A DPC Based Recommendation Algorithm for Internship Positions

Rui Zhang[1(✉)], Lingyun Bi[1], and Tao Du[2,3]

[1] School of Information Engineering, Shandong Management University, No.3500, Dingxiang Road, Changqing District, Jinan 250357, Shandong, China
951114224@qq.com

[2] School of Information Science and Engineering, University of Jinan, Jinan, China

[3] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan, China

Abstract. In the graduation internship activities, the choice of centralized internship based on school-enterprise cooperation units is limited, and the independent internship method will inevitably lead to the problem of blind job selection due to students' lack of social experience. In this paper, we mine historical practice data and propose a recommendation algorithm based on binomial association rules and density peak clustering analysis named DPC-RA. Through binomial association rules, students' internship results in different units are associated with the classification data of unit names, and the internship unit name data is numerically realized. DPC (density peak clustering) algorithm is used to cluster and mine internship data, and obtain multiple clusters. This method can provide accurate personalized work recommendations and improve the quality of internship. Experiments show that this method is effective.

Keywords: Recommendation · Density peak clustering · Digitization

1 Introduction

DPC algorithm relies on the concise data spatial distribution principle to capture the local and global data distribution, so as to obtain excellent clustering division effect [1, 2]. Therefore, this achievement has been published in the Journal of *Sciences* and widely cited. As an important branch of intelligent recommendation algorithm, recommendation system based on clustering algorithm has made a variety of attempts and practices to realize the basic recommendation function in various fields [3]. In the cases where clustering algorithm is applied in recommendation system, most of them use K-means for pattern mining [4–6]. As we all know, K-means algorithm is the representative of partition based clustering algorithm [7, 8]. DBSCAN algorithm is the most classic density based clustering algorithm, which can mine non spherical clusters, but the effect is very sensitive to parameters [10]. The clustering mining effect of non spherical spatial distribution data is not very satisfactory [9]. As an outstanding density clustering algorithm, DPC algorithm can find and mine non spherical and arbitrarily distributed clusters [11].

Y. Wang et al. (Eds.): ICSINC 2022, LNEE 996, pp. 292–299, 2023.
https://doi.org/10.1007/978-981-19-9968-0_35

SNNDPC introduces KNN similarity based on DPC to improve clustering effect, but at the same time, it improves the computational complexity [12]. Therefore, this paper uses DPC algorithm to mine and analyze the historical internship data, and establishes a recommendation model to realize the personalized internship position recommendation function for college students who are about to graduate internship.

2 Related works

In the fields of machine learning, data mining and big data analysis, data is the raw material and the soil for burying knowledge. However, the collected data not only need to complete the cleaning work, but also need to digitize and standardize a large number of classified feature data. Common numerical methods include one hot coding, dummy variable coding, label coding and so on. The first two use different bits of the vector to mark different categories, and the latter is jointly marked by multiple features in the vector, but these methods do not take into account the meaning of the feature itself, which will affect the depth of data mining to a certain extent.

The novel density based clustering method of DPC is used two important parameters δ and ρ to grasp the local and global distribution of the whole current sample space. Parameter ρ represents the local density and obtain local distribution and δ represents the distance from each sample to the nearest point which the density is higher [1]. The mathematical expression is shown in equations (1), (2) and (3):

$$\rho_i = \sum_{j=1}^{N} \chi\left(d_{ij} - d_c\right) \quad i, j \in [1, N] \tag{1}$$

$$\chi = \begin{cases} 1 \ \Delta d \leq 0 \\ 0 \ otherwise \end{cases} \tag{2}$$

$$\delta_i = \min\left(d_{ij}\right) \quad \rho_i < \rho_j \tag{3}$$

where N represents the total number of samples. $\Delta d = d_{ij} - d_c$. Next, drawing the decision graph with two parameters, the abscissa represents density and the ordinate represents distance. The cluster center and outliers can be clearly seen in the graph, this kind called classical decision graph. Taking DPC on UCI data set *spiral* [13] as an example, the classical decision graph is shown on the left of Fig. 1. According to the principle of the algorithm, the point with higher density and farther relative distance is likely to be the center point of high density, that is, the cluster center. In order to make it more central and more prominent, an improved decision graph is proposed, that is, the two parameters are multiplied ($\gamma_i = \rho_i * \delta_i$) to make the gap between samples bigger and make the discrimination greater. γ as ordinate to draw a improved decision graph. The specific figure is on the right of Fig. 1. Obviously, there are three centers in *spiral* .

