value prediction; the method is based on this model estimation algorithm based on the predicted values of the null value. Therefore, the set of neighbors generated source as the characteristics of the target object collection feature weight method can effectively improve the accuracy of the predicted value of the characteristics of the target object, the core idea is to enhance the "good object" to predict the results of positive impact, while reducing the negative impact of "bad objects" to predict results. As attributes of the data table is a vertical, horizontal two-dimensional data table object, so we can be considered from two aspects of the characteristics of a data set (Nelwamondo and Marwala, 2007).

So that the WWW larger the value of that property a more important. If known a sparse data set, the number of attributes if the probability distribution of attribute values dispersed more and more obvious, based on entropy weight. For example, let the data set S domain $U = \{e_1, \ldots, e_n\}$ attribute set $A = (a_1, \ldots, a_m)$

assumptions attribute a property value $v_a$ probability distribution is divided into two parts, where x represents probability distribution of the number of different, y represents the probability distribution of the number of the same (Butalia *et al*., 2007):

$$\lim_{x \to n} H_a = -\sum \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

Also, because the $H_{a,max} = \log_2 n$, Therefore, $\omega_\alpha = H_a/H_{a,max} = 1$ Similarly $\lim_{y \to n} H_a = -\sum 1 \log_2 1 = 0$, $\omega_\alpha = 0$ When the value of the property probability distribution is more dispersed, $\omega_\alpha$ value the larger the object the attribute preferences largest Similarly when y approaches n, $\omega_\alpha$ close to 0, the smallest object this attribute preferences.

On the other hand, due to the entropy weight departure from the property itself features to consider and not related to the relationship between the object and the object. Thus if the object j the prediction of the target object is very important, can be imparted to the object *j* higher weights, thereby improving the quality of the prediction result. Based on these ideas, the mutual information method to measure the correlation between different objects and apply it to the definition of the fitness function as the feature weights (Zhao *et al*., 2012). Its formula is as follows:

$$\omega_{i,j} = I(V_i;V_j)$$
$$I(V_i;V_j) = H(V_i) + H(V_j) - H(V_i,V_j)$$

$V_i$ $V_j$ are the property values of objects the *i* and object *j*, $H(V_i, V_j)$ is the joint entropy of the two objects (joint entropy) (Yin *et al*., 2011). Calculate the property value of the object is not all there, so only the attribute values of two objects are present in the property.

Based on the above two aspects, we propose a double feature weight method, from the data set, respectively, "horizontal" and "vertical" two considerations to consider the characteristics of the property itself weights, but also consider the degree of association between the object make up the case of the entropy weights law failure in small differences in property values, therefore, our similarity calculation formula (?) the following improvements:

$$sim(i,j) = \frac{\sum_{a \in A_{ij}} \omega_{i,j} \cdot (\omega_a \cdot R_{i,a} - \overline{R_i})(\omega_a \cdot R_{j,a} - \overline{R_j})}{\sqrt{\sum_{a \in A_i} (\omega_a \cdot R_{i,a} - \overline{R_i})^2} \sqrt{\sum_{a \in A_j} (\omega_a \cdot R_{j,a} - \overline{R_j})^2}}$$

Which $\omega_m$ mutual information objects *i* and object *j* weights $\omega_\alpha$ attribute an entropy weights. To $\omega_\alpha$, it will not be too low, limited $\omega_\alpha \in [0.5, 1]$.

**Algorithm Description:** The algorithm dataset sparse standards, according to the required accuracy of the actual data set from the line set.

