A prediction model for superconductor critical temperature using stepwise discriminant analysis based on feature extraction

P C Guo, Wei Li¹, Z Y Su
School of Science, Dalian Ocean University, Dalian, 116023, China.

¹ Email: wwli@dlou.edu.cn

Abstract. Although the critical temperature is a very important step for the extensive application of superconductors, it’s difficult to find the correlations of many kinds of physical superconductor properties and guarantee the accuracy in predicting the critical temperature. In this paper, an efficient prediction model using stepwise discriminant analysis based on feature extraction is provided to give the reduction of superconductor physical properties and predict the superconductor critical temperature. Firstly principal component analysis and clustering analysis are implemented to reduce the data dimension of superconductor physical properties and give the reductive clusters to complete the feature extraction. The 71 physical properties data of 1300 superconductors is efficiently reduced to 3 main components and 7 clusters. According to the extracted features and improved stepwise discriminant analysis based on the principle of binary search, bayesian discriminant function of each layer is established. At last, Python programming is designed to input the characteristic values and output the predicted the range of critical temperature to finish the efficient computer implementation of this model. Its result that only to the fifth layer, the superconductor critical temperature accuracy of 20,000×71 data matrix is 3. 125 has verified the efficiency of this model.

1. Introduction
In the era of big data, although large data can accurately reflect the overall situation, it is easy to suffer from "dimension disaster". In the existing data, there are more than 70 kinds of physical properties of superconductors and the correlations between many variables can’t be ignored. It increases the difficulties of data analysis. Recently, several properties with the greatest impact on the critical temperature are generally used to approximately replace all of properties according to the previous experience, and then the prediction is completed by multiple linear regression analysis[1]. Firstly, the selection of properties in the above process is not scientific enough, and its correlation is unknown,
which is easy to cause the loss of the information of the original data. Secondly in the multiple linear regression analysis, the model established can’t guarantee the accuracy in predicting the critical temperature and its margin of error can’t be calculated. It is easy to lead to the prediction results inconsistent with the facts. In this paper, an efficient prediction model using stepwise discriminant analysis based on feature extraction is provided to give the reduction of physical properties data of superconductors and predict the superconductor critical temperature to overcome the above limitations.

Superconductor is a conductive material with the properties of diamagnetism and resistance approaching 0 when approaching the critical temperature. Superconductivity was first discovered in 1911. During that time, the scientists found that at extremely low temperature, the resistance of some materials completely disappeared, showing a superconducting state. The temperature at which the resistance of a superconductor is zero is called the critical temperature of a superconductor[2]. Now the superconductor materials have extremely wide applications in the application of strong electricity, weak electricity and diamagnetism. Therefore, it is necessary to study its critical temperature for the more applications of superconductors.

This paper consists of four sections. Section 1 is the introduction, including the research purpose and significance; section 2 is the theoretical introduction of this prediction model including principal component analysis (PCA), cluster analysis and stepwise discriminant analysis; in Section 3, based on the theoretical guidance in section 2, the superconductor critical temperature is efficiently analyzed using this prediction model; Section 4 is the summary and discussion.

2. Methods

An efficient prediction model using stepwise discriminant analysis based on feature extraction is provided to give the reduction of physical properties data of superconductors and predict the superconductor critical temperature. Firstly PCA is implemented to reduce the data dimension of superconductor physical properties. Based on the principal component data, clustering analysis is used to give the reductive clusters to complete the feature extraction. Secondly, according to the extracted features and improved stepwise discriminant analysis, the bayesian discriminant function of each layer is established. Finally, Python programming is used to finish the efficient computer implementation of this model to obtain the predicted value of critical temperature.

2.1. Principal component analysis

PCA is a multivariate statistical method, which reduces the dimension of multivariate variables through linear transformation and selects several comprehensive indexes[4]. The above comprehensive indicators are the principal components after the linear transformation and combination of the original variables. Furthermore, there is no correlation or the correlation among the indicators is extremely low. It not only preserves the information contained in the original indicators, but also simplifies the problems in data analysis. Therefore it greatly reduces the workload and improves the analysis efficiency[5-6]. The algorithm steps are as follows:

(1) standardize the original data, unify its dimensions and order of magnitude, and obtain the standardized matrix X.
(2) use X to calculate the correlation coefficient matrix R.
(3) find the eigenvalue and eigenvector of R matrix by Jacobian algorithm.
(4) when the characteristic value is greater than 1 and the cumulative contribution rate is up to 80-95%, the principal component is determined and the principal component expression is obtained.
(5) establish the initial factor loading matrix and explain the principal components.