K-means is commonly used in recommendation system [4]. DPC has obvious advantages to K-means. The effect of the two algorithms on *spiral* are shown in the Fig. 2.

Facing to the irregular and non spherical distributed clusters, it is obvious that DPC algorithm is more practical than the partition based algorithm represented by K-means algorithm. In the application scenario of recommendation system, the mining task of local spatial structure is more suitable for density based clustering algorithm.

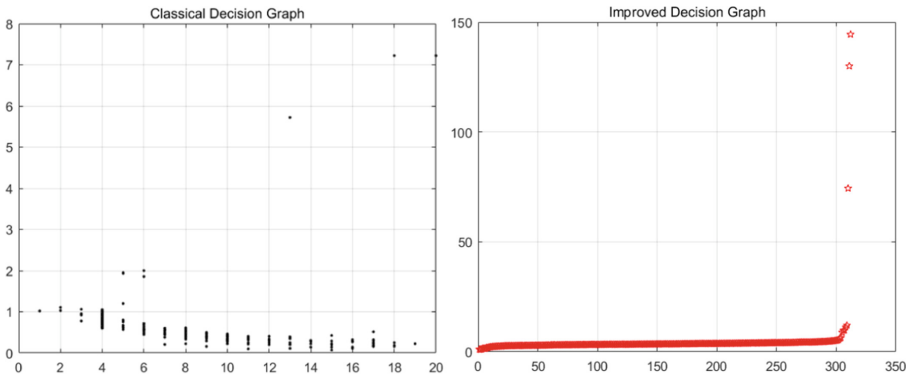

Fig. 1. Decision graph of DPC on spiral

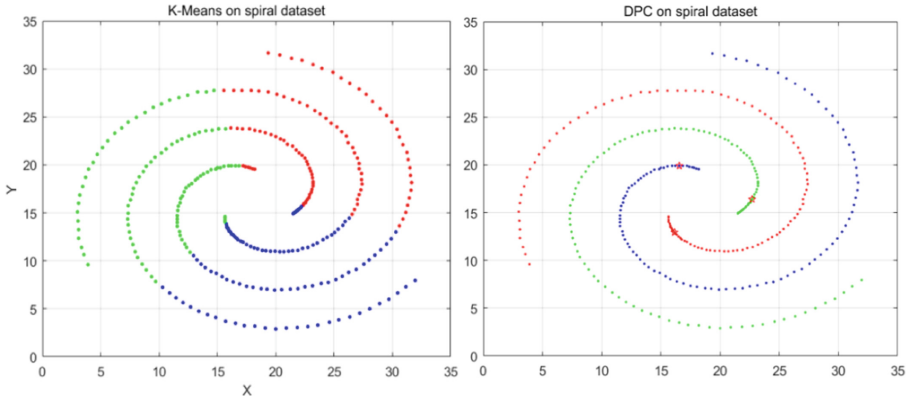

Fig. 2. Comparison of clustering results on spiral

3 Method

The research work of the algorithm is mainly divided into two parts: one is to digitize the classified data, the other is to use DPC algorithm to cluster the standardized data, establish a clustering model and realize the recommendation function. DPC based recommendation algorithm for internship positions function flow is shown in Fig. 3.

The work of this paper mainly includes the digitization of classified data and the implementation of recommendation algorithm based on DPC.

First, classified data digitized. Historical data is composed of students' school performance information, personal basic information and internship information, including student number, name, gender, class, major, courses performance, internship nature, internship performance, internship unit, grade and other dimensions. Internship unit needs to be digitized as sub-type data, and replaced by internship unit code first, measure and calculate the sample distance between internship units through the scores of internship units and students participating in each unit, that is, count the average scores of students participating in internship in each unit, and establish the distance matrix

