Study of Machine Learning Based Rice Breeding Decision Support Methods and Technologies

Yun-peng Cui[1](✉), Jian Wang[1], Shi-hong Liu[1], En-ping Liu[2], and Hai-qing Liu[2]

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences/Key Laboratory of Agri-Information Service Technology, Ministry of Agriculture, Beijing, People's Republic of China
{cuiyunpeng, wangjian01, liushihong}@caas.cn

[2] Institute of Scientific and Technical Information, CATS/Key lab of Tropical Crops Information Technology Application Research of Hainan Province, Danzhou, People's Republic of China
yjlep@163.com, haiqing3668720@163.com

Abstract. The Objective of the study is to Analyze and mining rice breeding data with data explore and machine learning algorithms to discover how rice biological characters influence the economic characters, explore effective methods and technologies for breeders and help them find appropriate breeding parents, and provide tools for parental selection in rice breeding. The author developed a B/S application with Python and Django, which implement real-time data mining of rice breeding data. Data analysis and processing result generated from decision tree algorithm can find effective breeding knowledge and patterns, and spectral biclustering algorithm can find required varieties with their local features follow certain patterns. The system can help breeders find useful knowledge and patterns more quickly, and improves the accuracy and efficiency of crop breeding.

Keywords: Machine learning · Rice · Breeding · Decision support

1 Introduction

Rice is one of the most important grain crops in the world, more than half population in the planet regard rice as their staple food [1]. So, the cultivation of new rice varieties is very important for human beings. Nowadays the cross breeding is still the most popular and effective rice breeding method [2], the parental selection is the key to a successful cross breeding. Through the long-term practice, scientists have found the key principles for parental varieties in crop cross breeding: the variety selected must have the target traits needed, no bad traits with the parental varieties, one of the parental varieties

Funding by scientific and technological innovation project of CAAS (CAAS-ASTIP-2016-AII) and Genotype of hot pepper plant in Hainan province – phenotype data depth mining technology research and application project (KJHZ2015-33).

Published by Springer Nature Switzerland AG 2019
D. Li and C. Zhao (Eds.): CCTA 2017, IFIP AICT 545, pp. 54–64, 2019.
https://doi.org/10.1007/978-3-030-06137-1_6

should be proved good variety locally, selected parent variety should have a dominant trait mark, the selected varieties should be able to produce a fruited progeny to avoid hybrid sterility. The five principles are the theoretical basis of parental varieties selection in cross breeding.

There is not much study in parental varieties selection decision support in rice cross breeding. Lifu Jiang uses FFNN and orthogonal genetic algorithm to forecast the quantity and traits of rice hybrids [3]. Dingchun Yan uses a knowledge model according constructed with the natural resources conditions to select the suitable planting location [4]. Yuliang Qi found there is significant positive correlation between use statistical methods for Hybrid F1 yield and effective spikes per plant (and grains per panicle), and put forward choosing principle between subspecies [5]. Since this century, molecular breeding becomes the main breeding method gradually. Marker-assisted selection (MAS) implements the locating of target genes by means of molecular markers for the material of the donor and receptor and predict hybrid parent breeding values of parental varieties through combing phenotype and marker together [6] GWS is developed on the basis of MAS, with zero returned best linear unbiased prediction method (RR-BLU), the accuracy rate is 18%–43% higher than MAS [7]. In actual hybrid rice breeding, parent selection should not depend on the genomics completely, because it will lead to artificial exaggeration of the target traits and affect the judgment for genetic stability. The trend of these years is to combine genomics and phenomics, mining patterns in traits and genes, and find varieties with higher ability and heritability [8], to provide accurate accordance for rice breeding. The study collects rice parental varieties and related phenotypic and economic traits data, mining the data with data exploring and machine learning technology, find how biological traits affect economic traits accurately and discover suitable methods and technology for rapid parental varieties seeking. So provide an operational tool for parental varieties matching. The core problem of the study is how to find the relationships between rice phenotype traits and economic traits, discover useful pattern and knowledge, and provide decision support tool for rice breeding.