2.2. Clustering analysis
The main principle of clustering analysis is to divide data sources into several categories based on their similarity size, so that the objects in the same category can be approximately regarded as one object, greatly improving the efficiency of data analysis[7-8]. In this paper, the system clustering method is used to analyze the main components obtained by PCA. The algorithm steps are as follows:
(1) divide n samples into n categories and form a distance matrix by calculating the distance between each data point.
(2) combine the nearest two classes into one class, leaving n-1 class. Repeat (1) operation to get a new distance matrix.
(3) repeat the operation (2) until the artificially specified number of classes is collected[9].

2.3. Stepwise discriminant analysis
Traditional Discriminant Analysis[3] is a Multivariate Statistical Analysis method to identify the type attribution of a research object based on various eigenvalues under the condition of classification determination. Stepwise discriminant analysis is a method with the bayesian discriminant function as criterion and the ability to variables selection. The main implementation principle is to introduce gradually the variables according to their importance and eliminate the relevant variables according to their correlations until all variables are important. The filtering is completed step by step. In this paper, the principle of binary search in programming[10] is used to improve the analysis method. The algorithm steps are as follows:
(1) determine the grouping variables, divide the samples into two groups according to the grouping criteria and establish the first-layer bayesian discriminant function.
(2) continue to classify one group downward and do the same for the other group to establish the second bayesian discriminant function.
(3) repeat the operation (2) until the desired effect is achieved, and establish the fifth layer bayesian discriminant function, as follows in figure 1:

![Figure 1. Structure of stepwise discriminant analysis.](image)

3. Results
Firstly, PCA is implemented to reduce the data dimension of superconductor physical properties, with SPSS software. Then clustering analysis is used to classify the main components in PCA to complete the feature extraction. Secondly, according to the extracted features and improved stepwise
discriminant analysis, the bayesian discriminant function of each layer is established. Finally, Python programming is used to finish the efficient computer implementation of this model to obtain the predicted value of critical temperature.

3.1 PCA Result
This part is the verification that the physical properties data of superconductors is suitable for PCA. The original data is from paper[1]. It is 71 physical properties of 21, 263 superconductors as Table 1.

| Material                  | La1.85Sr0.15Cu0.982Fe0.018O4 | La1.85Sr0.15Cu0.975Ga0.025O4 | ... | Y0.83Ca0.02Pt0.15Ba2Cu3O6 | Y1Ba2Cu4O8 |
|---------------------------|-----------------------------|-----------------------------|-----|-------------------------|------------|
| Number of elements        | 5                           | 5                           | ... | 6                       | 4          |
| Mean atomic mass          | 72.383174                   | 75.158774                   | ... | 81.12731667             | 76.4445625 |
| Wtd mean atomic mass      | 56.7890145                  | 56.83087779                 | ... | 51.908679               | 49.7159367 |
| Gmean Valence             | 2.352158045                 | 2.352158045                 | ... | 2.40187391              | 2.213363839 |
| Wtd gmean Valence         | 2.228543963                 | 2.229447743                 | ... | 2.069202467             | 2.054799318 |
| Entropy Valence           | 1.589026915                 | 1.589026915                 | ... | 1.748970751             | 1.368922361 |

We select the first 20,000 of the above data as samples and use the remaining data as the comparison samples for the discriminant analysis to obtain the 20,000 × 71 data matrix.

3.1.1 Factor extraction.
Solving factor solution is chosen as the main component analysis method. The basis for determining the number of factors is that the cumulative contribution rate of variance is over 85%, as shown in Table 2 and Figure 2:

| Total Variance | Variance percentage | Accumulation % | Total Variance | Variance percentage | Accumulation % |
|----------------|---------------------|---------------|----------------|---------------------|---------------|
| 1 1303.073     | 98.568              | 98.568        | 1303.073       | 98.568              | 98.568        |
In order to simplify the analysis without affecting the analysis results, 1300 samples of superconductors are randomly selected here.