**Prediction-EM algorithm:**
**Input:** incomplete information system containing a null value.

**Output:** fill the empty value after the incomplete information system.

- Input incomplete information systems S = <U, A, V, f>
- The calculation of the data set sparsity $\tau$
- If $0<\tau\leq k$, according to equation (4), the unknown values $i$ of estimation object $Pre_{i,p}$ turn to step 8, of $k<\tau\leq 1$ go to step 4;
- $i$ compatibility class the $S_B(i)$ (If there is no compatibility class selected followed by the number of null values of N objects as compatible), and entropy weights $\omega_c$ calculated according to the formula (9);
- Calculated according to the formula (10), all the $j \in S_B(i)$ the mutual information weights $\omega_j$ as well as the object $i$ and $j$ $sim(i, j)$ which $i \neq j$;
- $sim(i, j)$ values from largest to smallest, select a number of similar objects constitute a collection of $M_p = \{I_1, I_2, \ldots, I_v\}$
- Is calculated according to equation (8) $Pre_{i,p}$;
- According to the $Pre_{i,p}$ filling empty value, i.e., if the $Pre_{i,p} \neq ø$, otherwise $p(i) \leftarrow *$
- The object i uncertain attribute p' and $p' \in N_i$, go to step 7; otherwise go to step 10
- If S is incomplete information objects *i*', go to step 2; otherwise to step 11
- Outputs a comprehensive information system, the end.

Step 2-7 the calculation of the predictive value of the null values, wherein the step 3 is used in the case of a sufficiently large sparsity domain method, the use of prediction of null values, step 4-7 is the case in the sparse data set the next the null ability forecast based on mutual information entropy weights and objects between attributes weights. Step 8 4-7 to obtain a prediction value according to the step 3 or to a null value of the corresponding prediction fill and then recycled to the next empty prediction of the property. When a null value object does not exist, a complete non-object predicted to fill.

## EXPERIMENT ANALYSIS

In this study, the incomplete information table in Table 1. Contains a range of values for the attributes ($a_1$, $a_2$, $a_3$, $a_4$), setting the threshold of three sparsity of $k = 0.4$, $k = 0.5$, $k = 0.6$, respectively, to fill the empty values in the three sparse case result SIM-EM algorithm compared Meanwhile trained suitable Prediction-EM algorithm the optimal threshold value.

Table 1: Incomplete Data Set

| U | a1 | a2 | a3 | a4 | a5 |
|---|---|---|---|---|---|
| x1 | 3 | 2 | 1 | 0 | 0 |
| x2 | 2 | 3 | 2 | 0 | 0 |
| x3 | 2 | 3 | 2 | 0 | 1 |
| x4 | * | 2 | * | 1 | 0 |
| x5 | * | 2 | * | 1 | 1 |
| x6 | 2 | 3 | 2 | 1 | 1 |
| x7 | 3 | * | * | 3 | 0 |
| x8 | * | 0 | 0 | * | 1 |
| x9 | 3 | 2 | 1 | 3 | 1 |
| x10 | 1 | * | * | * | 0 |
| x11 | * | 2 | * | * | 1 |
| x12 | 3 | 2 | 1 | * | 0 |

Table 2: Incomplete Sparse Data Set

| U | a1 | a2 | a3 | a4 | a5 |
|---|---|---|---|---|---|
| x1 | 3 | * | * | 0 | 0 |
| x2 | * | 3 | * | 0 | * |
| x3 | 2 | * | * | 0 | * |
| x4 | * | 2 | * | 1 | 0 |
| x5 | * | 2 | * | 1 | * |
| x6 | 2 | 3 | 2 | * | 1 |
| x7 | 3 | * | * | 3 | * |
| x8 | * | 0 | 0 | * | 1 |
| x9 | 3 | 2 | * | * | * |
| x10 | 1 | * | * | * | * |
| x11 | * | 2 | * | * | 1 |
| x12 | * | 2 | 1 | * | * |