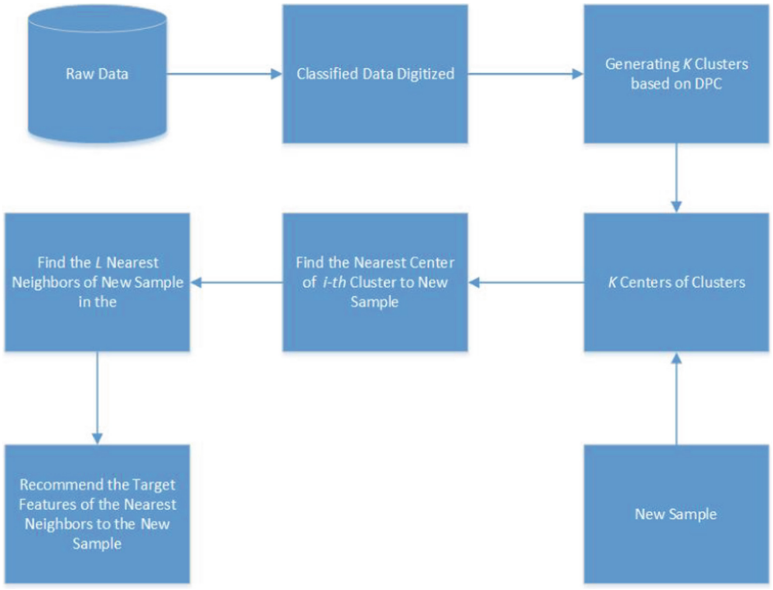

Fig. 3. Test paper generation process

between internship units. The distance calculation formula is as Eq. (4):

$$S\_matrix(i,j) = |Ave\_score_i - Ave\_score_j| = |\frac{Sum\_score_i}{Sum\_stu_i} - \frac{Sum\_score_j}{Sum\_stu_j}| \quad (4)$$

where i and j represent the numbers of two internship units respectively. Sum_score_i represents the total internship scores of all students interning in unit Sum_stu_i represents the number of students interning in unit i, so as to obtain the classification distance matrix(CDM) of internship unit dimension. Matrix CDM is a symmetric matrix with $N \times N$.

Second, the recommendation algorithm based DPC. The recommended function is implemented by DPC algorithm. The specific algorithm principle remains unchanged. The specific algorithm implementation is adjusted according to the creation of CDM matrix. That is, when calculating the Euclidean distance between two samples, it should be discussed separately when it comes to the number dimension of the internship unit. The details are shown in the Table 1 below.

After the Euclidean distance matrix is created, the clusters can be divided according to the original DPC clustering method. Referring to Fig. 4, after several clusters and corresponding high-density centers are obtained by DPC algorithm, the clustering model is obtained. The features of the data used for clustering include target features, such as internship units. For graduates who are about to participate in internship, there is no such data, so they only need to compare other feature data with each cluster center, divide them into the cluster where the nearest cluster center is located, and then find several nearest neighbors in the cluster, which are sorted from large to small according to the

Table 1. Creation of Euclidean distance matrix.

Algorithm 1

Input: Numerical data set $X=\{x_1,x_2,...,x_n\}$, *CDM* Matrix;
Output: Sample Euclidean distance matrix $D_{N\times N}$;
1. Create initial sample distance matrix $D_{N\times N;}$
2. For $i\leftarrow 1$ to N Do
3. For $j\leftarrow 1$ to N Do
4. IF $i != j$
5. For $k\leftarrow 1$ to M Do
6. IF $k == M$
7. $D(i,j)=D(i,j)+S\_matrix(index\_i,index\_j)^2$
8. ELSE
9. $D(i,j)=D(i,j)+(X_{ik},-X_{jk})^2$
10. END IF
11. END FOR
12. $D(i,j)=SQRT(D(i,j));$
13. $D(j,i)=D(i,j);$
14. ELSE
15. $D(i,j)=0;$
16. END IF
17. END FOR
18. END FOR

dimension of internship performance, Recommend the corresponding internship unit again to complete the recommendation function.

4 Experiments

4.1 Datasets

At present, this algorithm is recommended by a single major. The data set comes from the graduation practice data of graduates majoring in Internet of things application technology of our university last year, mainly including 38 dimensions of data, such as student number, name, gender, class, course results, practice unit, practice method and Practice results. The algorithm only cares about gender, the results of various subjects during school, the final practice results and the number of practice units. The gender dimension is a dichotomous attribute. Men and women are replaced by 1 and 0 respectively. In fact, the dimension of internship unit is mainly reflected by the students' comprehensive internship results, that is, it has the strongest correlation with the results, that is, it is numerically calculated with formula 4, and finally the data set available for the algorithm is obtained.

