How to Apply Data Mining Technology to the Study of Agricultural Information Data Resources?

Xindong WANG, Haoyue XU, Qian GAO, Haiyan CAI, Junhai LU, Min LI*

Institute of Agricultural Information and Economy, Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang 050051, China

Abstract This paper makes a brief description of the definition and methods of data mining. It describes the characteristics of agricultural data (value delivery, specialization, spatio-temporal bidimensionality) and the status of application of data mining technology in agriculture.

Key words Data mining, Agriculture, Information, Data resources, Research

In the process of agricultural research and practice, a large amount of agricultural data and related information are produced and accumulated, and these data are stored in a large number of agricultural science databases. Currently, agricultural science database has become the most comprehensive agricultural science information system in China with the largest amount of information, the most extensive discipline and the highest level of service. It is one of the basic facilities for the agricultural research work, and the resource for the agricultural scientific research. Agricultural science database involves chemistry, biology, natural resources, environment, atmosphere, plants and animals, and many other disciplines. With the constant deepening of research work and constant accumulation of scientific data, the amount of data becomes very large, and the use of these data for research has become increasingly difficult. Data mining technology provides scientists with an effective means to analyze data in the massive data and look for the knowledge of special relationship[1-3].

1 Definition and methods of data mining

1.1 Definition Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems[4].

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

1.2 Methods Data mining methods include supervised learning, unsupervised learning, affinity grouping (relational analysis), clustering and description. The supervised learning includes classification, estimation and prediction.

2 Characteristics of agricultural data

Due to various characteristics of agricultural data, such as many types of soil data, complex data of crop varieties, frequent pest and disease, and the relationship between fertilizer and climate, agricultural database and knowledge library have many features, such as large size, various dimensions, dynamic degree, low degree of standardization, incompleteness (missing values), uncertainty (unsystematic data or random noise), and sparsity (few or no useful records).

Therefore, to use the data mining tools, it is necessary to first deal with these characteristics of agricultural science and technology data. In addition, agricultural science and technology data have another feature: data are related to time.

2.1 Value delivery Information is of timeliness, and if it is obsolescent, it will become general information, with decreased or no value. The agricultural information mainly aims to reveal the laws of farming, with the characteristics of knowledge, even if it loses timeliness, it can still become a part of scientific theory in a certain subject area.

This is the value delivery of agricultural information. Agricultural information includes basic agricultural research, applied research, development research and other scientific researches. Such information will lose the value of timeliness as time goes, but as the basis for future agricultural science research, its value is still enduring.

2.2 Specialization With the development of science and technology, subdivision and deepening of professional disciplines, the penetration, crossing and fusion between different disciplines have generated many emerging disciplines. In the field of agricultural research, many emerging disciplines in recent years have reflected strong characteristics of specialization, such as functional genomics, proteomics, bioinformatics, precision agriculture, virtual farming, and digital agriculture. The pace of development is quicker than ever before.

2.3 Spatio-temporal bidimensionality The application object of agriculture is the organism, and the growth and development of

Received: August 29, 2013 Accepted: October 8, 2013

* Corresponding author. E-mail: nkylm@163.com

the organism is closely related with the environment. Different time and geographical conditions lead to different choices of agricultural technology. The plant and animal species suitable for growing in tropical environment, may not be suitable for growing in the temperate or boreal regions, and vice versa.

Similarly, some agricultural production techniques suitable for being used in the southern regions do not necessarily apply to the northern region, and vice versa.

Meanwhile, farming has strong features of timeliness and season, and there is a need of different information in different periods, which is a characteristic of agricultural data information research different from that of general information research.

3 Implementation process of data mining

The implementation process of data mining is divided into the following four steps[5]:

3.1 Data definition and data cleansing Creating metadata, defining the related fields, filling default value, smoothing the data noise, identifying and deleting the acnode, and handling inconsistent data.

3.2 Data gathering and data mining model development Integrating the data of multiple databases, cubes and files. Generalizing the data (using the high-level concept to replace the low-level "raw" data), standardization (scaling down the attribute data to a particular interval according to a certain proportion), attribute structuring, and data reduction (removing the weakly associated attributes, and condensing data). Selecting the knowledge discovery methods, to determine the model.

3.3 Data mining Using the definite data mining model to extract the knowledge needed by the users from the data, and using a particular way to signify or using the common expression ways to meet the needs of users, and maximize the work utility of data mining.

3.4 Knowledge assessment Knowledge assessment includes explanation, deduction and model adjustment. Manifesting the knowledge found in a way needed by the users, and carrying out the optimization of some processing stages during the process in accordance with the needs of users, to meet the needs of users as far as possible.

4 Application of data mining technology in agriculture

Agriculture is a complex giant system. China is vast country with multifarious soil types and abundant crop varieties. However, the pests and disease occur frequently, and the relationship among fertilizer, water, density and climate needs to be studied. Agricultural information data have the properties of large number, various dimensions, dynamic degree, incompleteness and uncertainty.

4.1 Application in agrometeorology Meteorological conditions have a major impact on the yield and quality of agricultural production, so forecasting meteorological disasters is of great significance to agricultural production. The data mining technology is used to analyze the local meteorological conditions stored by the meteorological departments over the years, such as temperature, humidity, rainfall, wind and sunshine, etc.

When certain meteorological parameter values are beyond a certain range or there are abnormal changes in some parameter values, it indicates the occurrence of weather disasters. It is necessary to provide users with targeted weather warning services, to provide a scientific basis for disaster prevention and mitigation[6-7].