Table 3: Results of SIM-EMS and Prediction-EM on Sparse Data Set

| | | | Prediction-EM | | |
|---|---|---|---|---|---|
| | The actual value | SIM-EM | 0.4 | 0.5 | 0.6 |
| v (x1, a1) | | | 1 | 1 | 1 |
| v (x1, a4) | 0 | 1 | 1 | 0 | 1 |
| v (x2, a2) | 3 | 3 | 3 | 3 | 3 |
| v (x2, a4) | 0 | 0 | 0 | 1 | 0 |
| v (x2, a5) | 0 | 1 | 1 | 0 | 1 |
| v (x3, a1) | 2 | 2 | 2 | 2 | 2 |
| v (x3, a4) | 0 | 0 | 0 | 0 | 0 |
| v (x3, a5) | 1 | 0 | 0 | 1 | 0 |
| v (x4, a2) | 2 | 1 | 1 | 1 | 1 |
| v (x5, a2) | 2 | 2 | 1 | 2 | 2 |
| v (x5, a5) | 1 | 1 | 1 | 1 | 1 |
| v (x6, a1) | 2 | 1 | 2 | 2 | 1 |
| v (x6, a2) | 3 | 3 | 3 | 3 | 3 |
| v (x6, a3) | 2 | 1 | 1 | 1 | 1 |
| v (x7, a1) | 3 | 3 | 2 | 2 | 3 |
| v (x7, a4) | 3 | 2 | 3 | 2 | 2 |
| v (x7, a5) | 0 | 0 | 0 | 0 | 0 |
| v (x8, a2) | 0 | 1 | 0 | 0 | 1 |
| v (x8, a3) | 0 | 0 | 0 | 0 | 0 |
| v (x9, a2) | 2 | 2 | 2 | 2 | 2 |
| v (x9, a5) | 1 | 1 | 1 | 1 | 1 |
| v (x10, a1) | 1 | 0 | 1 | 0 | 0 |
| v (x10, a5) | 0 | 0 | 0 | 0 | 0 |
| v (x11, a1) | 3 | 1 | 2 | 3 | 1 |
| v (x12, a2) | 2 | 2 | 2 | 2 | 2 |
| v (x12, a3) | 1 | 0 | 0 | 0 | 0 |
| v (x12, a5) | 0 | 0 | 0 | 0 | 0 |

**Case study:** Shown in Table 1 of the data set, the data table, from which to choose each time a non-null value is replaced with null values to estimates of the selected data element, respectively, using SIM-EM and Prediction-EM. In order to verify the valuation of the effectiveness of the algorithm proposed in this study, as shown in Table 1 data set processing, sparse dataset Table 2, SIM-EM and Prediction-EM estimate.

The accuracy of the evaluation criteria and the average absolute error MAE compare the valuation SIM-EM algorithm and improvement (Prediction-EM) algorithm proposed in this study.

Accuracy rate: refers to the correct estimation of the total number of attribute values with respect to the proportion of the total number of attribute values to fill, denoted as C, namely:

$$C = \frac{card(x \mid x \in U, a \in A, a(x) \neq \varnothing \wedge \Pr e(x,a) = a(x))}{card(x \mid x \in U, a \in A, a(x) \neq \varnothing)}$$

Average absolute error of the experimental select MAE (mean absolute error) was to evaluate recommendation algorithm quality standards to measure the accuracy of forecasts by calculating the deviation between the predicted score and the actual user ratings MAE The smaller the value, the recommended the higher the accuracy. MAE intuitively of recommended quality measure of ease of understanding is the most common form of recommended quality assessment methods. Average absolute error MAE formula is calculated as follows:

$$MAE = \frac{\sum_{j=1}^{|CI_i|} |p_{ij} - r_{ij}|}{|CI_i|}$$

wherein, $p_{ij}$ for the prediction object $u_i$ estimates of attributes $i_k$ $r_{ij}$ for dataset object $u_i$ the actual value of the property $i_k$, $|CI_i|$ to estimate the number of attribute values for the object $u_i$.

When the data set as shown in Table 1, the sparsity $\tau = 0.31$ data set for non-sparse data sets, SIM-EM the Prediction-EM algorithm estimates using the null value is identical to the domain estimation method, this when estimation accuracy rate is the same, namely the $C_S = C_p = 66.7\%$.