4.2 Algorithm Implementation and Comparison Test

Because the DPC algorithm has a high tolerance rate for the parameters of the truncation distance d_c. In a large range close to the optimal parameters, the impact of d_c on the clustering effect is small. We take one of the parameters to show the experimental effect, that is, $d_c = 1.8$. The corresponding classical decision degree and improved decision diagram are shown in the Fig. 4, It can be clearly seen that the distribution of 4 high-density centers is different from that of other samples.

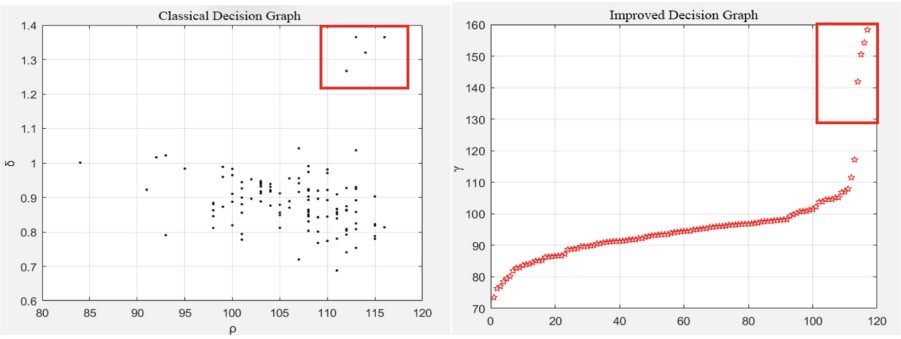

Fig. 4. Decision graph of DPC on *spiral*

It can be seen in Fig. 5 that when the interval of the truncation distance d_c is [1.0, 2.8], the number of suspected cluster centers is unstable at the beginning on the left, and it starts to be stable at 4 when d_c is equal to 1.8. According to the improved decision graph of nearby parameters, when d_c is 1.8, the cluster centers are the most obvious and concentrated, and the degree of discrimination is the highest. While, as can be seen in Fig. 6, when d_c is equal to 1.4 and 2.2 respectively, although the high-density centers can also be separated, they are not very obvious. so the algorithm selects the clustering result when d_c is 1.8 for analysis.

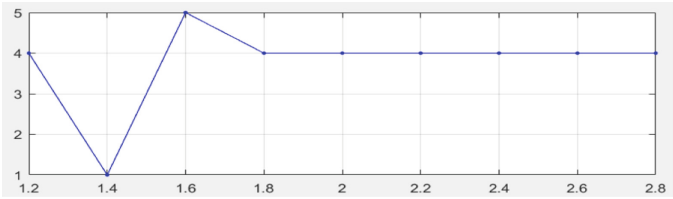

Fig. 5. Relationship between the number of suspected cluster centers and d_c

In order to test the effectiveness of DPC in recommendation effect, it is compared with K-means, DBSCAN and SNNDPC algorithms respectively. The comparative experiment mainly evaluates the clustering effect and the numerical effect of classified data through the tightness (*CP and CT*) index. The indexes can reflect the compactness within

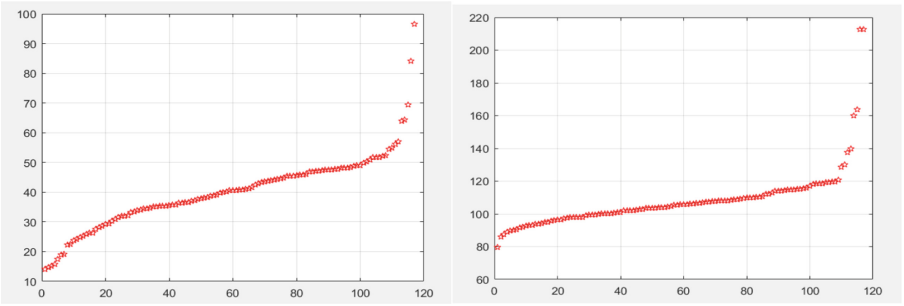

Fig. 6. Improved Decision Graph of DPC with $d_c = 1.4$ and $d_c = 2.2$ respectively