2 Data Process and Analysis

2.1 Data Collecting and Processing

The data used in the study comes from rice genetic resources characteristic evaluation database of National Agriculture Science data sharing center, outstanding rice germplasm repository, crop varieties examination, and approval database, rice germplasm database, rice bred varieties and pedigree database, main crop seed vigor monitoring database, main crop variety regional experiment database, approved rice data from national and provincial rice data center. Due to the different data quality and future missing, combine with the data analysis subject and data integration requirements, 19 features is selected during the data preprocessing period, include varieties, available spike, plant height, spike length, total grains amount, filled grains amount, seed rate, thousand seeds weight, grain length, length-width ratio, brown rice percentage, milled rice percentage, head rice rate, chalky percentage, chalkiness degree, gel consistency, amylose content, whole growth period and actual output, and filter the data according to data quality. Finally, 147 male parent data, 133 female parent data and 134 new varieties data are selected.

2.2 Visual Data Exploring

Visual data exploring show data to decision makers in interactive graph to enhance data browse and analyze [9], help them know datasets deeply, form hypothesis, and it's in fact the basis for further data mining and analytics [10].

Because the restriction of human visual sense, usually data exploring only process unitary data, binary data and ternary data. Unitary data analysis is always used for data sample distribution observation, and multivariate data analysis used for discovering mutual effect and dependency. It is hard to visualize the data higher than ternary with conventional charts. They must be transformed into low level data or visualized in other methods. Usual unitary data exploring methods include counting, percentage, variance, standard deviation, average, median, skewness and kurtosis etc., the visualization includes histogram, box chart, pie chart, curve and line chart etc. Usual multivariate data exploring methods include Z test. T test, chi-square test, covariance, regression etc., and visualization include scatter, stacked histogram and also the combination of several charts.

The system developed for this project generate interactive histogram and ridge regression for arbitrary two features and explore the possible correlation relationships among different features.

Ridge regression is a kind of Biased estimate regression analysis method use in collinearity data, it's essentially a kind of improved least square estimation method, the objective function of ridge regression is:

$$\min_{\omega}\|X_{\omega}-y\|_2^2+\alpha\|\omega\|_2^2$$

Ridge regression obtain higher numerical stability through the loss of unbiasedness, and get higher computational accuracy, the algorithm has more practical value in collinearity problem and study with more error data.

The system observes the distribution of every feature with drawing histogram of them, and finds the possible correlation relationships though ridge regression between any two features. As showed in Fig. 1, the two features here have obvious linear correlation relationship.

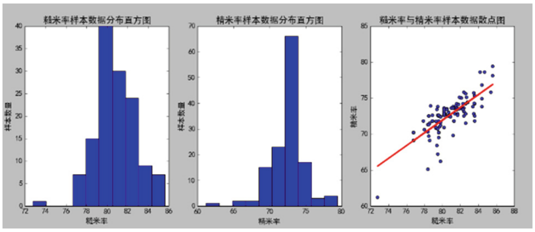

Fig. 1. Binary analysis of rice breeding data

2.3 Mining for Rice Data

The data mining algorithms used for rice data in the study are CART (Classification and Regression Tree) algorithm and Spectral BiCluster.

2.3.1 Decision Tree

Decision Tree also called judging tree. It's a flow chart tree structure with two or more branches [11]. There are numerous implementations for decision tree, like ID3, C4.5, CART, SLIQ and SPRINT etc. CART algorithm is used in this project. CART uses binary recursive partitioning technology, it always divides the sample datasets into two sub datasets, the according to determine the sample partition next time is GINI coefficient (also called GINI impurity): $gini(T) = 1 - \sum p_j^2$. Which pj means the probability of a sample belongs to class j, All the samples belong to a single class when gini (T) = 0, when all the class appear in all the nodes with same probability, gini(T) reach the maximum value, the value is (C−1)C/2. If the GINI coefficient of all the features is calculated, the GINI information gain can be calculated. In this study, features' data is input into CART, and the common sense, rules and knowledge for rice breeding varieties selection and judgement can be found.