3.1.2 Scree plot

According to figure 2, the first principal component is significantly distinguished from other eigenvalues. The first three principal component eigenvalues account for 99.819% of all information and the majority of all information. From the fourth principal component, it’s basically flat. Therefore, we obtain 3 principal components to express the vast majority of the information that can be expressed by all the original indicators[1].

3.1.3 Score of each component

Calculating the score matrix of each component with SPSS yields the following results in table 3.

| composition         | 1   | 2    | 3    |
|---------------------|-----|------|------|
| La1. 85Sr0. 15Cu0. 982Fe0. 018O4 | .001| -.005| -.007|
3.2 Clustering analysis of Table 3

Based on the principal component results, with SPSS software, clustering analysis is implemented to further analyze the data. The number of cluster for the data of 3 principal components of 1300 superconductors is 7, as table 4:

| Case | 7 clusters | Case | 7 clusters |
|------|------------|------|------------|
| 1    | 1          | 13   | 4          |
|      |            | 14   | 3          |
| 8    | 2          |      |            |
| 9    | 7          | 68   | 6          |
|      |            | 71   | 1          |

As table 4, based on the three principal components, 71 physical attribute features in the data are successfully classified into 7 categories. Based on the characteristics of the association between various attributes of objects, one item of each category can be selected as the indicator to represent the class, which is respectively: Range_atomic_mass, mean_fie, wtd_mean_fie, wtd_range_fie, wtd_range_atomic_radius, wtd_mean_ThermalConductivity, range_ThermalConductivity. Now the feature extraction operation is completed.

3.3 Stepwise Discriminant Analysis

Based on the seven features extracted by PCA and clustering analysis, the stepwise discriminant analysis is performed. We use the above seven characteristics as independent variables to predict the dependent variable (critical temperature T). The symbols of independent variables are a, b, c, d, e, f, and g, respectively. Firstly, the samples are divided into the analysis samples and the comparison samples for the following analysis. The former is to take the critical temperature as a known data sample to construct the bayesian discriminant function model [11], and the latter is to take the critical temperature as an unknown data sample to test whether the model is accurate. Secondly, to analyze multiple binary classification for the critical temperature T of sample, namely (1) and (2). Because of
the selected data in T within 0 ~ 100 °C, the first layer of the interval length is 50 °C, namely 0 ~ 50 °C for (1) class, 50 ~ 100 °C for (2) class; the second layer is 25 °C, namely 0 ~ 25 °C for (1) class, 25 ~ 50 °C for (1) class, 50 ~ 75 °C for (2) class; 75 ~ 100 °C for (2) class. . . and so on, until the fifth layer interval length by 3. 125 °C, namely 0 ~ 3. 125 °C for (1) class, 3. 125 ~ 6. 25 °C for (2) class. . . 93. 625 ~ 96. 875 °C for (1) class, 96. 875 ~ 100 °C for (2) class. Until this point, the classification of data is completed. The algorithm is as follows:

Table 5. Discriminant analysis data table.

|   |   |   |   |   |
|---|---|---|---|---|
| 607 | 781. 46 | . | 399. 97342 | 0. 2 | 0 | . . | 0 |
| 607 | 745. 38 | . | 399. 97342 | 0. 7 | 0 | . . | 0 |
| 81 | 748. 0666 | . | 399. 97342 | 1. 5 | 0 | . . | 0 |
|   |   |   |   |   |   |   |   |
| 981 | 769. 1 | . | 399. 97342 | 100 | 1 | . . | 1 |

Based on table 5, SPSS is used to construct the bayesian discriminant function layer by layer. First, the typical discriminant function of the first layer is tested:

Table 6. Typical discriminant function test table.

|   |   |   |   |   |
|---|---|---|---|---|
|   |   |   |   |   |
| Werker Lambda |

Function test | Werker Lambda | chi-square | Degrees of freedom | significance |
|---|---|---|---|---|
| 1 | . 673 | 772. 129 | 7 | . 000 |
|   |   |   |   |   |
| As shown in table 6 significance test, the effect of discriminant function is significant. The classification function coefficient of the first layer:

Table 7. Coefficient of classification function

|   |   |   |   |
|---|---|---|---|
| Group1 |
|   |   |   |   |
| 0 | 1 |
|   |   |   |   |
| range_atom_mass | 2. 244 | 2. 277 |
|   |   |   |   |
| mean_fie | . 293 | . 290 |

7
As the tables 5, table 6, table 7, the first Bayesian discriminant function is constructed, namely:

\[
T = 2.244a + 0.293b + 4.373c - 1.468d + 6.871e + 3.391f - 0.378g - 2121.348
\]

\[
T = 2.277a + 0.290b + 4.416c - 1.477d + 6.881e + 3.484f - 0.382g - 2166.998
\]

The Bayesian discriminant function established layer by layer. Repeating the above operation yields the five levels of classification, a total of 62 Bayesian discriminant functions:

1. \( T = 2.720a - 0.464b + 6.340c - 2.607d + 10.015e + 3.213f - 0.899g - 2394.941 \)
2. \( T = 2.710a - 0.469b + 6.331c - 2.606d + 10.020e + 3.203f - 0.896g - 2382.983 \)
3. \( T = 3.883a - 0.059b + 12.264c - 5.069d + 15.878e + 6.065f - 0.0g - 5202.347 \)
4. \( T = 3.899a - 0.075b + 12.229c - 5.049d + 15.855e + 6.084f - 0.0g - 5197.540 \)

Until this point, all the Bayesian discriminant functions are constructed.

### 3.4 Python Programming

Using Python to achieve binary search, the part of the original code is shown as figure 3:

![Implementation Code](image)

**Figure 3. Implementation Code**

The above seven values can be input to obtain the estimated range of the critical temperature. Since the purpose of this paper is to verify the feasibility of this model, only the fifth layer is achieved, that is, the interval length is 3.125. Therefore the accuracy is 3.125. If you keep going down, the length of the interval shrinks exponentially and the accuracy increases correspondingly. The results verifies that the prediction data of this model is more effective and accurate.

### 4. Conclusions
In this paper, an efficient prediction model using stepwise discriminant analysis based on features extraction is provided to give the reduction of superconductor physical properties and predict the superconductor critical temperature. With PCA and cluster analysis, the 71 physical properties data of 1300 superconductors is efficiently reduced to 3 main components and 7 clusters. Based on the above extracted features, a improved stepwise discriminant analysis is provided to find the superconductor critical temperature accuracy. The principle of binary search is used to improve stepwise discriminant analysis to complete the classification in the form of a binary tree and set up the Bayes discriminant function of each layer. Finally, Python programming is designed to input the characteristic values and output the predicted the range of critical temperature to complete the efficient computer implementation of this model. Using this model, the superconductor critical temperature accuracy 20,000×71 data matrix is 3. 125. The results verify that this model is more efficient because it overcomes the limitations of the traditional prediction. The model is simpler and more accurate tool for the prediction of superconductor critical temperature and provides the potential analytical value for the application and promotion of superconductors.

5. Acknowledgments
Natural Science Foundation of Liaoning Province of China(20170540103), National Natural Science Foundation of China(11501076).

6. References
[1] Kam H 2018 Comp. Mater. Sci. 154 346
[2] Zhang L Y 2016 Beyond Freedom: the Magic Superconductor (Beijing: Science Press)
[3] Zhao D F 2007 Mathematical model and computation (Beijing: Science Press)
[4] He X Q 2015 Multivariate Statistical Analysis (Renmin University of China Press)
[5] Su W H 2000 Study on the theory and method of multi-index comprehensive evaluation (Xiamen University)
[6] Guo Y J 2002 Comprehensive evaluation theory and method (Beijing: Science Press)
[7] Du Q, Jia L Y and Yan X F 2014 Spss statistical analysis from introduction to mastery (People's Post Office)
[8] Fang K T and Pan E P 1982 Clustering analysis (Beijing: Geological Publishing House)
[9] Li P Z 2008 The application of cluster analysis and principal component analysis in the evaluation of regional comprehensive consumption level (Nankai University)
[10] Kernighan B W, Pike R , etal 2000 The Practice of Programming (Beijing: Machinery Industry Press)
[11] Yang S Y 2011 Pattern recognition and intelligent computing–matlab technology implementation (Electronic Industry Press)