TELKOMNIKA, Vol. 11, No. 12, December 2013, pp. 7290~7294
e-ISSN: 2087-278X

Multidimensional Data Mining using a K-mean Algorithm based on the Forest Management Inventory of Fujian Province, China

Yanrong Guo[1], Baoguo Wu*[1], Yang Liu[2]
[1]School of Information Science and Technology of Beijing Forestry University, Beijing 100083, China
[2]Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing 100083, China
*Corresponding author, e-mail: wbaoguo@yeah.net

Abstract

To determine relationships between stand volume and site factors in the absence of information about stand age and density, a classification pattern was established using a clustering analysis algorithm and applied to China fir in Fujian Province. The results showed that slope position, elevation, elevation and humus depth were important factors affecting the stand volumes of young/immature forests, near-mature forests, and mature/overmature forests, respectively. The K-mean algorithm could be used to evaluate the influences of site factors on stand volume under different stand age groups and density conditions.

Keywords: *data mining, K-means algorithm, site factor, forest management inventory*

1. Introduction

Forest resource data play important roles in forest management and decision making. Generally, the forest resource data mainly consist of national forest inventory, forest management inventory and investigation of fixed sampling. These data contribute to sustainable forest management, but rules for huge datasets have not been defined. Existing data cannot be mined for rules, preventing the prediction of future trends. Rapid and efficient data mining has become necessary to enable forest harvesting.

Knowledge is mined and then analyzed from multiple angles, aiding in decision support, process control, and information management [1]. Because of these benefits, data mining is used in many industries. It has been applied to urban residential loads [2], ecological environment compensation [3], intelligent design systems [4], and forestry [5]. In forestry, data mining techniques benefit long-term forest management.

China fir is an important coniferous plantation tree species in Fujian Province, where the climate is arid and sub-tropical. China fir plays an important role because it is the main tree species for afforestation, providing wood that economically benefits the region. We expect to improve China fir growth and efficiency. The relationships between stand volume and site factors must be defined clearly to enable proper management. Previous papers have reported relationships between tree growth and site conditions [6-8], but these analyses have been unidimensional.

However, in the real process of China fir growth, the stand volume is affected by the age, density and site condition. Determining the productivity level of China fir is important because it provides a basis for thinning management. To understand tree growth, realize multidimensional data analysis, and find out the rules, we introduced data mining to forestry and forest management. In this paper, we sought to identify relationships between stand volume and site factors under different stand age and density conditions. The results provide more accurate decision support for tree growth evaluation in forest resource management.

Received June 18, 2013; Revised July 12, 2013; Accepted August 12, 2013

2. Materials and Methods

2.1. Data Collection

In this study, plots were identified using the forest management inventory of Fujian Province. Plots of China fir throughout the province were selected using the following survey data requirements: availability of data on the different site conditions and afforestation times, consistent stand management measures, and relatively little destruction of the stand by humans. Threefold standard deviation was used to eliminate abnormal data, and 52,920 China fir sample plots were chosen for random sampling analysis. Annual data for these plots were distributed as uniformly as possible.

The main survey factors were the compartment, subplot, stand age, dominant tree species, tree species composition, and stand average height. The components of the environmental variables were contained. These variables were landform; elevation; slope, slope direction, slope position; soil type, texture, and structure; humus thickness; stand age; management measure; health level; site type; and afforestation time (1956–2006).

2.2. Data Mining Framework

Figure 1 depicts the design of the assessment system, including data mining. The assessment process was divided into the following steps: (i) data preparation, (ii) clustering analysis for data mining, and (iii) categorization of volume and site conditions for different stand ages and densities.

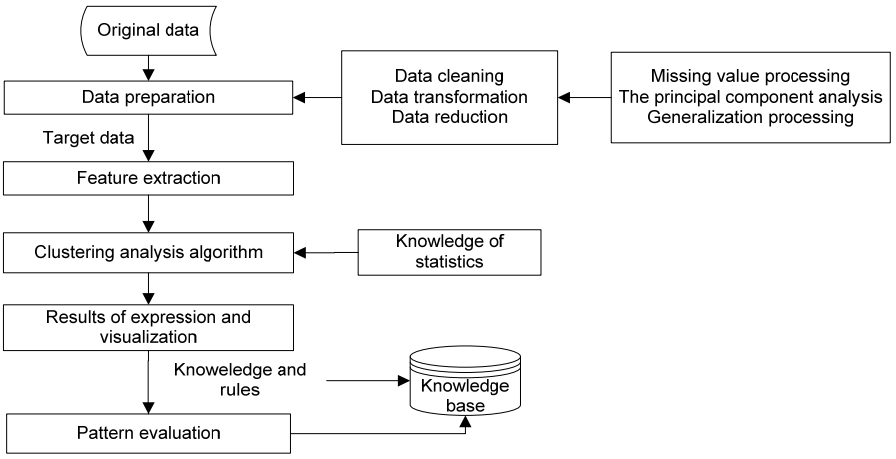

Figure 1. Flow Chart of China Fir Data Mining

2.3. Data Preparation

2.3.1. Data Cleaning

The data contained some outliers, noise, and missing or inconsistent values. In such cases, we replaced data points with mean values of the corresponding variables.