2.3.2 Spectral BiCluster

Spectral BiCluster is a cluster method through eigenvalue decomposition. The algorithm is closely related with graphic partitioning. The algorithm can cluster samples on the arbitrary shape of the sample space, and always converge to the global optimal solution.

Figure 2 shows the difference between K-means and spectral cluster. K-means select clustering center randomly first, and gather sample points to the nearest center, then calculate the mean of the points as a new clustering center, repeat the process until the clustering center converge to a stable point. The spectral clustering first calculate the similarity matrix W though Gaussian similarity function: $W_{ij} = e^{\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}}$, then calculatenixlacian) trix L $(L = D - W)$, D means degree matrix, which is a diagonal matrix, eacrepresented in the matrix represent the sum of every single row in matrix W. The algorithm calculates extract minimized k eigenvectors and combine the k vectors into a n × k matrix, each row in the matrix is A vector in the k dimension space. The

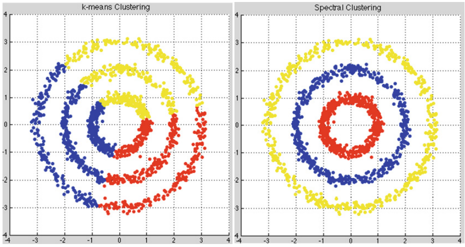

Fig. 2. Comparison between spectral cluster and k-means cluster (Source: Scikit-Learn official Site)

algorithm then cluster the vector with K-means, The class of each row in a cluster belongs is the node of the original graph, which is the class which the original data point belongs.

Traditional cluster just cluster data in one direction (row or column) in a matrix, that is, cluster in global pattern, so K-means just find the global information of the dataset and drop much partition information [12]. While BiCluster algorithm clustering at both row and column direction of the matrix, it not only cluster global information but also find effective partition information in high dimension data [13–17].

Spectral cluster is integration of spectral cluster and bicluster. The algorithm suppose the input data matrix has a hidden board structure, the rows and columns of The matrix can be divided into one or more sub biclusters, to make the count of the cartesian product of row clusters and column clusters in any sub bicluster is roughly constant. For instance, there is a 2 × 3 sub bicluster, then each row in the matrix belongs to three subclusters, and each column belongs to two subclusters. The algorithm re-divide the rows and columns of the matrix into subclusters, to make board matrix corresponding to the subclusters approximate the original matrix better.

The spectral BiCluster algorithm is used in this study to analyze rice breeding data to find approximate parent varieties with partition excellent traits.

2.4 Training Strategy of the Algorithm

Because the samples we collected in this study are limited, in order to get better training result, this study adopts ten-folder cross validation for data training. During each data training, The data is divided into 10 even parts randomly, each single part of the data is taken out in turn, and other 9 parts of the data join the training, the part taken out used to calculate the error rate. Repeat the process 10 times, and each time uses different datasets, finally calculate a comprehensive error rate. The experience proved that 10 times are the best choice for getting more accurate error rate [18].

3 Result

For the convenience of data processing and data visualization, a system is developed by Python with Django framework, to realize the online rice breeding data analysis and mining.

3.1 Result of CART

The maximum amounts of the layers of the decision tree generated should be input before running the CART algorithm. Figures 3, 4 and 5 show the visual results in different layer parameter. A higher or lower layer parameter could result in overfitting or under fitting, so users should adjust the parameter and watch the result to get the best output. The default value of the parameter of the system is 5, which can get reasonable output in most situations.

