Investigación

Use of decision trees algorithm for the territorial logistic planning

Uso del algoritmo árboles de decisión en la planificación logística territorial

Jorge Quijada-Alarcón, Nicoletta González,
Francisco Soler y Alberto Camarero

Revista de Investigación

Volumen III, Número 2, pp. 051–058, ISSN 2174-0410
Recepción: 14 Mar'13; Aceptación: 31 Jun'13

1 de octubre de 2012

Abstract

Data mining, and in particular decision trees have been used in different fields: engineering, medicine, banking and finance, etc., to analyze a target variable through decision variables. The following article examines the use of the decision trees algorithm as a tool in territorial logistic planning. The decision tree built has estimated population density indexes for territorial units with similar logistics characteristics in a concise and practical way.

Keywords: Data mining, decision trees algorithm, territorial logistic planning.

Resumen

La minería de datos, y en particular los árboles de decisión han sido utilizados en diferentes campos: ingeniería, medicina, banca y finanzas, etc., para analizar una variable objetivo a través de variables de predicción. El siguiente artículo examina el uso del algoritmo de árboles de decisión como una herramienta en la planificación logística territorial. El árbol de decisión construido ha estimado índices de densidad de población para unidades territoriales con similares características logísticas en un modo conciso y práctico.

Palabras Clave: Minería de Datos, Arboles de decisión, planificación logística territorial.

1 Introduction

In the last decades, it has been developed numerous techniques for analysis and modeling of data in different areas of statistics and artificial intelligence. Data mining is the process of discovering actionable and meaningful patterns, profiles, and trends by sifting through your data using pattern recognition technologies [1]. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns [2]. Data mining it's also considered in the intersection of the artificial intelligence (machine learning) and the statistics techniques, even though there is some remarkable differences between them, i.e., artificial intelligence has been more concerned to offer algorithmic solutions with an acceptable computational cost, while the statistic has been concerned over the power of generalization of the results obtained, i.e., be able to infer results to more general than the studied situations [3]. In the late 1980s and mostly on 1990s data mining begins to be as we know it, as a result of the increase of the computing power, a faster data collection and the apparition of new techniques of data learning and storage.

Any problem with stored data it is a problem to be dealt with using data mining. Some of these problems are: search of the unexpected by description of the multivariate reality, search for associations, typologies definition, detection of temporal sequences and prediction [3]. However, it's important to do a process of data preparation. Zhang et al. [4] argue for the importance of data preparation at three aspects: (1) real-world data is impure; (2) high-performance mining systems require quality data; and (3) quality data yields high-quality patterns. Some areas with data mining applications are: civil engineering [5, 6, 7], medicine [8, 9, 10]; education [11, 12, 13]; Banking and Finance [14, 15, 16].

One of the data mining techniques is decision tree. A decision tree is a logical model represented as a binary (two-way split) tree that shows how the value of a target variable can be predicted by using the values of a set of predictor variables. The first widely-used program for generating decision trees was "AID" (Automatic Interaction Detection) developed by Morgan and Sonquist [17]. AID was followed by many other decision tree generators including THAID [18], and ID3 [19] and, later, C4.5 [20]. The theoretical underpinning of decision tree analysis was greatly enhanced by the research done by Breiman et al. [21], embedded in a program they developed CART® (a registered trademark of Salford Systems). Recent advancements in decision tree analyses include the TreeBoost method [22] and Decision Tree Forests [23].

In particular, some decision trees applications in planning of logistic node in the latest research are: modeling on the environmental impact of airport deicing activities to determine important explanatory variables for predicting levels of chemical oxygen demand and dissolved oxygen in the airport's waterways [24]; evaluating countermeasures to secure cargo at United States southwestern ports of entry [25]; spatial decision tree application to traffic risk analysis [26]; Deriving decision rules to locate export containers in container yards, including a decision tree from the set of the optimal solutions to support real time decisions [27]. Others researches include concepts of territorial planning using decision trees: consider the classification of location contexts [28]; primary and secondary road network analysis using decision trees algorithm [29].

2 The territorial logistic planning and the general trend analysis of logistic platforms.

Since the first experience in Land Use Planning (LUP) in 1933, to develop the Tennessee River System in the United States of America, and integrate the water control with the conservation and preservation of the land resources [30, 31], the LUP's definitions [32, 33] have considered the perspective of sustainability. This planning process requires consider each logistic node impact on the spatial economy and regional development, and the possible synergies between them, even though the planners often allocate the nodes on the territorial space without considering the one effect over the others. The territorial logistic planning is an essential part of a sustainable regional development.