When the data set as shown in Table 2, the sparsity $\tau = 0.54$, the data sets for sparse data sets in Table 3, respectively, uses the results of the valuation of the SIM-EM the Prediction-EM algorithm.

Experiments were taken of this section, the three different threshold values shown in Table 2 sparsity 0.54 data set for training, therefore to evaluate the analysis of the training results by using the two indicators of the accuracy and the average error MAE observed located given the impact of different threshold values on the experimental results, the finally obtained the optimal threshold for the data set. Table 3 shows that the sparsity threshold 0.4, 0.5, 0.6 when after the end of the data set to fill the accuracy of 68.9, 73.3, 65.6%, respectively. Order to observe the exact rate of change of the entire data set to fill with sparse degree change sparsity as abscissa, ordinate accurate rate, three different accuracy curve $l_{c1}$, $l_{c2}$, $l_{c3}$ in three different threshold values were obtained, as shown in Fig. 1, training thus been the optimal threshold sparse.

Figure 1 shows that, when the sparsity $\tau$ is 0, it indicates a null value of the data set has been completely filled, the exact rate obtained:

K = 0.4 C = 68.9%

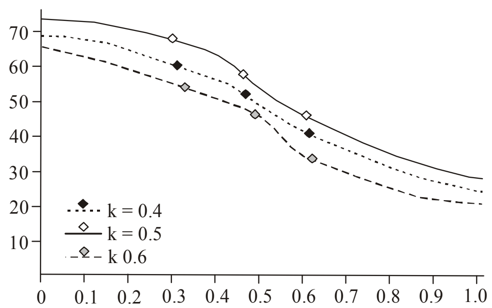


Fig. 1: Comparison of the accuracy under different thresholds

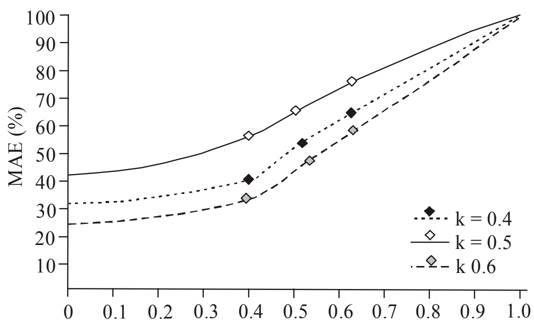


Fig. 2: Comparison of the average error under different thresholds

K = 0.5  C = 73.3%
K = 0.6  C = 65.6%

When sparsity $\tau$>0.6, three threshold values are the data set in this case as the non-sparse data set, so the curve, $l_{c1}$, $l_{c2}$, $l_{c3}$, also consistent; the When sparsity 0.5<$\tau$≤0.6, k = 0.4, k = 0.5 two curves willet this time, the data set as a non-sparse data set, so the curve $l_{c2}$, $l_{c3}$ consistent; Similarly, when the sparsity $\tau$≤0.4, k = 0.4, k = 0.5, K = 0.6, all of the data set in this case as a sparse data set, Therefore curve $l_{c1}$, $l_{c2}$, $l_{c3}$,, the difference increases. Based on the above situation, the curve $l_{c2}$ always maintains a high accuracy and therefore, according to the accuracy metric threshold is about 0.5 for optimal threshold.

Table 3, the sparsity threshold value 0.4, 0.5, 0.6 when data sets to fill after the average error MAE is 31.0, 24.1, 41.4%, respectively. Size in order to compare with the sparsity changes the entire data set to fill the average error sparsity abscissa MAE average error for the longitudinal coordinates were obtained in three different threshold three different MAE curve $l_{MAE1}$, $l_{MAE2}$, $l_{MAE3}$, Fig. 2 and then compare the optimal threshold for the training set.