X[N] (N = 1, 2, …, 18) Means The Features of the Data, X[1] Means the Feature Effective Panicle, X[2] Plant Height, X[3] Ear Length, X[4] Total Grain Number, X[5]

Grain Number, X[6] Seeding Rate, X[7] the 1000 Kernel Weight, X[8] Grain Length, X[9] Length-Width Ratio, X[10] Brown Rice Percentage, X[11] Head Rice Rate, X[12] Whole Head Rice Percentage, X[13] Chalky Rice Rate, X[14] Chalkiness Degree, X [15] Gel Consistence, X[16] Amylose Content, X[17] Whole Growth Period, X[18] Actual Yield

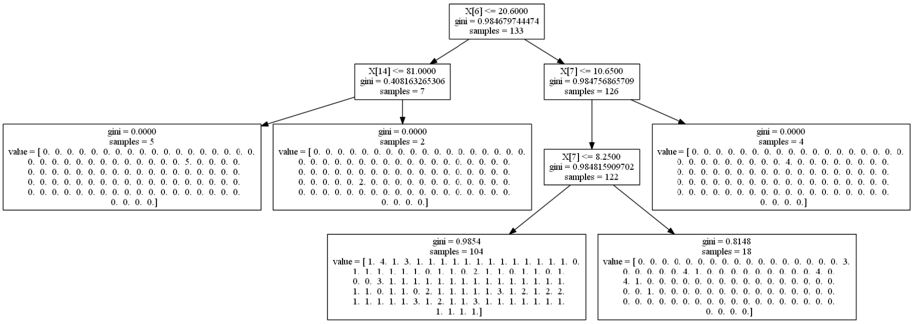

Fig. 3. The 3 layers decision tree generated by CART algorithm processing the rice breeding data

For example, the following rules or common sense can be recognized from the decision tree shows in Fig. 4: there are 100 samples for setting rate > 20.6 and 4.85 $\leq$ thousand seed weight $\leq$ 7.9:10 samples for setting rate $\leq$ 20.6 and thousand seed weight $\leq$ 8.25 and chalk rice grade > 70 and polished rice rate $\leq$ 62.4%. Other samples are scattered across different rules. But the corresponding sample number is small. Changing the number of layers of the decision tree may result in different results. For example, set the maximum layer to 3, the result changes as follow: 104 samples for setting rate $\leq$ 20.6 and thousand seed weight $\leq$ 8.25, 18 samples for setting rate 20.6 and thousand seed weight > 8.25. Therefore, along each path from the root node to leaf node of the tree, there is a rule or a common sense, means a knowledge or a pattern, effective knowledge and patterns can provide valuable references for varieties breeding decision making.

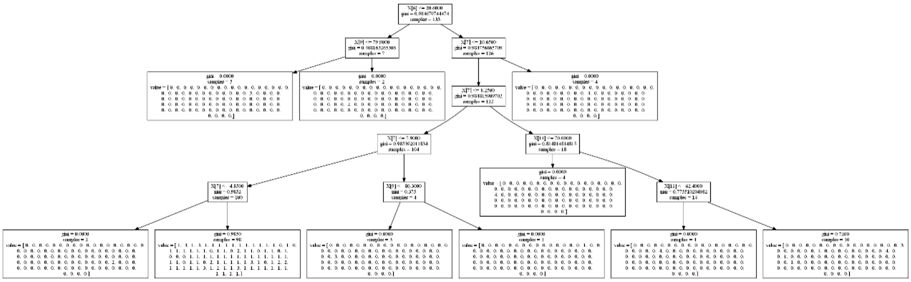

Fig. 4. The 5 layers decision tree generated by CART algorithm processing the rice breeding data