The general trend analysis of the exploitation and planning of nodes or logistic platforms is to compare the ratios and parameters of the international literature. One of the most used bibliographies [34] discusses that according to the Growth Poles Theory that the development is not uniform and is carried out in specific places around activities of agglomeration. The infrastructures around an activity can improve the accessibility to suppliers and customers. The transport terminals, therefore are a kind of economic forces to generate links to other sectors of the economy and become sources of economic activity, so are often considered as growth poles; e.g. a terrestrial terminal growth strategy turns around the formation of logistics platforms, where the distribution centers share installations and also have a better access to the transport terminal. There are two major groups of techniques in the analysis of efficiency and performance of transport infrastructure: known as Data Envelopment Analysis DEA [35, 36, 37] has been traditionally used for the relative efficiency estimation of a set of peer entities called Decision Making Units (DMUs). Econometric estimation of distances functions [38, 39, 40, 41] its an empirical estimation of cost functions. The methodology of DEA models generalizes the traditional analysis of activity ratios allowing consider simultaneously multiple inputs and/or outputs. DEA and the estimation of frontier functions are alternatives to calculate the production boundary and therefore the efficiency of the production and costs. DEA is a non-parametric method based on linear programming while the estimation of frontier functions use econometric methods (parametric methods).

3 Methodology

In order to aim the research purpose to determine the relationship between a target territorial logistic variable and their decision variables, it was developed the following methodology in two steps: step 1 to determine the work stage and step 2 to develop the artificial intelligence model.

Step 1: Determination of the work stage

Diagnosis and state of the art. It consists on the review of the State of the art to identify the set of variables to characterize the target territorial logistic variable of study using specialized search services in two steps.

1. *Determination of physical and functional decision variables*: i.e. a systematic study of all the physical and functional decision variables (of ports, airports, road network, railroads, etc) susceptible to research for our target territorial logistic variable. It includes the study of the variables of territorial context as population, environment and regional economic grow.

2. *Getting the value of the decision variables*: Once the decision variables to be studied are known get its variables values using different information sources. The variable values have to be referred to territorial units identified in the country. Data set is established as follows:

$$(n, M) = (n_1, n_2, n_3, ..., n_k, M) \quad (1)$$

   The dependent variable M is the target variable and the vector n is composed of the predictor variables $n_1, n_2, n_3, ..., n_k$. This data set is called the learning or training dataset and is needed to build a decision tree model.

Step 2: Construction of the model of artificial intelligence

The software chosen for to build the decision tree is DTREG [42]. The process DTREG uses to build and prune a tree is complex and computationally intensive. Here is an outline of the steps:

1. Build the tree.
   (a) Examine each node and find the best possible split.
      - Examine each predictor variable.
         - Examine each possible split on each predictor.

   (b) Create two child nodes.
   (c) Determine which child node each row goes into. This may involve using surrogate splitters.
   (d) Continue the process until a stopping criterion (e.g., minimum node size) is reached.
2. Prune the tree.
   (a) Build a set of cross-validation trees.
   (b) Compute the cross validated misclassification cost for each possible tree size.
   (c) Prune the primary tree to the optimal size.

The method for evaluating the quality of splits when building classification trees is Gini, where each split is chosen to maximize the heterogeneity of the categories of the target variable in the child nodes.

$$GiniIndex = 1 - \sum_j p_j^2 \tag{2}$$

Where the equation 2 contains values of probability of p_j for a class j.

The method used to determine the optimal tree size is V-fold cross validation, a technique for performing independent tree size tests without requiring separate test datasets and without reducing the data used to build the tree.

4 Results and discussion

The target logistic variable selected for the study is the Population Density (Population Per Square Kilometer) for explain the distribution of the population according to the distribution and development of the logistics nodes in the different territorial units that conform the country.

Figure 1 shows the principal child nodes and its predictor variables of the decision tree built for the target variable. After complete the process of build the tree, the predictor variables chosen by the decision tree algorithm were:

- Panama Canal Influence (binary 1/0)
- Secondary road-network (Kilometer Per Square Kilometer)
- Primary road-network (Kilometer Per Square Kilometer)

The principal predictor variable is the Panama Canal Influence that is a binary variable: 1 for the territorial units under the Panama Canal Influence, and zero for the others.