2.3.2. Data Transformation

In this study, data transformation consisted of generalization and normalized processing. Generalization processing replaced the lower levels of data objects with more abstract concepts. Stand age was defined as young growth, middle-aged forest, near-mature forest and mature. In normalized processing, attribute data were projected proportionally onto a specific small scale. This process was used in data mining to eliminate deviations among the different attribute data. The dimensions of the attributes were not consistent or comparable. The standardization method was applied to solve the problem of non-unified dimensions in all indicators and then compared with the assessment index:

$$H_{ij}' = H_{ij} - H_{ijmin} / H_{ijmax} - H_{ijmin'} \quad (1)$$

Where H_{ij}' is the standardization value, H_{ij} is the observed value, H_{ijmax} is the maximum of all observed values, and H_{ijmin} is the minimum of all observed values.

2.3.3. Data Reduction

Much time is wasted on the analysis of large and complex datasets. To avoid this problem, data reduction methods must be researched. Complex datasets containing some correlation can be reduced to a few indicators that fully reflect the original information and are independent of one another. In other words, this technology can maintain the integrity of the dataset while allowing efficient data mining and improving the quality of results.

Extensive research has been performed on efficient algorithms that can manage high dimensionality [9-12]. High-dimensional data are often transformed into lower-dimensional data by principal component decomposition [13]. Principal component analysis was used for data reduction in this study.

2.4. Cluster Algorithms

The distance measure was used to compute cluster similarity for most clustering algorithms. In data mining, clustering is a discovery process that groups or compartmentalizes a dataset to maximize intra-cluster and minimize inter-cluster similarity. In cluster analysis, the K-mean algorithm is one of the most efficient and widely used methods in practice [13, 14].

The K-mean algorithm is initialized from some random or approximate solution, as follows [15, 16]: (i) K objects are selected randomly as initial cluster centers from n data objects, (ii) the distance of each object from the mean of each clustering object (cluster center) was calculated and a new partition is created using the minimum distance, (iii) new cluster centers are computed, and (iv) steps (ii) and (iii) are iterated until no change occurs in any cluster.

The specifics of the K-means algorithm are described below. Each repetition assigns each point to its nearest cluster, and points belonging to the same cluster are then averaged to derive new cluster centers. Each repetition successively improves the cluster centers until they become stable [10, 13]. The algorithm uses the equation:

$$E = \sum_{i=1}^{K} \sum_{P \in C_i} |p - m_i|^2 \quad (2)$$

Where E is the sum of squared errors for all objects in the database, p is the data matrix, and m_i is the centroid of cluster C_i. In the K-mean method, the k cluster must be kept as compact as possible in the interior of the cluster, and clusters must be kept as distant from one another as possible.

3. Results of Expression and Visualization

3.1. Determination of Site Factors

The principal component decomposition method was used for data reduction in this study. The eigenvalues of the eight main components exceeded 1, and the accumulative contribution rate reached 86.17% (Table 1). The main components were landform, elevation, slope, slope position, exposure, soil type, humus depth, and soil thickness.

Table 1. Statistics of the Main Components

Main component	Eigenvalue	Contribution rate	Accumulative contribution rate
1	2.98275402	18.64	18.64
2	2.35229036	14.70	33.34
3	2.11518018	13.22	46.56
4	1.88209686	11.76	58.33
5	1.63179546	10.20	68.53
6	1.23453778	7.72	76.24
7	1.02892308	6.43	82.67
8	1.00982300	3.50	86.17

3.2. Classification Determination

According to important level, the orders in which the eight main components affected stand volume were: slope position, slope, exposure, soil thickness, elevation, humus depth, soil type, and landform for young forests; elevation, slope position, soil thickness, exposure, humus depth, slope position, landform, and soil type for immature timber; elevation, humus depth, soil thickness, soil type, landform, slope position, exposure, and slope for near-mature forests; and humus depth, elevation, soil thickness, slope position, soil type, landform, slope, and exposure for mature and overmature forests (Table 2).