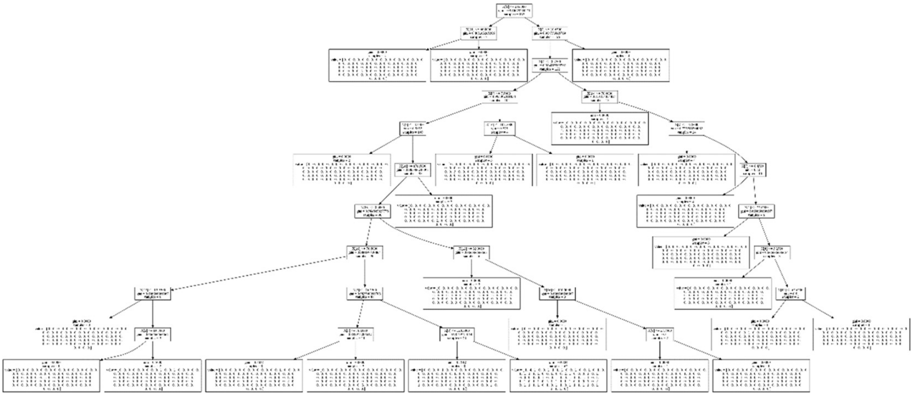

Fig. 5. The 10 layers decision tree generated by CART algorithm processing the rice breeding data

3.2 Result of Spectral Bicluster

The number of rows and columns generated by the spectral Bicluster algorithm must be input previously before running the algorithm (Fig. 6).

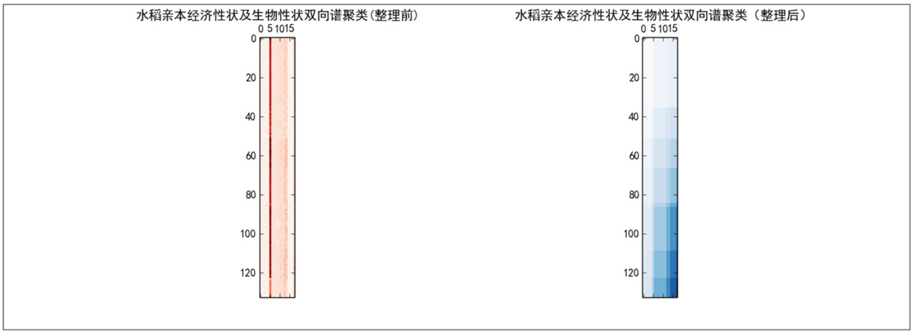

Fig. 6. Visual output of spectral BiCluster analysis

The number on top of the shape represents the feature number. The number to the left of the shape represents the sample number.

Tables 1, 2 and 3 show 3 biclusters generated by the algorithm with parameters rows = 8 and columns = 6. There are 48 biclusters generated in this training, but some of them are meaningless (e.g. Some biclusters only have one column include), we can get different biclusters through changing the value of rows and columns parameters. The value of the biclusters depends on the evaluation of breeders. But the algorithm provides a tool for trying different parameters and training the data anytime, until the breeder gets satisfactory results.

Table 1. Analysis output by spectral BiCluster – data in BiCluster (1)

Time	Effective spike	Length of spike	Thousand seed weight	Chalk rate	Amylose content
Yuefeng A	19.1	20.0	22.1	5.0	14.3
Peiai 64	21.3	22.2	23.1	6.0	16.1
1318	20.6	19.8	24.0	10.0	15.8
Nongken 58S	22.6	23.5	22.4	12.0	14.7
Zhending28A	23.5	25.0	30.0	8.0	17.3
Yanjing187	28.0	21.1	24.5	9.0	17.6
Jiaao935	18.9	18.7	22.5	14.0	17.2
Jiahe212A	23.1	21.4	24.8	2.0	15.0
Qianjiang1A	23.6	20.6	20.0	18.0	16.6
Chuanlu389A	23.7	25.5	26.5	12.0	17.6
YuetaiA	22.2	19.9	24.5	11.0	16.4
Guang8A	23.7	22.1	23.9	8.0	15.7
Zhenshan97A	22.6	19.7	20.0	9.0	16.2
Kongyu110	20.5	21.3	22.0	10.0	15.8
GuiyangaiC17	21.6	22.7	21.4	16.0	16.9
Zhu1S	21.6	20.8	24.3	16.0	15.8
RongfengA	19.8	22.4	21.8	14.0	16.1
WufengA	22.8	19.7	25.8	13.0	14.9
Zhong9A	25.6	22.1	24.0	8.0	14.8

From the data in these biclusters, we can see that the samples in each bicluster are generally closer (the distance between samples). In addition, features extracted from each bicluster follow certain patterns. So results from spectral biclustering algorithm are some closer samples with their features follow certain patterns.