The territorial units that conforms the node 3 in the figure 1 correspond to the strip of land adjacent to the Panama Canal, with special economic areas, container ports, transatlantic railroad and Tocumen International airport. This strip of land include Panama City and Colon City

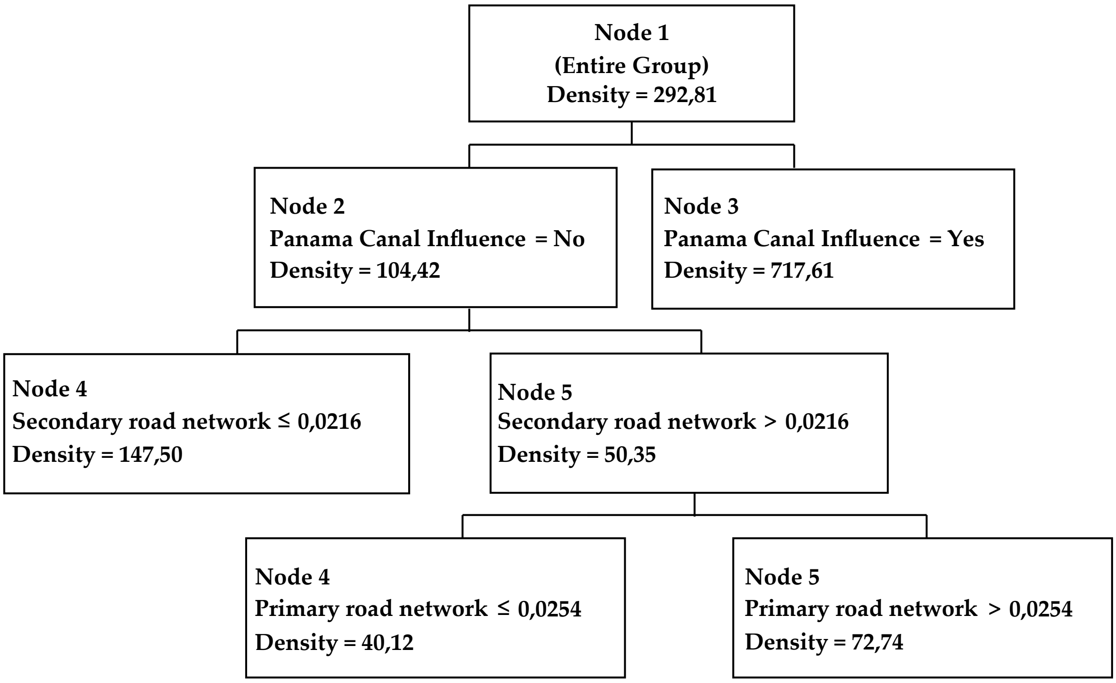
Node 1
(Entire Group)
Density = 292,81
Node 2
Panama Canal Influence = No
Density = 104,42
Node 3
Panama Canal Influence = Yes
Density = 717,61
Node 4
Secondary road network ≤ 0,0216
Density = 147,50
Node 5
Secondary road network > 0,0216
Density = 50,35
Node 4
Primary road network ≤ 0,0254
Density = 40,12
Node 5
Primary road network > 0,0254
Density = 72,74

Figure 1. Decision three.

and has the highest population density index. The activities of transport, storage, communication, wholesale and retail trades, and real estate activities make the greatest contribution to the provincial GDP.

For the rest of the country where the logistics nodes development is lower than in the strip of land adjacent to the Panama Canal the principal characteristic is the relationship between the road network and the population density.

5 Conclusions

The decisions tree built for the target variable Population Density chose three predictor variables which explain the distribution of the population in the country according to the distribution and development of the logistics nodes. This allows to territorial planners to take in consideration the present scenario.

The decisions tree built has estimated Population Density index for territorial units with similar logistics characteristics in a concise and practice way. The characteristics of each group of territorial units has been analyzed using expert criteria.

The predictor variable "Panama Canal Influence" chosen by the decision tree algorithm, it is the principal predictor variable and allows to the territorial planners divide the country at least in two principal zones: the first under the Panama Canal influence with a highest logistics nodes development, and the second without the Panama Canal influence and a lowest logistics nodes development.

References

[1] MENA, J., *Data mining your website*, Digital Pr, 1999.

[2] HAND, D. J., MANNILA, H., and SMYTH, P., *Principles of data mining*, The MIT press, 2001.

