Comprehensive Economic Strength Assessment Based on Decision Tree Algorithm

Yuan Hou, Min Wu[*]

Shanghai Key Lab of Trustworthy Computing, Software Engineering Institute, East China Normal University, Shanghai 200062, P.R.China

Received 23 November 2014, www.cmnt.lv

Abstract

In this paper, we apply some computer software techniques in socio-economics to study an economic decision-making problem, the comprehensive economic strength assessment problem. To achieve delicacy socio-economical decision, we first conduct an assessment of various influencing factors involved in the decision problem based on analytic hierarchy process, and then apply two decision-tree algorithms to establish the appropriate models with corresponding classification rules. We also use Matlab to fit the data for the purpose of prediction and validation. The experiments show the effectivity and accuracy of decision-tree algorithms in supplying valuable suggestions for economic decision-making problems.

Keywords: Complex Decision; Decision Tree Algorithm; Data Mining; ID3 Algorithm; C4.5 Algorithm

1 Introduction

Decision-making[1], refers to the process of determining the optimal selection for an action on the basis of some possessed information and experiences, through applying certain methods, techniques and tools for analysis and calculation on various factors affecting the achievement of objectives. However, in real life, many factors are often contained in complex decision-making problems, which may involve a large amount of information, and complicated interactions and overall performance. So, making accurate decisions become very difficult.

In early studies of sociology, decision-making was fulfilled through single-case studies or through simple summarization, and more emphasis is put on proposing specific measures. With the development of statistics and computer tools, more and more statistical tools, such as SPSS (Statistical Product and Service Solutions), have been applied in quantitative analysis of sociological research to extract hidden statistical laws to aid in decision-making. However, statistical methods often assume that the various factors are independent variables or there exists only a simple linear or quadratic relation among the variables. Therefore, for complex decision-making problems which involve complicated relations among various affecting factors, statistical analysis merely can only yield results with large deviation from the actual situation. In recent years, a new idea [2] of qualitative comparative analysis (QCA) has emerged in sociological research field. The QCA method can produce highly accurate results, but due to combinatorial exploration, it cannot handle decision problems of large size (usually appropriate for problems with 5 or 6 factors). Complex decisions usually involve many factors, and, moreover, it is often very difficult to extract through subjective judgments the delicate relationships or rules among factors. Currently, a hot research topic in the field of computer science is data mining [3] technology. Data mining is a non-trivial process of revealing implicit, previously unknown and potentially useful information from large amounts of data. Data mining, mainly based on artificial intelligence, machine learning, pattern recognition, statistical methods, and etc, can carry on highly automatic analysis on the data and make inductive reasoning. In recent years, data mining has aroused great concern in information industry, and is now widely applied in business management, production control, market analysis, engineering design, etc.

The research in this paper is motivated from an interdisciplinary project of sociology and computer science -- *The Design and Development of a Decision Support System for Comprehensive National Strength.* In this paper, we focus on the modeling and analysis of a country's comprehensive economic strength assessment problem, by combing the AHP method, two decision tree algorithms ID3 ([4]) and C4.5 ([5,6]) and data fitting techniques.

2 Problem formulation

In this paper we will study a complicated economical decision problem, i.e., how to assess a country’s comprehensive economic strength based on various economic indicators.

At first, let us determine the various important factors that affect a country's economic strength. Based on preliminary investigation, a country's economy falls into four main aspects: consumption, finance, investment and fiscal. Applying the expert’s opinion elicitation method (or, DELPHI method), we can break these four indicators down further into three basic indicators respectively. Therefore there are totally twelve basic indicators, listed as the consumer price index, inflation, personal remittances, official exchange rate, real interest rates, broad money, domestic credits provided by financial sectors, external debt stock, foreign direct investment, claims on other sec-

[*] ***Corresponding author’s*** e-mail: mwu@sei.ecnu.edu.cn

tors, tariff and revenue. In addition, we select the gross domestic product (GDP) as a comprehensive assessment indicator of economic strength.

To clarify the hierarchical relationships among the above thirteen indicators, we use the analytic hierarchy process (Analytic Hierarchy Process)[7] to construct the basic framework of a country's comprehensive economic strength. We obtain a 3-layer analytic hierarchy model of comprehensive national strength (see Figure 1). Then, using some objective methods like principal component analysis, or the subjective methods such as Delphi method and importance sorting method, we can attach each attribute from the second layer an extent of importance relative to the upper layer. Consequently, we construct the determination matrix and calculate the weight vector. At last, to determine the coefficients for each factor, we carry on a consistency test until the calibration coefficient is less than 0.1.

