2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.101

Land Indexes Selection Based on Fuzzy Measures

Jinfeng Wang[1a], Yueming Hu[2], Xiangning Ren[2], Wenzhong Wang[3]

[1]Department of Mathematics and Informatics, South China Agricultural University, Guangzhou, 510642, China

[2]Key Laboratory of the Ministry of Land and Resources for Construction Land Transformation, Guangzhou, 510642, China

[3]Department of economics and management, South China Agricultural University, Guangzhou, 510642, China

[a]wangphoenix@163.com

Keywords: Land index, Fuzzy measure, L1-Norm Regularization

Abstract. In land consolidation, it is very important to construct an effective index system. Land data is characteristic of big volume, complex varieties and more indexes. We need select a group of good index for some goals of land consolidation according to concrete demand. In this paper, Fuzzy Integrals is adopted to finish the feature selection. Fuzzy Integral is a kind of infusion tool based on fuzzy measure which can describe the importance of each feature or feature subset. Some researchers can obtain the optimal solution for Fuzzy measure using soft computing tools. When the Fuzzy integrals can be transformed to a linear equation, L1-norm regularization method is applied to solve the linear equation system and find a solution with the fewest nonzero values for fuzzy measure. The solution with the fewest nonzero can show the degree of contribution of some features or their combinations for decision. This method provides a quick and optimal way to determine the land index system for preparing the following land research.

Introduction

In land consolidation, the land index system is important for land evaluation. So the selection of land indexes affects the results of evaluation and decision model. Currently, many researchers have focused on the optimization and selection of land index system. T.L. Saaty proposed an index selection method based on the analytic hierarchy process with weight[1]. Then he proposed Least Square Method(LSM) and Least Logarithm Square Method(LLSM) for confirming the previous weight[2]. But land indexes are very multiple and complicated. It may be related to society, economics and ecology. Traditionally, land index system was constructed according to experts' experience. Due to human factors, the evaluation results lost the objectivity and consistency. It is too hard to obtain a set of accurate weight in the analytic hierarchy process.

In this paper, a method based on computational tool-fuzzy measure is proposed for land index selection. It can avoid the human factors' effect and confirm the final index subset objectively. Fuzzy measure can represent the importance of index and index combination for decision[3]. We can solve the values of fuzzy measure by using L1-Norm method to obtain a sparse vector. Those indexes with non-zero fuzzy measures will be kept in final index set.

This paper is constructed as follows. The introduction has been given in section 1. The fuzzy integral model for solving fuzzy measure is given in section2. In next section the land indexes are described in detail. Section 4 shows the experiments and results. Conclusions will be drawn in section 5.

Model based on fuzzy measure

We are given a data set consisting of L example records, called training set, where each record contains the value of a decisive feature, Y, and the value of predictive features $x_1, x_2, \ldots, x_n$. Positive integer L is the data size. The classifying feature indicates the class to which each

example belongs, and it is a categorical feature with values coming from an unordered finite domain. The set of all possible values of the classifying feature is denoted by $C = c_1, c_2, \ldots, c_m$, where each c_k, $k = 1, 2, \ldots, m$, refers to a specified class. The feature features are numerical, and their values are described by an n-dimensional vector, ($f(x_1), f(x_2), \ldots, f(x_n)$). The range of the vector, a subset of n-dimensional Euclidean space, is called the feature space. The jth observation consists of n feature features and the classifying feature can be denoted by $(f_j(x_1), f_j(x_2), \ldots, f_j(x_n), Y_j)$, $j = 1, 2, \ldots, L$. Before introducing the model, we give out the fundamental concepts as follows.

Fuzzy Measure. Let $X = x_1, x_2, \ldots, x_n$, be a nonempty finite set of feature features and $P(X)$ be the power set of X.