Table 2. Categorized Results of Relationships between Site Factors and Volume for Different Ages and Densities

Age groups	Density (tree·ha^{-1})	Landform	Elevation	Slope	Slope position	Exposure	Soil type	Humus depth	Soil thickness
Young forest	1000-4500	—	-0.025	-0.073	-0.076	-0.049	-0.005	0.013	0.037
Immature timber	600-4200	-0.068	-0.257	-0.069	-0.241	-0.158	-0.029	0.076	0.237
Near-mature forest	450-3600	-0.169	-0.345	-0.010	-0.100	-0.097	-0.222	0.317	0.249
Mature/ overmature forest	450-3300	-0.185	-0.419	-0.144	-0.230	-0.123	-0.189	0.505	0.390

4. Discussion and conclusion

Stand density was negatively correlated with forest illumination and temperature in forests of different densities, but tree growth was positively associated with these components. Because the densities of young forests and immature timber were high, growth of China fir was limited mainly by illumination and temperature. Consistent with this situation, exposure and slope position impacted forest illumination and photosynthesis, thereby affecting tree growth. Although gentle slopes (<25°) were conducive to China fir growth [17], sample plots with slopes >25° accounted for 83.23% of young forests. Thus, slope may be a key factor in the stand volume of young forest.

The density of immature timber was sufficiently high and then result in illumination affected the tree growth. The illumination of upper slopes was adequate, which promote the tree growth (Table 2). Slope position was a key factor influencing the stand volume of immature timber.

The densities of near-mature, mature, and overmature forests were relatively low. Such stands should have adequate illumination and abundant shrubs, herbs, and forest litter. The soil condition is the most important direct factor because soil thickness influences the root system capacity and fertilizer absorption. Thus, soil thickness was a key factor influencing the stand volumes of near-mature, mature, and overmature forests.

The elevations of young forests, immature timber, near-mature forests, and mature/overmature forests were 140–990 m, 108–1225 m, 150–1115 m, and 150–1225 m, respectively. There are significant differences in the stand volume among the forest types for four age groups (Table 2). Thus, an elevation had a large impact on the growth of China fir.

In this study, because all plots were in low and middle mountain areas, landform did not significantly influence tree growth. If the study plots had been in different areas, such as hills and low, middle, and high mountain areas, the results would have differed. In low and middle mountain areas, plant residues decomposed rapidly under mild conditions, producing loose and fertile soil. These areas had greater annual rainfalls, which enhanced tree growth. In high mountain areas, tree growth was inhibited by lower temperatures and higher evaporation rates.

In conclusion, under different stand age groups and density conditions, stand volume was influenced by the different components of site factor for the young, immature, near-mature and mature/overmature forests of China fir. Given the relationships between stand volume and basic tree factors, such as diameter, tree height, crown width, and branch height, we should

consider that site factors also affected these basic factors. To further evaluate the influences of site factors on tree height and diameter under the same age groups but different density conditions, stand density should be divided into more classes.

Acknowledgement

This study was supported by the national natural science foundation-funded project (No. 31170513) and the national high technology research and development program (863 program) (2012AA102003).

References

[1] PN Tan, Steinbach M, Kumar V. Introduction to data mining. WP Co. 2006.
[2] Y He, Y Wang, T Luo, A He, J Wang. Urban Residential Load Combined Forecast Model Based on Data Mining Techniques and Panel Data Theory. *Journal of Computational Information Systems*. 2010; 6(6): 1801-1808.
[3] JQ Xiang. Research on mining development in yunnan under ecological environment compensation. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(5): 2853-2859.
[4] T Jing, Y Yuan. Intelligent design system of mechanical products based on data mining and knowledge based engineering. *Journal of Theoretical and Applied Information Technology*. 2012; 46(1): 237-244.
[5] CP Chen, BG Wu, YG Jia , DD Lu. A Study on Applying Techniques of Data Mining in Forest in From Ation Management. *He Bei Journal of Forestry and Orchard Research*. 2004; 19(2): 149-153.
[6] JC Lu. Affection of terrain factors on growth of Plantation Chinese fir in high altitude mountains in eastern Fujian Province. *Journal of Fujian Forestry Science and Technology*. 2006; 33(2): 120-128.
[7] J Huang, W Min, CC Cai, SH Lu. Effects of different densities on the cunninghamia lanceolata in middle age. *Application of Statistics and Management*. 2006; 25(1): 111-116.
[8] JH Meng. Building forestry data warehouse for forest management enterprise and their application PhD dissertation.China: Beijing Forestry University, Beijing. 2011.
[9] CC Aggarwal, PS Yu. Finding generalized projected clusters in high dimensional spaces. ACM. 2000.
[10] Hinneburg, Keim DA. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. Citeseer. 1999.
[11] C Ordonez. Clustering binary data streams with K-means. In, Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. 2003; 12-19.
[12] H Wei, XJ Li, Y Guan, et al. On the model checking of the spacewire link interface. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(2): 740-746.
[13] C Ding, X He. K-means clustering via principal component analysis. In, Proceedings of the twenty-first international conference on Machine learning. ACM. 2004.
[14] JA Hartigan, MA Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*. 1979; 100-108.
[15] AK Jain, Dubes RC. Algorithms for clustering data. Prentice-Hall, Inc. 1988.
[16] AK Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 2010; 31(8): 651-666.
[17] HB Zhang. Fujian Forest. Beijing: China Forestry Press.1993.