Predictive models of economic systems based on data mining

José Cazal
Universidad Nacional de Asunción
Email: jcazalw@gmail.com

***Abstract*—Data election to build a representative model able to explain socio-economic phenomena is a challenge within the model construction stage itself. Knowing what data to include within the studies and what to discard is a challenge, and again, at the same time, a great amount of possible factors affecting each variable behavior must be found. In complex phenomena, the number of factors affecting a variable is enormous, and isolating a variable can become a hopeless effort. Besides, there are also factors that are difficultly observable or inherently not observable that must be considered, those ones known as errors or perturbations in a relation that have influence in the constructed model outputs. Techniques applied in data mining can give support to the studies in the moment of analyzing the socio-economic phenomena and demonstrate results obtained through a scientific and reliable way. Data mining is proposed as a valid option in the study of indicators contrasting the traditional methodology (econometrics). An experiment was conducted to contrast two cultures in the use of statistical modeling. One assumes that the data are generated by stochastic GIVEN data model (Data Modeling Culture). The other one uses algorithmic models and treats the data as unknown mechanism (Algorithmic Modeling Culture).**

1. Introduction

Econometry is the science that studies and tries to explain socio-economic phenomena. Econometry is a branch of Economics that applies mathematical and statistical techniques when analyzing economics theories and solving economic problems through models.

Explained by Wooldridge [1], econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, evaluating and implementing government and business policy. Nowdays, the analysis of these phenomena requires specialists with a high technical preparation and considerable experience level, which is an own restriction due to the studies of the techniques used in Econometry in the studies of economic phenomena.

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the web, other massive information repositories, or data streams [2].

It makes use of artificial intelligence, automatic learning, statistics and data base systems. The general objective of the data mining process consists of eliciting information from a data set and transforming them into a comprehensible structure for its posterior use.

Owed to data mining can be of a great help when it comes to analyzing information, what we intend to do it is to apply data mining techniques to solve socio-economic problems. Econometrics, it represents the Data Modeling Culture. The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from:

response variables = f (predictor variables, random noise, parameters)

The analysis in Algorithmic Modeling Culture considers the inside of the box complex and unknown. Their approach is to find a function f(x) algorithm that operates on x to predict the responses y.

2. Tool based on Data Mining

In order to achieve the objective, a tool is developed as main core the Weka API. Weka is a collection of machine learning algorithms for data mining tasks developed by the "Machine Learning Group at the University of Waikato" [3]. The algorithms can be either applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

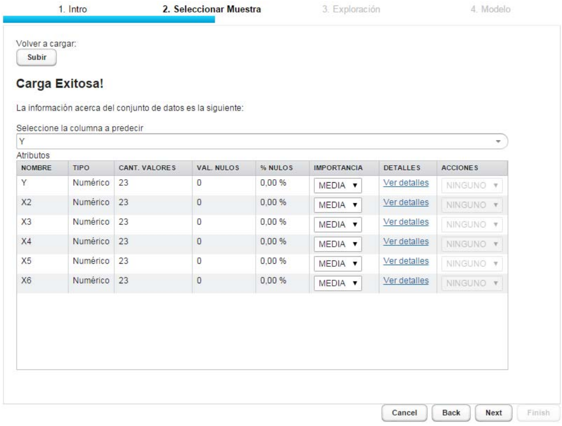
Based on SEMMA model (SEMMA is an acronym that stands for Sample, Explore, Modify, Model and Assess) [4], different techniques are applied in each phase of the K.D.D process to help us to reach the most reliable predictive model possible.

The tool must be semi-automatic, this means that the users as an economic domain expert, must only know the set of data. In each step, this tool will collect information from the user and will apply this knowledge extracted to cover needs at every step.

3. The SEMMA Phases.

Sample: The process starts with data sampling, e.g., selecting the data set for modeling. In this

978-1-4673-8111-6/15 $31.00 © 2015 IEEE
DOI 10.1109/DMIA.2015.20



Volver a cargar:

Carga Exitosa!