the cluster and the overall compactness, and can be calculated by Eq. (5) and Eq. (6) respectively:

$$CP_i = \frac{\sum_{j=1}^{N_i} \left| X_{ij} - 1NN_i \right|}{N_i} \quad j \in [1, N_i] \tag{5}$$

$$CT = \sum_{j=1}^{N} \left| X_j - 1NN_j \right| \Big/ N \tag{6}$$

where N_i represents the number of samples in cluster i. $1\mathrm{NN_i}$ means the nearest neighbor of X_i and it is in the same cluster with X_i. The smaller the value of CP and CT the higher the similarity in the cluster and it illustrate the better effect of the clustering.

Table 2 shows the CP values of each cluster of the four algorithms under the same conditions of clustering four clusters. The experimental results show that for the CP index of each cluster, the results of DPC algorithm are significantly better than those of K-means, DBSCAN and SNNDPC. *Ave_CP* is the average of *CPs*. Although the result of SNNDPC is the best, there is little difference between DPC and it. The CT value of DPC algorithm is the best compared with the other three algorithms. Although the effect of SNNDPC is not much different from that of DPC, the time complexity is much higher due to SNN. The above also proves that the clustering effect is improved after the binomial statistical digitization of the classified data of internship units through relevant dimensions.

Table 2. Comparison between DPC and other methods.

Algorithm	CP_1	CP_2	CP_3	CP_4	*Ave_CP*	CT
DPC ($d_c = 1$)	**0.8241**	**0.8456**	0.8620	**0.8861**	0.8545	**0.8622**
K-means ($k = 4$)	0.8398	0.8544	0.8740	0.8861	0.8635	0.8657
SNNDPC ($k = 8$)	0.8357	0.8539	**0.8419**	0.8861	**0.8544**	0.8623
DBSCAN (*MinPts = 5*, $d_c = 1$)	0.8398	0.8544	0.8740	0.8861	0.8636	0.8908

5 Conclusion

Because of the unknowness and uncertainty, compared with supervised learning or classification algorithm, unsupervised clustering analysis has a low proportion in practical application, while unsupervised learning can mine interesting knowledge in practical application. In this paper, DPC algorithm is introduced into the recommendation system to realize the recommendation function of general internship post selection. Otherwise, a numerical method based on statistical correlation is proposed for classified data. Retain and improve the amount of information in the original data set, so as to improve the depth and quality of data mining. Practice has proved that this idea is effective and has practical application value.

References

1. Rodriguez, A., Laio, A.: Machine learning. Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
2. Jiang, P., Zeng, Q.: An improved density peak clustering algorithm based on grid. Comput. Appl. Softw. (2019)
3. Song, Z.: Research and Application of Clustering Algorithm in Recommendation System. Nanjing University of Posts and Telecommunications (2018)
4. Jaafar, B.A., Gaata, M.T., Jasim, M.N.: Home appliances recommendation system based on weather information using combined modified k-means and elbow algorithms. Indones. J. Electr. Eng. Comput. Sci. **19**(3), 1635 (2020)
5. Jain, S., Sharma, M., Kumar, P.: Recommendation system for breast cancer treatment using k-means clustering algorithm. Indian J. Public Health Res. Dev. **10**(4), 202–207 (2019)
6. Himel, M.T., et al.: Weight based movie recommendation system using K-means algorithm. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, pp. 1302–1306 (2017)
7. Kanungo, T., et al.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
8. Bai, L., et al.: Fast density clustering strategies based on the k-means algorithm. Pattern Recogn. **71**, 375–386 (2017)
9. Xiao X. Research on selective Clustering Fusion Algorithm Based on fractal dimension. Hefei Polytechnic University (2015)
10. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. International Conference on Knowledge Discovery & Data Mining (1996)
11. Rui, Z., et al.: Adaptive density-based clustering algorithm with shared KNN conflict game. Inf. Sci. **565**(5), 344–369 (2021)
12. Liu, R., Wang, H., Yu, X.: Shared-nearest-neighbor-based clustering by fast search and find of density peaks. Inf. Sci. **450**(1), 200–226 (2018)
13. Chang, H., Yeung, D.-Y.: Robust path-based spectral clustering. Pattern Recogn. **41**(1), 191–203 (2008)