[3] ALUJA BANET, T., *La minería de datos, entre la estadística y la inteligencia artificial*, Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, vol. 25, Nº 3, pp. 479–498, 2001.

[4] ZHANG, S., ZHANG, C., and YANG, Q., *Data preparation for data mining*, Applied Artificial Intelligence, vol. 17, Nº 5-6, pp. 375–381, 2003.

[5] CHUNCHUN, H., NIANXUE, L., XIAOHONG, Y., and WENZHONG, S., *Traffic flow data mining and evaluation based on fuzzy clustering techniques*, International journal of Fuzzy Systems, vol. 13, Nº 4, pp. 344–349, 2011.

[6] LEE, W. H., TSENG, S. S., SHIEH, J. L., and CHEN, H. H., *Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services*, IEEE Transactions on Intelligent Transportation Systems, vol. 12, Nº 4, pp. 1047–1056, 2011.

[7] GUO,Y., HU, J., and PENG, Y., *Research on CBR system based on data mining*, Applied Soft Computing, vol. 11, Nº 8, pp. 5006–5014, 2011.

[8] SHOUMAN, M., TURNER, T., and STOCKER, R., *Using data mining techniques in heart disease diagnosis and treatment*, in Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on, IEEE, 2012, pp. 173– 177.

[9] AARTS, S., VOS, R., VAN BOXTEL, M., VERHEY, F., METSEMAKERS, J., and VAN DEN AKKER, M., *Exploring medical data to generate new hypotheses: an introduction to data and text mining techniques in epidemiology*, Multimorbidity in general practice: Adverse health effects and innovative research strategies, 2012.

[10] KOH, H. C. AND TAN, G., *Data mining applications in healthcare*, Journal of Healthcare Information Management, vol. 19, Nº 2, p. 65, 2011.

[11] HUNG, J.L., HSU, Y.-C., and RICE, K., *Integrating data mining in program evaluation of k-12 online education*, Educational Technology Society, vol. 15, Nº 3, pp. 27–41, 2012.

[12] YADAV, S. K., BHARADWAJ, B., and PAL, S., *Mining education data to predict student's retention: A comparative study*, arXiv preprint arXiv:1203.2987, 2012.

[13] ASIF, R., MERCERON, A., and PATHAN, M., *Mining student's admission data and predicting student's performance using decision trees*, ICERI2012 Proceedings, pp. 5121–5129, 2012.

[14] RAVI, V., NEKURI, N., and RAO, C. R., *Privacy preserving data mining using particle swarm optimisation trained auto–associative neural network: an application to bankruptcy prediction in banks*, International Journal of Data Mining, Modelling and Management, vol. 4, Nº 1, pp. 39–56, 2012.

[15] KWAK, W., SHI, Y., and KOU, G., *Bankruptcy prediction for korean firms after the 1997 financial crisis: using a multiple criteria linear programming data mining approach*, Review of Quantitative Finance and Accounting, pp. 1–13, 2012.

[16] PRASAD, U. D., and MADHAVI, S., *Prediction of churn behavior of bank customers using data mining tools*, Business Intelligence Journal, vol. 5, Nº 1, 2012.

[17] MORGAN, J. N. AND SONQUIST, J. A., *Problems in the analysis of survey data and a proposal*, (Original AID), 1963.

[18] MORGAN, J. N. AND MESSENGER, R. C., *Thaid, a sequential analysis program for the analysis of nominal scale dependent variables*, Survey Research Center, University of Michigan, 1973.

[19] QUINLAN, J.R., *Induction of decision trees*, Machine learning, vol. 1, Nº 1, pp. 81-106, 1986.

[20] QUINLAN, J. R., *C4. 5: programs for machine learning*, Morgan kaufmann, 1993.

[21] BREIMAN, L., FRIEDMAN, J., STONE, C. J., and OLSHEN, R. A., *Classification and regression trees*, Wadsworth International Group, 1984.

[22] FRIEDMAN, J. H., *Stochastic gradient boosting*, Computational Statistics Data Analysis, vol. 38, Nº 4, pp. 367–378, 2002.

[23] BREIMAN, L., *Decision tree forests*, Machine Learning, vol. 45, Nº 1, pp. 5–32, 2001.

[24] FAN, H., TARUN, P. K., SHIH, D. T., KIM, S. B., CHEN, V. C. P., ROSENBERGER, J. M., and BERGMAN, D., *Data mining modeling on the environmental impact of airport deicing activities*, Expert Systems with Applications, vol. 38, Nº 12, pp. 14899–14906, 2011.