To further understand the practical meaning of the Fuzzy Measure, let us consider the elements in a universal set X as a set of predictive features to predict a certain objective. Then, for each individual predictive feature as well as each possible combination of the predictive features, a distinct value of a Fuzzy measure is assigned to describe its influence to the objective. Due to the nonadditivity of the Fuzzy Measure, the influences of the predictive features to the objective are dependent such that the global contribution of them to the objective is not just the simple sum of their individual contributions. Set function μ is nonadditive in general. If $\mu(X) = 1$, then μ is said to be regular. The monotonicity and non-negativity of fuzzy measure are too restrictive for real applications. Thus, the signed fuzzy measure, which is a generalization of fuzzy measure, has been defined [12, 13] and applied.

A signed Fuzzy measure allows its value to be negative and frees monotonicity constraint. Thus, it is more flexible to describe the individual and joint contribution rates from the predictive features in a universal set towards some target. Let f be a real-valued function on X. The fuzzy integral of f with respect to μ is obtained by

$$\int f d\mu = \int_{-\infty}^{0} [\mu(F_\alpha) - \mu(X)] d\alpha + \int_0^{\infty} \mu(F_\alpha) d\alpha \tag{1}$$

where $F_\alpha = \{x | f(x) \geq \alpha\}$, for any $\alpha \in (-\infty, \infty)$, is called the $\alpha - cut\ of\ f$.

To calculate the value of the Fuzzy integral of a given real-valued function f, usually the values of f, i.e., $f(x_1), f(x_2), \cdots, f(x_n)$, should be sorted in a nondecreasing order so that $f(x_1') \leq f(x_2') \leq \ldots \leq f(x_n')$, where $(x_1', x_2', \ldots x_n')$ is a certain permutation of $(x_1, x_2, \ldots x_n)$. So the value of Fuzzy integral can be obtained by

$$\int f d\mu = \sum_{i=1}^{n} [f(x'_i) - f(x'_{i-1})] \mu(\{x'_i, x'_{i+1}, \ldots x'_n\}), \quad \text{where } f(x_0') = 0 \tag{2}$$

The Fuzzy integral is based on linear operators to deal with nonlinear space.

Transformation of Fuzzy integral. To be convenient, Wang [8] proposed a new scheme to calculate the value of a Fuzzy integral with real-valued integrand by the inner product of two $(2^n - 1)$-dimension vectors as $\int f d\mu = \sum_{j=1}^{2^n - 1} z_j \mu_j$ (3)

where $z_j = \begin{cases} \min\limits_{i: frc(\frac{j}{2^i}) \in [\frac{1}{2}, 1)} f(x_i) - \max\limits_{i: frc(\frac{j}{2^i}) \in [0, \frac{1}{2})} f(x_i), & \text{if it is bigger than zero or } j \text{ is } 2^n - 1; \\ 0, & \text{otherwise.} \end{cases}$ (4)

for $j = 1, 2, \cdots, 2^n - 1$

with a convention that the maximum on the empty set is zero. Here, $frc(\frac{j}{2^i})$ denotes the fractional part of $\frac{j}{2^i}$. In the above formula, if we express j in the binary form $j_n j_{n-1} \cdots j_1$, then $\left\{ i \middle| frc(\frac{j}{2^i}) \in [\frac{1}{2}, 1) \right\} = \{ i | j_i = 1 \}$ and $\left\{ i \middle| frc(\frac{j}{2^i}) \in [0, \frac{1}{2}) \right\} = \{ i | j_i = 0 \}$.

A significant advantage of this new calculation scheme is that it can easily discover the coefficients matrix of a system of linear equations with the unknown variables μ when the Choquet integral is applied in further applications, such as regression and classification [4, 8, 9]. In those practical applications, values of the signed Fuzzy measure are usually considered as unknown parameters which are to be estimated using the training data sets[5, 6]. The adoption of this new scheme make it convenient for using an algebraic method, such as the least square method, to estimate the value of μ, and furthermore, to reduce complexity of computation.

After having this transformation, we can obtain the Fuzzy measure for a known dataset by using L1-norm Regularization.

Solutions of the Fuzzy Measure. For determining the Fuzzy Measure, researchers have proposed many methods. In our past work, we used GA to learn the value of Fuzzy measure for each concrete dataset. In this paper, we propose a new method based on L1-norm regularization.