The algorithm used in rice breeding can help breeders find closer samples and phenotype features from large amount of samples rapidly, and help breeders to make the right decision.

4 Discussion

The study explore rice breeding data with unitary and binary analysis method, help breeders discover the influence between biological traits and economic traits, reveal hidden patterns, knowledge or rules in the data, discover closer samples with their features follow certain patterns through Spectral Biclustering algorithm, all these technology can promote the accuracy and efficiency of rice breeding.

The essence of mining the rice breeding data is to reveal the pattern and knowledge hide in the data, discover the influence of how phenotype traits affect the economic traits, and provide reference for rice varieties selection and breeding decision making.

Table 2. Analysis output by spectral BiCluster – data in BiCluster (2)

Name	Amylose content	Length of spike	Thousand seed weight	Chalk rate	Amylose content
Zhenshan97A	23.3	23.6	26.0	4.0	15.2
R432	19.8	21.8	25.0	8.0	16.3
Aijiaonante	21.2	22.4	26.5	8.0	16.2
Zhe04A	19.2	20.6	25.0	12.0	15.0
ChuanXiang28A	25.4	19.3	24.8	13.0	20.8
SimiaoA	20.5	18.6	21.2	10.5	22.7
HD9802S	20.5	19.7	24.2	6.0	15.2
Xiangfeng70A	19.8	23.0	26.0	43.0	14.5
Xiangling750S	22.4	24.0	26.0	5.0	11.8
KexiangA	11.6	22.5	27.0	40.0	17.0
XinA	21.8	18.9	25.0	4.0	15.2
Ewan11	20.6	14.7	26.0	15.0	16.43
ZhongguA	19.4	21.8	24.0	5.0	24.7
Zhong2A	23.4	20.5	26.5	16.0	17.8

Table 3. Analysis output by spectral BiCluster – data in BiCluster (3)

Name	Grain length	Length-width ratio	Chalk rate
HD9802S	8.0	2.9	0.6
R432	7.6	3.2	0.8
Aijiaonante	7.3	3.4	0.8
Chuanxiang28A	6.9	3.2	1.2
Ewan11	5.9	1.9	1.4
KexiangA	7.6	3.8	3.2
SimiaoA	5.6	2.3	4.5
Xiangfeng70A	6.8	3.1	3.4
Xiangling750S	7.0	3.3	0.2
XinA	7.1	3.3	0.6
Zhe04A	5.1	1.9	2.4
Zhenshan97A	7.2	3.1	0.4
Zhong2A	6.8	2.9	0.8
ZhongguA	6.3	3.1	0.8

5 Conclusion

Use data mining to analyze rice breeding data can help breeders discover the underlying patterns and knowledge hidden in these data, the patterns and knowledge can help breeders make a more accurate decision rapidly. Through the study, the method is proved feasible, and the method will be the major trend in crop breeding area in the future.

But from the perspective of the current discipline development, this study needs further investigation, Follow-up studies should focus on the following two aspects:

(1) The data dimension should be further enriched and the data integrity should be improved. The data used in this study did not include the characteristics of the rice planting environment, such as altitude and planting area, but these factors tend to have an important effect on the economic traits of rice. In addition, the amount of data available for this study is to be expanded. The quality of data and the richness of the data characteristics need to be improved, to ensure the effectiveness of the patterns and knowledge extracted from the data.

(2) further researches should be made on bioinformatics. With the development of life science and bioinformatics, the study of phenotypology has also been developing rapidly Phenotypology mainly studies the patterns of phenotypic traits, such as the physical and chemical properties of biology, that vary with mutation and environment [19]. Phenomics studies the phenotype system at the genome scale, hope to explains the unknown function of the genome [20]. The simple analysis of crop phenotype data only shows the relation and function of crop biological traits and economic traits at best, but with the genome data add in, the decisive action and relationships between the specific genetic segment and the specific traits of the crop can be excavated. It is possible to change a gene precisely and make it possible for the offspring to have good traits.