La información acerca del conjunto de datos es la siguiente:

Seleccione la columna a predecir: Y

Atributos

NOMBRE	TIPO	CANT. VALORES	VAL. NULOS	% NULOS	IMPORTANCIA	DETALLES	ACCIONES
Y	Numérico	23	0	0,00 %	MEDIA	Ver detalles	NINGUNO
X2	Numérico	23	0	0,00 %	MEDIA	Ver detalles	NINGUNO
X3	Numérico	23	0	0,00 %	MEDIA	Ver detalles	NINGUNO
X4	Numérico	23	0	0,00 %	MEDIA	Ver detalles	NINGUNO
X5	Numérico	23	0	0,00 %	MEDIA	Ver detalles	NINGUNO
X6	Numérico	23	0	0,00 %	MEDIA	Ver detalles	NINGUNO

Figure 1. EMMA explore phase

TABLE 1. SETTING ACTIONS BY IMPORTANCE AND NULL VALUES TABLE

	% Nulls values		
Importance	More of 50%	Between 50% y el 25%	Less than 25%
HIGH	Suggest delete	Suggest impute value	Suggest impute value
MED	Delete	Suggest delete	Suggest impute value
LOW	Delete	Delete	Suggest delete

phase we need to know whether the data is a time series or cross-sectional . According to it, we filter in the first instance the possible algorithms we can use.

Explore: This phase covers the understanding of the data by discovering anticipated and unanticipated relationships between the variables, and also abnormalities. When obtaining the attributes (depending variables of the model) it is required that the user points out the level of importance the attribute has in the set of data, according to him/her. There are 3 levels: HIGHT, MED y LOW. (High, medium, low).

Modify: The Modify phase contains methods to select, create and transform variables in preparation for data modeling. Once these levels are specified (It is not obligatory but important) they are used as axis to know what actions to set in the following cases (see Table 1. Setting actions by importance and null values table).

According to the amount of null values we manage 3 possible actions to set. In this phase, cleaning process removing extreme values will be implemented, affecting null and discrete numerical values (The options aren't obligatory). Once data were optimized, we move into the next phase, the modeling one.

Model: In the Model phase the focus is on applying various modeling (data mining) techniques on the prepared variables in order to create models that possibly provide the desired outcome.

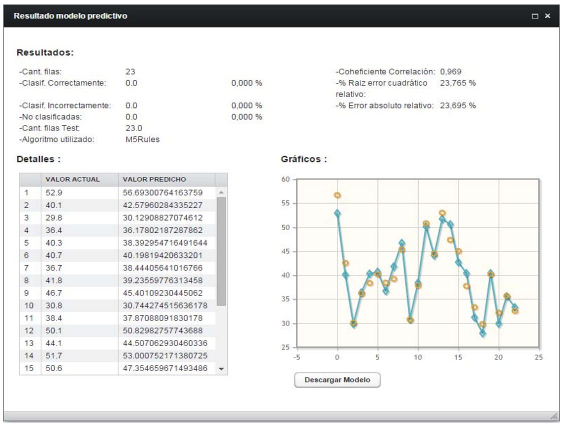
In this phase we can apply "Attribute selection" techniques. Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. Reducing the number of attributes can not only help speeding up runtime with algorithms (some algorithms runtime are quadratic in regards to number of attributes), but also help avoid "burying" the algorithm in a mass of attributes, when only a few are essential for building a good model. Running these algorithms (attribute selection) are also optional.

At the beginning (in simple phase) the user decided whether the set of data was one of the time-series type or cross-sectional, and according to the case a series of predictive algorithms are executed. The algorithms used for time series data are :

1) *Linear Regression*: use linear regression for prediction.
2) *Support Vector Machine*: Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.
3) *Gaussian Processes*: Implements Gaussian processes for regression without hyperparameter-tuning. To make choosing an appropriate noise level easier, this implementation applies normalization/standardization to the target attribute as well as the other attributes. Missing values are replaced by the global mean/mode. Nominal attributes are converted to binary ones
4) *Multilayer Perceptrons*: A Classifier that uses backpropagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the the output nodes become unthresholded linear units).