Shown in Fig. 2, the null value of the data set has been completely filled, get MAE when the sparsity $\tau$ is 0:

K = 0.4  MAE = 31.0%
K = 0.5  MAE = 24.1%
K = 0.6  MAE = 41.4%

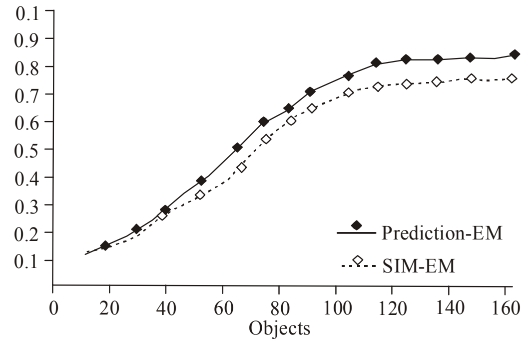


Fig. 3: Comparison of the accuracy using to fill the data set

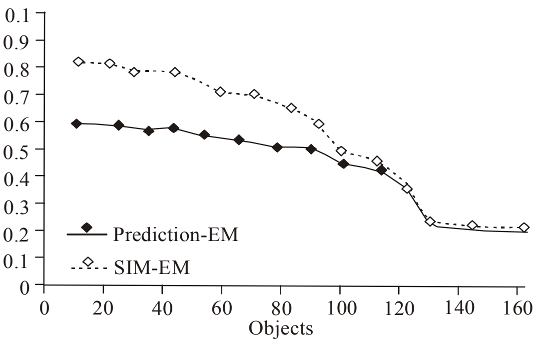


Fig. 4: Comparison of recommended error using to fill the data set

When the sparse degree $\tau$ = 1 when special circumstances that all data sets is empty at this time to fill the average error MAE = 1 the MAE curve $l_{MAE1}$, $l_{MAE2}$, $l_{MAE3}$ consistent trend 1. Other two curves, the curve of the value of $l_{MAE2}$ is lower than in the process with sparse degree increments When k = 0.5, the error of the data set to fill a minimum, therefore, according to the average error metric threshold is about 0.5 for optimal threshold.

K = 0.5-oriented experimental data set by the accuracy of the evaluation of C and the average error MAE get optimal threshold. The next section, we fill the 0.5 threshold Prediction-EM results as the best result of the fill and fill the results were compared with the SIM-EM algorithm.

## COMPARATIVE ANALYSIS

To further analyze the fill effect of the improved algorithm, this study uses the common data provided by the Minnesota State University the GroupLens study group as experimental source data set MovieLens. SIM-EM and Prediction-EM algorithm were used to fill the empty value, the final calculation of the two ways to get the accuracy of the results and the average absolute error. Just as Shown in Fig. 3.

Prediction-EM sparse data set null value prediction to fill, so that the attribute values are dependent on padding value more fully, while using the two heavy right value method corrected similarity calculation results, to improve the prediction accuracy, filling after the end the Prediction-EM algorithm accuracy was 83.6% and SIM-EM algorithm is only 75.1%, so the Prediction-EM increased by 8.5% relative to the SIM-EM algorithm accuracy.

Figure 4 Prediction-EM relative to the SIM-EM data sets recommended filling the effect of a certain improvement in the MAE user rating data are extremely sparse, that users rated project between 0-110, the improved algorithm the accuracy of the prediction score of candidate projects have been greatly enhanced. Although With User Rating intensive, that advantage is waning, but the improved algorithm to solve the sparsity problem of recommender system effectively.

Accurate comparison of the rate of C and the average error MAE evaluation graph can be drawn, in the estimation of sparse data set, the proposed algorithm is more effective and accurate.