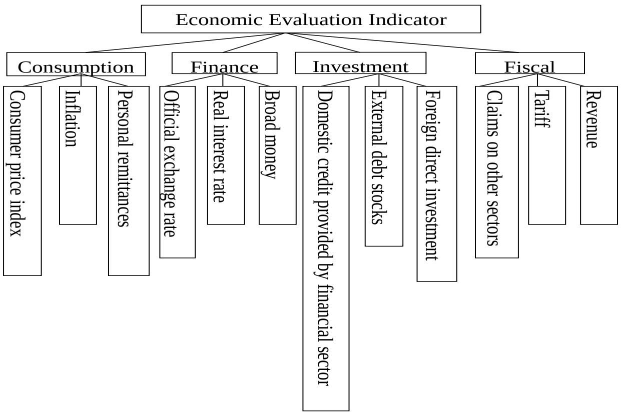

FIGURE 1Hierarchical relationships among various indicators

3 Decision tree algorithms

In this section, we will discuss how to apply decision tree algorithms to excavate the relationships among the factors of economic decision-making problem and to extract classification rules contained in data.

Decision tree algorithms are typical classification methods in data mining and are designed to infer classification rules (represented as a decision tree) from a set of unordered and irregular instances.

3.1 ENTROPY

In his great paper [6] published in 1948, Shannon systematically discussed how to quantify and encode information.

To measure uncertainty of information, Shannon defined the amount of information (entropy) of an event a_i[7] as follows

$$Entropy(a_i) = p(a_i)\log_2 \frac{1}{p(a_i)},$$

where$p(a_i)$ denotes the probability of occurrence of the event a_i. Specially, we define $Entropy(a_i) = 0$ if $p(a_i) = 0$ Let a_1 , a_2 $\cdots\cdots a_n$ be a set of n incompatible events, meaning that each time one and only one event occurs. Then the entropy of a_1 , a_2 $\cdots\cdots a_n$ is defined as

$$Entropy(a_1, a_2, ..., a_n) = \sum_{i=1}^{n} Entropy(a_i) = \sum_{i=1}^{n} p(a_i)\log_2 \frac{1}{p(a_i)}.$$

3.2 ID3 ALGORITHM

ID3(Iterative Dichotomiser 3) algorithm[4] is a classifycation prediction algorithm proposed by R.Quinlan. The main idea of ID3 algorithm is to use information gain to split attributes: the higher the information gain is, the more classified information the attribute contains.

Let S be a collection of training samples. The information gain of an attribute A relative to the sample set S is defined as

$$Gain(S, A) = Entropy(S) - Entropy(S, A) \ ,$$

where

$$Entropy(S, A) = \sum \frac{|S_v|}{|S|} \cdot Entropy(S_v),$$

in which |S| represents the number of elements in a set S, S_v is the subset of S consisting of all samples whose attribute A value equals v, i.e.,

$$S_v = \{s \in S \mid A(s) = v\}.$$

From the above definition, the higher the information gain Gain(S,A) is, the more classified information the attribute A provides. In ID3 algorithm, in each step during the process of constructing a decision tree, we select the attribute of highest gain value as the split point. ID3 algorithm is then a recursive attribute splitting process. There are two recursion terminating criterions: either all the data belong to the same class or there is no attributes that can be split further.

The main steps of ID3 algorithm are illustrated in the following Figure 2.

ID3 algorithm has a clear computational logic, low complexity and fast computing speed, and then is of strong practical value. However, ID3 algorithm can handle only discrete attribute values, and moreover, the algorithm may render biased results since it adopts information gain as the split criterion.

As mentioned above, ID3 algorithm can only handle discrete attribute values, while almost all the basic indicators in economics are continuous attributes. Therefore, before applying ID3 algorithm in the economic decision making problems, we need to discretize the continuous data first [8]. There are usually two means of discretization of continuous data: 1) use K-means clustering algorithm to classify the data which belong to the same cluster as one class; 2) use the discretization function supplied by the data mining platform WEKA[9]. In this paper, we choose the latter method to discretize the economic indicators. There are two reasons for this: one is that a large amount of data processing are involved in our decision making problem, and the other is that the supervised discretization method provided by WEKA is more interactive and easier to operate.