Algorithms for cross-sectional (transversal) data are:

1) *J48*: for generating a pruned or unpruned C4.5 decision tree.
2) *Nave Bays*: classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data.



Resultado modelo predictivo

Resultados:

-Cant. filas: 23
-Clasif. Correctamente: 0.0 0,000 %
-Clasif. Incorrectamente: 0.0 0,000 %
-No clasificadas: 0.0 0,000 %
-Cant. filas Test: 23.0
-Algoritmo utilizado: M5Rules
-Coheficiente Correlación: 0,969
-% Raiz error cuadrático relativo: 23,765 %
-% Error absoluto relativo: 23,695 %

Detalles :

	VALOR ACTUAL	VALOR PREDICHO
1	52.9	56.69300764163759
2	40.1	42.57960284335227
3	29.8	30.12908827074612
4	36.4	36.17802187287862
5	40.3	38.392954716491644
6	40.7	40.19819420633201
7	36.7	38.44405641016766
8	41.8	39.23559776313458
9	46.7	45.40109230445062
10	30.8	30.744274515636178
11	38.4	37.87088091830178
12	50.1	50.82982757743688
13	44.1	44.507062930460336
14	51.7	53.000752171380725
15	50.6	47.354659671493486

Gráficos :

Figure 2. EMMA results window

3) *Nearest Neighbors*: K-nearest neighbors classifier.
4) *Support Vector Machine*: Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.

Asses: Once the predictive algorithms were executed, outputs are stored and then compared one another, to finally chose the one with the best performance.
Incase of chosing a selection of attributes, there are also 3 algorithms for attribute selection for each data set, the difference is that once chosen the the attribute selection (for each algorithm) the prediction algorithm is executed (all the algorithms with the new resulting data set from the attribute selection), for the process in very extensive, but the processing of data is fast, not taking much time.
Once obtained a model, the user or analyst can keep it in a file, and use it any time it is required an analysis or for making a prediction.
The tool displays the precision indicators for each resulting predictive Model. From here on, it depends on the analyst whether these numbers are the minimal required ones(amount of achievements, mistakes, absolute errors, absolute relative errors, untities, and else, etc).

4. The experiment road map.

4.1. Scope definition

Scope 1 . Assess the efficiency and effectiveness of data mining techniques against techniques used in econometrics.
Scope 2 . Evaluate the level of technical training required for economic analysis using a tool based on data mining and other tools based on econometric analysis.

4.2. Context selection

Experimental objects consist of a tool based on data mining (EMMA), and another tool that subjects already have knowledge and experience when making use econometric analysis, the tool is Eviews(version 8).
Two study case was defined, used as evaluation tests with the third year students of economics. The cases are the "Chicken consumption" and "Wire consumption", so it is considered that the cases are not very complex but are for serious studies.
The case study "Chicken consumption" has 23 samples and "Wire consumption" has 16.
The 31"%" of the samples were used as tests, the remaining 69"%" of the samples were used to generate the predictive model(training). The project will be implemented in a "off - line" context where subjects are employees of the same organization (professional subjects). These subjects are now part of the Department of Economic research of a company dedicated to finances and normally works to study micro and macro quantitative indicators that come from continuously surveyed businesses, shops and other public bodies.
These data are processed and published later. The subjects never used datamining [1]

TABLE 2. EXPERMINT GQM TABLE

Object of study:	Predictive model based on data mining techniques.
Purpose:	Assess the efficiency and effectiveness of data mining techniques against the techniques used in econometrics.
Quality focus:	Assess the efficiency and effectiveness of data mining techniques against the techniques used in econometrics.
Perspective:	Economic analyst.
Context:	Department of Economic Research of a company dedicated to finances.
Objects:	Tool based on data mining (EMMA). Tool used for the econometric analysis (Eviews).
Subjects:	Classified by level of expertise in the domain: They can be classified to HIGH level (Senior Econ. Analysts) , Medium (Junior economic analysts) , Low (Students). According academic training (Economic Sciences): High (PhD , Masters) , Mid (College) , Low (Student).