In many regression problem, the most popular function used is the Least Squares estimate, alternately referred to as minimizer of the residual sum of squared errors (RSS)[7]: $RSS = \sum_{i=1}^{n} (y_i - \omega_0 - \sum_{j=1}^{p} x_{ij} \omega_i)^2$. Regularization addresses the numerical instability of the matrix inversion and subsequently produces lower variance models. It is easy to see that the following penalized RSS function with respect to ω and ω_0: $\sum_{i=1}^{n} (y_i - \omega_0 - \sum_{j=1}^{p} x_{ij} \omega_i)^2 + \lambda \sum_{j=1}^{p} \omega_j^2$. This is referred to as L2 regularization. In order to simplify the notation used, we reduce it to the following problem (in matrix notation): $\|X\omega - y\|_2^2 + \lambda \|\omega\|_2^2$. While L2 regularization is an effective means of achieving numerical stability and increasing predictive performance, it cannot address another important problem with Least Squares estimates, parsimony of the model and interpretability of the coefficient values. It does not encourage sparsity in some cases [10]. So a trend has been become to replace L2-norm with an L1-norm recently. This L1 regularization has many of the beneficial properties of L2 regularization, but obtains sparse solutions that are more easily interpreted [7]. This property is what our algorithm wants. In Fuzzy integrals, determining the Fuzzy measure is the key procedure in the whole model. Fuzzy measure represents the importance of features and the interaction degree of features combined.

We hope get a solution of Fuzzy measure with the fewest nonzero values to find the most important features and feature combinations. Using L1-norm regularization, we can minimize the following formula to reduce the size of nonzero in Fuzzy Measure: $\left\| \sum_{j=1}^{2^n - 1} z_j \mu_j - y \right\|_2^2 + \lambda \|\mu\|_1$. We can control the condensation compress degree for Fuzzy measure by adjusting the parameter λ. Author of [11] proposed the Least Absolute Selection and Shrinkage Operator (LASSO) model based on Gauss-Seidel method. The obvious advantages of the Gauss-Seidel approach are its simplicity and its low iteration cost. We applied this kind of LASSO to solve the above L1-Norm problem. Finally, the optimal Fuzzy measure can be obtained.

Experiments and Analysis

In first, we must investigate by collecting materials, spatial image recognition, field investigation, and questionnaire for the land potential evaluation. All factors which include land-use state, economics, social factors, ecological environment and policy have been considered. The results would be summarized and analyzed so that the whole situation of the 'three old' project is acknowledged precisely. All indexes considered are described as Table 1.

Our model is applied to the shunde's data for obtaining the key index system. Several classical evaluation models are adopted for testing the feature selection results. But the current number of indexes of three old data is rather large for Fuzzy integrals to deal with. It will take very long time to learn the Fuzzy Measure. So the feature selection is a necessary step. Based on previous research, reduct in Rough Sets is adopted to process the data before index selecting and classifying. The feature subsets selected are shown in Table 2. We can see the size of feature subsets from Rough Sets is greatly smaller than original one. It can greatly promote the efficiency of Fuzzy integrals because the time of learning the signed Fuzzy measure is reduced greatly.

Table 1. All indexes of three old land

Criteria layer	Sub criteria layer	Evaluation indexes
Land-use(A)	landscapes	Building coordination
		Block crush degree
	Building situation	Building age
		Building structure
	Development strength	Volume ratio
		changing of building density
Economical factors(B)	Basic land price	Basic land price
	Investment strength	changing degree of investment amount
	Per capita net income	Per capita net income
Social factors(C)	Population density	Population density
	Social welfare	Medical and sanity
		Education
		Public welfares(park , square)
	Basic facilities	Traffic connectivity
	Green degree	Green ratio
Ecological factors(D)	Ecological environment	Noisy pollution
		Air pollution
		Water pollution
Policy(E)	Compensation and emplacement	compensation
		emplacement
	Responding	Responding activity
	Management	Public participation

The parameter λ in L1-Norm method is used for controlling the degree of compression for Fuzzy Measure. We set the value of λ as 0, 1, 5, 10, 20, 50 and 100 respectively. The larger the value of λ is, the fewer the number of zero in solution is. The compressing the Fuzzy measure simplify the computation of Fuzzy integrals at the cost of performance. It needs to select an appropriate value for λ to balance the complexity and the performance. Finally, the value of λ is determined as 100. The binary forms corresponding to fuzzy measure with values are {10000000} and {1111100} after compressing by L1-Norm, which means keeping indexes from x1 to x5. Based on fuzzy measures, fuzzy decision tree is applied to land data. All results with feature selection and fuzzy measures are listed in Table 2. We can see that the size of tree is compressed as the number of features is decreased and the performance is improved.