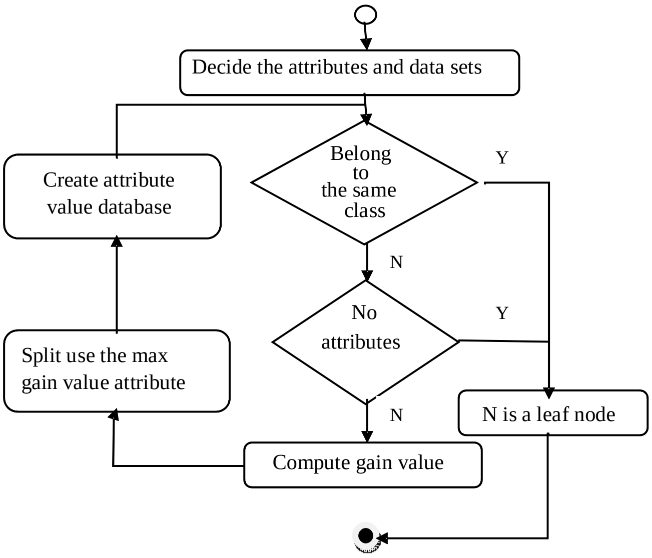
Decide the attributes and data sets
Belong to the same class
Y
N
No attributes
Y
N
Create attribute value database
Split use the max gain value attribute
N is a leaf node
Compute gain value

FIGURE 2 Steps of ID3 algorithm

3.3 C4.5 ALGORITHM

For control experiments, we will also apply C4.5 algorithm, another decision-tree algorithm from data mining, to do the data analysis. C4.5 algorithm ([5,10]) was proposed in 1993 by Quinlan as an improvement of ID3 algorithm. Unlike ID3 algorithm, C4.5 algorithm splits the attributes according to information gain ratio (rather than the gain in ID3)，which can avoid generating biased results to a large extent.

Before introducing the notion of information gain ratios, let us first define the split information of an attribute A relative to the sample set S to be

$$SplitInfo(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|},$$

where S_i represents the ith sample subset split by the attribute A. Then, the information gain rate is defined to

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}.$$

be the ratio of the information gain and split information of the attribute A：

Then C4.5 algorithm recursively selects the attribute of highest gain rate as the split point, until all the training samples are classified and the decision tree is constructed.

4 Experiments

In this section, we will apply the data processing and analysis method to study the assessment of comprehensive national strengths based on thirteen economic indicators described in Section 2. The sampling data has been collected from the World Bank Open Data website[11]. We will apply ID3 and C4.5 algorithms to process, analyze, train and predict the dataset, so as to extract certain hidden classification rules.

4.1 DATA COLLECTION

In order to reduce the complexity of the data and to maintain data integrity while ensuring the accuracy of the results obtained from data training and classification, we will focus on the thirteen factors in Figure 1. We have collected from the World Bank Open Data website the relevant data of 30 countries of small or medium size (e.g., Afghanistan, Algeria, etc) over a five-year period (2008-2012) (c.f. Table 1 and Appendix in [12]). Due to space limit, we cannot list all the dataset. The reader can refer to [12] for more details on the dataset, the output of the algorithms and the corresponding conclusions of the decision-making problem.

4.2 APPLICATION OF ID3 ALGORITHM

We first analyze the dataset of 2012. Using the platform supplied by WEKA([9]), we discretize the thirteen attribute values in Figure 1. The discretized values are shown in Table 1(c.f. also Figure 4-8 in [12]). Each column of Table 1 represents respectively the discretized values of the thirteen factors in Section 2, in which L (Low) and H (High) are the discretized attribute values, and P (Relatively Poor) and R (Relatively Rich) represent values of the comprehensive indicator GDP.

TABLE 1 Discretized Data

Data sets	1	2	3	···	11	12	13
	CPI	Inf	Off	···	Tar	Rev	GDP
Afghanistan	L	L	L	···	L	L	R
Albania	L	L	L	···	L	L	R
···	···	···	···	···	···	···	···
SierraLeone	H	H	L	···	H	L	P
SouthAfrica	H	H	L	···	H	H	P
Thailand	L	L	L	···	L	L	P

Based on the discretized value of Table 1, we apply ID3 algorithm to construct the decision tree model, as illustrated in Figure 3.