4.2 Application in pest control decisions Data mining technology is used to predict the likelihood of the occurrence of pests in a certain period of time, as well as the occurrence of other diseases, to provide support to pest control.

4.3 Application in the agricultural market information research With agricultural market price database and Chinese import and export trade data warehouse of agricultural products as the data sources for data mining, the data mining is carried out to process these data sources, to extract the implicit, credible, and unknown potential useful information.

The correlation analysis is used to find the relevant association rules; the statistical techniques are used to infer the price trend of agricultural prices; the classification prediction is used to predict the future prices. The cluster analysis is used to carry out objective classification of agricultural products; acnode analysis technique is used to identify the small probability events with great influence, such as the disaster, financial events, import and export issues[8]. Through deductive analysis, we can carry out research on the laws and trends of variation of prices of agricultural products.

4.4 Application in agricultural production Using data mining technique to predict the growth of greenhouse crops, it is found that in comparison with the traditional forecasting methods, the data mining methods is more accurate.

The data mining technology is used for the pest observation data record, in order to find out the pests forecasting model to provide guidance on the prevention of diseases and pests. It is necessary to use data mining technology for the pest control data, to find out effective pest control management model[9-10].

Using data mining technology for soil database, the soil fertility evaluation rules are made, in order to guide farmers to scientifically use fertilizer and improve soil fertility in the field production process.

4.5 Application in the agricultural environment The data mining is carried out on a large amount of accumulated agri-environmental data, to find the major cause of the difference in the environmental conditions in various regions and find data support to this cause.

References

[1] MA L, CHEN GF. Study on the decision tree's application based on data mining[J]. Agriculture Network Information, 2008, (11): 45. (in Chinese).

[2] HAN YQ, ZHANG Y, YAO YX. Application of data mining based on association rule[J]. Agriculture Network Information, 2008, (11): 73. (in Chinese).

(To page 125)

grant workers, to expand employment channel of migrant workers. For migrant workers, a city is a different social environment, so living and working in a city needs various knowledge and skills. In this situation, government should reinforce training effort of migrant workers, strengthen training of migrant workers in vocational skills, legal awareness and health care, and improve their living and working ability. Training expenses can be shared by government, enterprises and migrant workers, to practically reduce burden of migrant workers.

5.3 Providing housing subsidy for migrant workers and safeguarding education right of migrant workers' children

According to the survey, housing is still the major problem most migrant workers hope to solve. High house rental is a problem troubling migrant workers. Therefore, government should provide moderate housing subsidy for migrant workers, reduce rentals of migrant workers, and lighten their living burden. The *Report to the Eighteenth National Congress of the Communist Party of China* states that we should ensure that children of rural migrant workers in cities have equal access to education. In recent years, Panyu District government practiced the "integration admission scheme" for compulsory education of migrant workers' children. This scheme won high opinion of most migrant workers, but its threshold is too high and only a small portion of migrant workers can be benefited. Based on this, government should continue to improve this scheme, allocate education fund for migrant workers who take their school-age children together, moderately expand public school resources and liberalize admission condition, to satisfy fair compulsory education requirement of migrant workers' children in public schools. Furthermore, it is recommended to build migrant workers self-established schools and provide adequate subsidy in tuition fees, to safeguard compulsory education right of migrant workers' children.

References

[1] ZONG CF, ZHU QZ. Empirical analysis of subsistence conditions of peasant workers——Investigation and analysis on peasant workers of 897 samples in Nanchang[J]. China Rural Survey, 2007(1):47 – 52. (in Chinese).

[2] HAN J, WANG ZH, CUI CY, *et al*. Supply situation of China's migrant workers and the trend in the twelfth five-year period[J]. Reform, 2010(9): 5 – 14. (in Chinese).

[3] XU CX. Falling to the ground, but taking no roots – the urban adaptability of the peasant-workers of the new generation[J]. South China Population, 2007,22(4):52 – 59. (in Chinese).

[4] LI GF. Guangzhou farmer worker existing situation and social security guarantee problem research[J]. Special Zone Economy,2011(3):35 – 37. (in Chinese).

[5] GUO XY, WU P. Analysis and review on rural land use right transfer mode [J]. Modern Management Science,2011(1):77 – 105. (in Chinese).

(From page 121)

[3] BI CG, CHEN GF. Study on the naive Bayes algorithm application based on data mining[J]. Agriculture Network Information, 2010, (3): 19. (in Chinese).

[4] YUAN AX. The improvement and application of C4.5 algorithm for data mining[J]. Journal of Shandong Agricultural University (Natural Science Edition), 2008, 39(3): 461 – 464. (in Chinese).

[5] SUN WW, LIU CX. Data warehouse and data mining [J]. Agriculture Network Information, 2005, (1): 36 – 38. (in Chinese).

[6] ZHENG B, LI J. Elementary discussion on the application of data from agro-product market[J]. Journal of Anhui Agricultural Science, 2008, 36 (18): 7935 – 7936. (in Chinese).

[7] XING PP, XIONG FL. Efficient application of data mining in agriculture databases[J]. Computer Engineering and Applications, 2001, 37(2): 4 – 6. (in Chinese).

[8] XIA ZJ, LI PP, HU YG. The application of data mining in the prediction of crop growth in the greenhouse[J]. Journal of Jiangsu University (Natural Science Edition), 2003, 24(2): 20 – 22, 31. (in Chinese).

[9] Wikipedia. http://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98.

[10] ZHANG JA, ZHANG LY. Research of data mining[J]. Journal of Jilin Agricultural Science and Technology College, 2008, 17(2): 27 – 29. (in Chinese).