Table 2. The results with feature selection and compressing by fuzzy measure

Types / performance	All features	with feature selection	with fuzzy measure
Prediction accuracy	89.12%	93.06%	**94.34%**
Selected features	all	{4,6,8,9,10,11,15}	{4,6,8,9,10}
Number of leaves	10	7	**4**
Size of tree	19	13	**7**

Conclusions

Land data usually includes many indexes which may be not useful for decision. Feature selection must be executed before prediction. Due to the great number of features, the computational complexity of determining Fuzzy measures is very large. Finding the values of each fuzzy measure is a hard work. In this paper, we use the L1-norm method to solve the problem of complexity. The Fuzzy measure with the fewest nonzero values can be obtained by compressing using L1-norm regularization, which can reduce the complexity greatly with promoting performance. Experimental results show that the indexes selection can help to reduce the complexity and improve performance. Selecting one optimal value of parameter λ can keep a balance between complexity and performance. The detailed values of fuzzy measure can be confirmed to describe the interaction of features with respect to contribution for decision.

Acknowledgement

This research is supported by Ministry of Key Projects in the National Science & Technology Pillar Program during the Twelfth Five-year Plan Period(No. 2013BAJ13B05), the National Natural Science Foundation of China(No. 61202295), and the National Social Science Foundation of China (Projects No. 10CJY024).

References

[1] Saaty, Thomas L.; Peniwati, Kirti. Group Decision Making: Drawing out and Reconciling Differences. Pittsburgh, Pennsylvania: RWS Publications (2008) .

[2] Saaty, Thomas L. . Principia Mathematica Decernendi: Mathematical Principles of Decision Making. Pittsburgh, Pennsylvania: RWS Publications(2010).

[3] M. Sugeno: Theory of Fuzzy Integrals and Its Applications. Doctoral Thesis, Tokyo Institute of Technology (1974).

[4] W. Wang, Z.Y. Wang , G. J. Klir: Genetic Algorithm for Determining Fuzzy Measures from Data. Journal of Intelligent and Fuzzy Systems. Vol. 6, pp. 171-183 (1998).

[5] Z.Y. Wang , G.J. Klir: Fuzzy measure Theory. New York: Plenum (1992)

[6] Z.Y. Wang, K.S. Leung, J. Wang: A Genetic Algorithm for Determining Nonadditive Set F unctions in Information Fusion. Fuzzy Sets and Systems, Vol. 102, pp. 463-469 (1999).

[7] T. Hastie, R. Tibshirani and J.H. Friedman: The Elements of Statistical Learning. Spring, (2001).

[8] Z. Wang: A new genetic algorithm for nonlinear multiregressions based on generalized Choquet integrals, Proc. 12th IEEE Intern. Conf. Fuzzy Systems, vol. 2, pp. 819-821 (2003)

[9] K. S. Leung, M. L. Wong, W. Lam, Z. Wang and K. Xu: Learning nonlinear multiregression networks based on evolutionary computation, IEEE Trans. On Systems, Man and Cybernetics, Part B, vol. 32, no. 5, pp. 630-644 (2002)

[10] R. Tibshirani: Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288(1996)

[11] Shirish Krishnaj Shevade and S. Sathiya Keerthi: A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics, 19(17):2246–2253 (2003).

[12] T. Murofushi, M. Sugeno, M. Machida: Non monotonic Fuzzy Measures and the Choquet integral. Fuzzy Sets and Systems, Vol. 64, 73-86 (1994)

[13] M. Grabisch, T. Murofushi, M. Sugeno (editors). : Fuzzy Measures and Integrals: Theory and Applications, Physica-Verlag(2000).