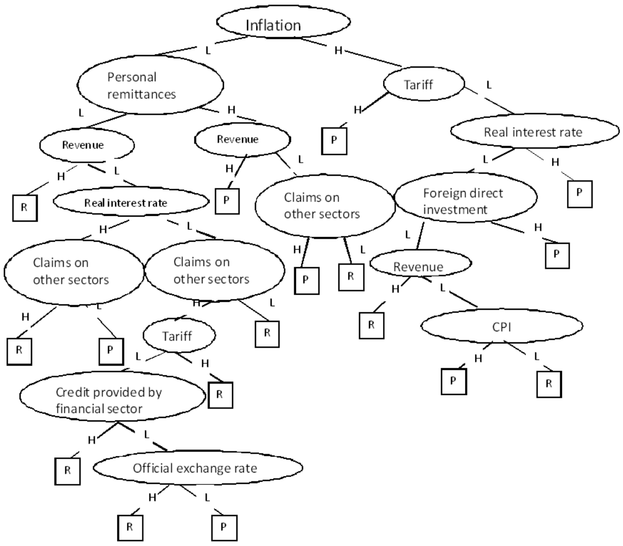

FIGURE 3 ID3 Decision tree

The ID3 decision tree in Figure 3 has17 leaf nodes, which represent 17 rules of the form IF…AND…THEN…

As an example, we list below one of these 17 rules:

IF the annual inflation rate is H (high) AND the tariff is H (high), THEN GDP is P (relatively poor).

Other rules can be elicited in the same way.

4.3 APPLICATION OF C4.5 ALGORITHM

In order to compare with ID3 algorithm, we also use the discretization method supplied by C4.5 algorithm in Matlab. We take the twelve continuous attribute values as the training sample set input of C4.5 algorithm, and obtain a decision tree shown in Figure 4.

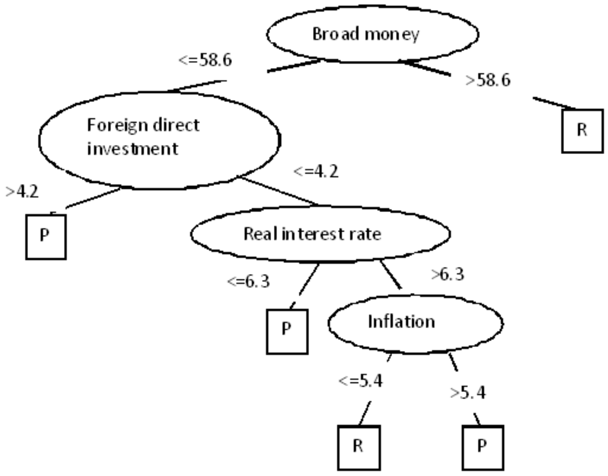

FIGURE 4 C4.5 Decision tree

Compared with the decision tree in Figure 3 constructed using ID3 algorithm, the decision tree in Figure 4 has 5 leaf nodes, which represents 5 rules. As an example, we list below one of these 5 rules:

IF the board money is <=58.6 AND the foreign direct investment >4.2, THEN GDP is P (relatively poor).

Other rules can be elicited in the same way.

5 Comparison and analysis

In Section 4, we apply ID3 and C4.5 algorithms to train a dataset containing 13 attributes. The decision tree obtained from ID3 algorithm has 16 non-leaf nodes and 17 leaf nodes, and the maximum number of layers is 8, while the one obtained from C4.5 algorithm has only 4 non-leaf nodes and 5 leaf nodes, and the maximum number of layers is 4. The comparison between the two results is pictured in Figure 5. For our example, the ID3 decision tree is more complicated than C4.5 decision tree, in that the numbers of non-leaf nodes, leaf node and layers in ID3 decision tree are 4, 3.4 and 2 times of the C4.5 decision tree. Moreover, the rules generated by C4.5 algorithm is much simpler, clearer, and easier to understand than those generated by ID3 algorithm.

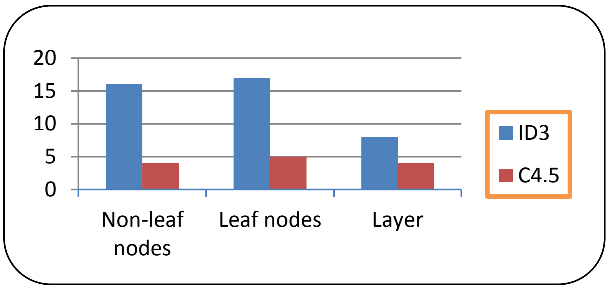

FIGURE 5 Comparison of different decision trees

For the purpose of prediction and validation, we also apply Matlab for data fitting on the five-year dataset from 2008 to 2012. We construct a series of second-order expressions and use them to make a prediction of the 2013 data. The predicted attribute values can be further used to construct the validation database.

Use the 2012 data as a training sample dataset, we construct the 2012 ID3 and C4.5 decision trees and the corresponding 2013 prediction results through rule matching. A comparison between the predicted results and the real data is shown in Table 2 (c.f. Table 4-18 in [12]).