Therefore, the development of future breeding intelligent decision technology must be the close combination of data analysis technology and bioinformatics. Concrete will be reflected in more detailed and rich, continuous, a huge amount of phenotypic data collection, combined with crop genome comparison mining, and provide methods and tools for accurate directional breeding.

References

1. Zhu, R., Deng, J., Li, Y.: The response of rice yield and nitrogen fertilizer utilization to different formula fertilizer. Mod. Agric. (10), 17–21 (2008). (in Chinese)
2. Zhu, R.S., Deng, J.S., Li, Y.: Response of rice yield and nitrogen fertilizer utilization rate to different recipes fertilizer. Mod. Agric. (10), 17–21 (2008). (in Chinese)
3. Xia, R.B.: A study on the breeding science and technology of rice in contemporary China. Nanjing Agricultural University (2009). (in Chinese)
4. Che, S.F., Dai, K.K., Cao, F.L.: The hybrid training algorithm for feedforward neural networks and its application. J. China Univ. Metrol. (4), 424–431 (2014). (in Chinese)
5. Yan, D.C., Zhu, Y., Cao, W.X.: A knowledge model for selection of suitable variety in rice production. J. Nanjing Agric. Univ. (04), 424–431 (2014). (in Chinese)
6. Qi, Y.L., et al.: Interspecific superiority analysis of two rice series subspecies and study of rice parent selection. Henan Agric. Sci. (10), 33–36 (2005). (in Chinese)
7. Gupa, P.K.: Marker-assisted wheat breeding: present status and future possibilities. Mol. Breeding **26**, 145–161 (2010)
8. Guo, Z.: Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor. Appl. Genet. **124**, 261–275 (2012)

9. Yang, J.J., Jin, C.X., Ma, H.C.: Consideration of traditional cross-breeding parent selection factors and its application of modern breeding techniques. Gansu Agric. Sci. Technol. (01), 61–64 (2015). (in Chinese)
10. Chen, C.M.: Information Visualization: Beyond the Horizon, pp. 10–25. Springer, London (2004)
11. Yu, H.M., Liang, Z.P.: Visual data exploration and its applications. Inf. Sci. (04), 599–603 (2007). (in Chinese)
12. Zhao, R.: Design and implementation of decision tree classifier based on WEKA. Central South University (2007). (in Chinese)
13. Hu, Y., Miao, D.Q., Wang, R.Z.: A biclustering algorithm based on rough K-means. Comput. Sci. **34**(11), 174–177 (2007). (in Chinese)
14. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. U.S.A. **97**(22), 12079–12084 (2000)
15. Yang, J., Wang, W., Wang, H., et al.: δ-clusters: capturing subspace correlation in a large data set. In: Proceedings of the 18th IEEE International Conference on Data Engineering (2002)
16. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. Bioinformatics **18**(Suppl 1), S136–S144 (2002)
17. Kluger, Y., Basri, R., Chang, J.T., et al.: Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. (13), 703–716 (2003)
18. Cano, C., Adarve, L., Lopez, J., et al.: Possibilistic approach for biclustering microarray data. Comput. Biol. Med. **37**(10), 1426–1436 (2007)
19. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn, p. 150. Morgan Kaufmann Publishers, Burlington (2016)
20. Gowen, C.M., Fong, S.: Phenome analysis of microorganisms. In: Edwards, D., Stajich, J., Hansen, D. (eds.) Bioinformatics Tools and Applications. Springer, New York (2009). https://doi.org/10.1007/978-0-387-92738-1_14
21. Li, H., Wei, X.L.: Phenomics: a science of unravelling the genotype-phenotype relationship. Biotechnol. Bull. **7**, 41–47 (2013). (in Chinese)