[25] BAKIR, N. O., *A decision tree model for evaluating countermeasures to secure cargo at united states southwestern ports of entry*, Decision Analysis, vol. 5, Nº 4, pp. 230–248, 2008.

[26] ZEITOUNI, K. AND CHELGHOUM, N., *Spatial decision tree application to traffic risk analysis*, in Computer Systems and Applications, ACS/IEEE International Conference on. 2001, 2001, pp. 203–207.

[27] KIM, K. H., PARK, Y. M., and RYU, K. R., *Deriving decision rules to locate export containers in container yards*, European Journal of Operational Research, vol. 124, Nº 1, pp. 89–101, 2000.

[28] SANTOS, M. Y. AND MOREIRA, A., *Automatic classification of location contexts with decision trees*, 2006.

[29] QUIJADA-ALARCON, J., GONZÁLEZ CANCELAS, N., CAMARERO ORIVE, A., and SOLER FLORES, F., *Road network analysis using decision trees algorithm: A case of study of Panama*, in proceedings in Advanced Research in Scientific Areas. The 1st Virtual International Conference, EDIS Publishing Institution of the University of Zilina, Slovak Republic, 2012.

[30] TENNESSEE WRITERS' PROJECT, *Tennessee: a guide to the state*, Scholarly Pr, 1939.

[31] MONOD, J., and DE CASTELBAJAC, P., *L'aménagement du territoire*, Presses universitaires de France, 1987.

[32] AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, *Guidelines for land-use planning*, Food & Agriculture Org., vol. 1, 1993.

[33] MINISTRY OF AGRICULTURE AND LAND AFFAIRS OF THE REPUBLIC OF SOUTH AFRICA, *White Paper on Spatial Planning and Land Use Management*, 2001.

[34] RODRIGUE, J. P., COMTOIS, C., and SLACK, B., *The geography of transport systems*, Taylor & Francis, 2009.

[35] ROLL, Y., and HAYUTH, Y., *Port performance comparison applying data envelopment analysis (DEA)*, Maritime Policy and Management, vol. 20, Nº 2, pp. 153-161, 1993.

[36] TONGZON, J., *Efficiency measurement of selected Australian and other international ports using data envelopment analysis*, Transportation Research Part A: Policy and Practice, vol. 35, Nº 2, pp. 107-122, 2001.

[37] BONILLA, M., CASASUS, T., MEDAL, A., and SALA, R., *An efficiency analysis with tolerance of the Spanish port system*, International Journal of Transport Economics= Rivista Internazionale de Economia dei Trasporti, vol. 31, Nº 3, 2004.

[38] LIU, C. I., JULA, H., and IOANNOU, P. A., *Design, simulation, and evaluation of automated container terminals*, Intelligent Transportation Systems, IEEE Transactions on, vol. 3, Nº 1, pp. 12-26, 2002.

[39] TOVAR, B., JARA-DÍAZ, S., and TRUJILLO-CASTELLANO, L., *A multioutput cost function for port terminals: some guidelines for regulation*, World Bank Policy Research Working Paper No.3151, 2003.

[40] TOVAR, B., TRUJILLO, L., and JARA-DÍAZ, S., *13 Organisation and Regulation of the Port Industry: Europe and Spain*, Essays on microeconomics and industrial organisation, vol. 1262, pp. 189, 2004.

[41] TOVAR, B., JARA-DÍAZ, S., and TRUJILLO, L., *Funciones de producción y costes y su aplicación al sector portuario. Una revisión de la literatura*, Documento de trabajo, vol. 6, 2004.

[42] SHERROD, P., *Dtreg predictive modeling software*, Software available at http://www.dtreg.com, 2003.

Sobre los autores:

Nombre: Jorge Quijada-Alarcon
Correo electrónico: jorge.quijada@alumnos.upm.es
Institución: Universidad Politécnica de Madrid, España.

Nombre: Nicoletta González Cancelas
Correo electrónico: n.gcancelas@upm.es
Institución: Universidad Politécnica de Madrid, España.

Nombre: Francisco Soler Flores
Correo electrónico: f.soler@upm.es
Institución: Universidad Politécnica de Madrid, España.

Nombre: Alberto Camarero Orive
Correo electrónico: alberto.camarero@upm.es
Institución: Universidad Politécnica de Madrid, España.