TABLE 2 Prediction classification and actual ranking

Decision Results	ID3	C4.5	Reality
Afghanistan	P	P	R
Algeria	R	R	R
…	…	…	…
South Africa	R	R	R
Thailand	R	R	R

From Table 2, we see that ID3 algorithm has predicted successfully 19 countries among 30 countries, with the prediction accuracy rate 63.3%, while C4.5 algorithm has predicted successfully 21 countries, the prediction accuracy rate being 70%, 6.7% higher than ID3.

Similarly, we apply the ID3 and C4.5 algorithms on the five year data from 2008 to 2012. The accuracy rate for each year can be found in Table 3.

TABLE 3 Accuracy Rates for 2008-2012

Accuracy Rate	ID3	C4.5
2008	56.60%	76.60%
2009	63.30%	63.30%
2010	73.30%	70%
2011	63.30%	63.30%
2012	63.30%	70%

To sum up, for our comprehensive economic strength assessment problem based on 13 economic indicators, C4.5 algorithm yields a decision tree model which is much simpler and of higher accuracy. This decision tree model can supply auxiliary suggestions on how to assess a country's comprehensive economic strength. For example, we can measure the country's economic strength through rule matching once we obtain certain data of the economic key indicators.

6 Conclusion and future work

In this paper, by combining socioeconomics and computer software techniques including decision-tree classification algorithms and data fitting for prediction and validation, we provide a new thought on how to supply auxiliary technical and delicate support for economic decision-making problem.

Through experimental results in this paper, we find that C4.5 algorithm can generate simple, easy-to-use and accurate rules, and so can provide valuable auxiliary suggestions for social-economic decision problems.

This study of our paper comes from an interdisciplinary project of sociology and computer science. For future work, we hope that more in-depth study on the data processing and data mining techniques and more hints on how to formulate the relationship among different factors will produce more valuable results on the study of refined decision-making problems.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 11371143 and 61321064, and the Fundamental Research Funds for the Central Universities.

References

[1] Reason J 1990 *Human Error* Cambridge University Press

[2] Charles R 1987 *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies* University of California Press

[3] Frawley W, Piatetsky-Shapiro G and Matheus C 1992 *A.I. Magazine*, 213-228

[4] Quinlan J R 1986 *Machine Learning*, **1**(1): 81-106

[5] Breiman L, Friedman J, Olshen R and Stone C J 1984 *Classification and regression trees*. Chapman & Hall/CRC

[6] Shannon E 1948 *Bell System Technical Journal*, **27** (3): 379–423

[7] Saaty T L 2000 *Fundamentals of the Analytic Hierarchy Process*

[8] RWS Publications

[9] Belton V 1986 *European Journal of Operational Research*, **26**(1):7-21

[10] Holmes G, Donkin A and Witten I H 1994 *Proc. 2nd Australia and New Zealand Conference on Intelligent Information Systems*, 357-361

[11] Quinlan J R 1994 *Machine Learning*, 16(3):235-240

[12] World Bank Open Data. http://www.worldbank.org.cn, **2014**-07-01.

[13] HOU Y 2014 *Modeling and Simulation for Data Sets of Complex Decision-Making Problems*. Undergraduate Thesis, East China Normal University, Shanghai, China, **2014**. (in Chinese)

Authors

Yuan Hou, March 27, 1993, Zhangjiajie Hunan, P.R.China

Current position, grades: Master Student
University studies: East China Normal University
Research interest: Trustworthy Computing
Experience: In July 2014 Bachelor Degree from Software Engineering Institute, East China Normal University and in the same year, he was recommended for a direct entry into graduate study in East China Normal University exempt from the Graduate Entrance Examination.

Min Wu, December 11, 1976, Xinyu Jiangxi, P.R.China

Current position: Associate Professor
University: East China Normal University
Research interest: Trustworthy Computing
Publications: 30 papers in high quality international journals and conferences, including the International Symposium on Symbolic and Algebraic Computation (ISSAC), Journal of Symbolic Computation, Science in China, Frontiers of Computer Science, Communications in Nonlinear Science and Numerical Simulation, Abstract and Applied Analysis, etc.
Experience: In 2005, graduated from Institute of Systems Science, Chinese Academy of Sciences (China) and INRIA Sophia Antipolis (France), and obtained the PhD degrees Doctor of Science from Chinese Academy of Sciences, China and Docteur en Mathematiques from Université de Nice-Sophia Antipolis, France. In 2006, she joined the East China Normal University.