## CONCLUSION

This study first introduces a rough set approach non-null values of the incomplete information system, divided into delete tuples, padded, does not deal with three types of methods and missing data highlights null value data on the basis of classification filled the principle of the method; Second, the detailed steps of the algorithm and the advantages and disadvantages of the valuation of the SIM-EM algorithm; once again, the introduction of the collaborative filtering score predicted knowledge, collaborative filtering, sparse data combine to empty value forecast to solve the problem of sparse data, sparsity control the valuation algorithm merit-based, similar weight to ensure the accuracy of the null value predicted improvements null value estimation method based on similarity relations and on this basis. Finally, through the classic data sets and real data sets of the proposed algorithm validation, the contrast of the old and new methods, verify the superiority of the algorithm.

The null value of the predicted value is used in the proposed method of rough set to estimate, in the method on the basis of the original SIM-EM algorithm has been improved. Advantages compared with the original SIM-EM algorithm, due to the introduction of the concept of similarity, the proposed method is not only to maintain the SIM-EM algorithm to solve the sparse data sets is dependent on the data brought too little empty value estimate. Denial or problem cannot be estimated. In addition, this study also proposed the double feature weight method and its integration into the similarity calculation process, final predictive value contains the attributes of the data set characteristics compared with the traditional null value prediction method proposed in this study the predicted value obtained by the method is more consistent with the needs of real data.

## ACKNOWLEDGMENT

The study is supported by the Fundamental Research Funds for the Central Universities (HEUCF101601).

## REFERENCES

Butalia, A., M.L. Dhore and G. Tewani, 2008. Application of rough sets in the field of data mining. Proceeding of the 1st International Conference on Emerging Trends in Engineering and Technology, pp: 298-503.

Krasowski, H., 1988. Aircraft pilot performance evaluation using rough sets. Ph.D. Thesis, Dissertation of Technical University of Rzeszow, Poland.

Li, C., 2001. Where Agent-based knowledge representation metric theory. Comput. Sci., 28(8).

Li, X. and K. Shi, 2006. One kind of attribute reduction algorithm in incomplete information system based on knowledge granularity. Comput. Sci., 2006: 169-199.

Liu, Y.Z. and S.L. Yang, 2001. Null value estimation method based on rough set theory to study. Comput. Eng., 27(10): 41-43.

Miao, D.Q. and G.R. Hu, 1999. A heuristic algorithm of knowledge reduction. Comput. Res. Dev., 36(6): 681-684.

Nejman, D., 1994. A rough set based method of handwritten numerals classification. Institute of Computer Science Reports, Warsaw University of Technology, Warsaw.

Nelwamondo, F.V. and T. Marwala, 2007. Rough set theory for the treatment of incomplete data. Proceeding of the IEEE International Conference on Fuzzy Systems.

Tzung-Peihong, T. Li-Huei and C. Been-Chian, 2010. Mining from incomplete quantitative data by fuzzy rough sets. Expert Syst. Appl., 2010(5): 2644-2653.

Yin, Y., X.D. Zhu and L. Xu, 2011. A media sensitive cache replacement algorithm. IEIT J. Adapt. Dyn. Comput., 1(1): 9-14.

Yu, K., Z. Wen, X. Xu and M. Esters, 2001. Feature weighting and instance selection for collaborative filtering. Proceeding of the 2nd International Workshop on Management of Information on the Web, In Conjunction with the 12th International Conference on DEXA' 2001, Munich, Germany.

Zhang, W., Y. Liang and W. Wu, 2003. Information Systems and Knowledge Discovery. Beijing Science Press, China.

Zhao, Z.L., B. Liu and W. Li, 2012. Image clustering based on extreme k-means algorithm. IEIT J. Adapt. Dyn. Comput., 2012(1): 12-16.

Zong, N., A. Skowron and S. Ohsuga, 1999. New Directions in Rough Sets, Data Mining and Granular Soft Computing. Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, 1711: 73-413.

Zou, X., W. Du, S. Wang and C. Wei, 2001. Fault tolerance relations in the rough set theory and its constructor. Comput. Eng. Appl., 2001(15): 135-137.