4.3. Subjects selection

In the experiment they involved 5 employees and 3 college interns (between second and third year). All subjects, career (students and professionals), should be economics.
The subjects were classified based on:

- Econometrics Knowledge.
  - Specialization in econometrics (High).
  - Obtained at the university(Mid)
  - No knowledge (No).

[1] Goal/Question/Metric layout, used to specify the Experimet Goals

- Academic level.
  - Master - Phd (High).
  - College(Mid)
  - Student (Low).
- Time experience (years)

The subjects were classified using preliminary surveys.
Subjects will receive training for the data mining based tool, once in conditions and perform the process to arrive at a predictive model based on KDD. Subjects already have experience and knowledge in Eviews.
If each subject knew ecometrics, they would applie both treatments (Otherwise they, can only manipulate data mining based tool). In both cases, time for applying both methods will be measured, and the final results based on the use of econometric model will be considered by the expert judgment of the subject.

TABLE 3. SUBJECT CLASSIFICATION BY KNOWLEDGE AND EXPERIENCE IN ECONOMETRICS

Subjects	Academic Level	Knowledge	Experience (Years)
1	Master (High)	Hight	2
2	College (Mid)	Mid	0
3	College (Mid)	Mid	1
4	College (Mid)	Mid	1
5	Master (High)	Mid	2
6	Student (Low)	No	0
7	Student (Low)	No	0
8	Student (Low)	No	0

The subject is given to choose the study case “Chicken consumption” or “Wire consumption” whatever he feels comfortable with.

4.4. Experiment Design

The type of design will be a factor with two treatments [5]. The factor would reach a predictive model of the variables mentioned in the study case and the treatment would be the use of a KDD process or econometric analysis.

4.5. Hypothesis

As hypothesis we compare the effectiveness of both techniques in the percentage error introduced both real data and efficiency with the time it takes to reach a solution .
The levels of the independent variable are both technical (Econometrics and Datamining) and the dependent variables are the percentage of error (Relative absolute error) and the time it takes to generate a predictive model (Relative Absolute Error(”%”) / Time(minutes)).

- Effectiveness:
  - Let μE Econometrics Resultant Model Relative Absolute Error (RAE)
  - Let μD Datamining Resultant Model Relative Absolute Error (RAE)
- Efficiency:
  - Let τE Econometrics Resultant Model Relative Absolute Error (RAE)/ Time(minutes)
  - Let τD Datamining Resultant Model Relative Absolute Error (RAE)/ Time(minutes)
- Then:[2]
  - $H1_0 \rightarrow \mu E = \mu D$
  - $H1_1 \rightarrow \mu E \neq \mu D$
  - $H2_0 \rightarrow \tau E = \tau D$
  - $H2_1 \rightarrow \tau E \neq \tau D$

4.6. Instrumentation

In this task, the instrumental objects consisted in the appropiate functioning of the data mining based tool and the correct installation of Eviews. The study case was well and clearly specified by the correct understanding. To record everything that has been done by users, Camtasia software was used. It would record in video format all activity performed by the subjects for further analysis. A brief presentation slides was prepared to train users on data mining tool. It was designed a short questionnaire before classify the level of knowledge and experience of the subject in econometrics.

4.7. Preparation

The PCs of users were inspected to verify they have installed the required software (Eviews , camtasia , EMMA) and proper operation there fore.

4.8. Execution

The experiment was conducted after a questionnaire and the short training were given, in the same places where subjects performs their daily working activity and finally proceeds to start recording the subject’s activity through Camtasia and will not be interrupted under any circumstances until after the claims that the econometric model was completed , according to expert judgment in a good way and with appropriate coefficients.
The subjects deliver the results obtained by Eviews and proceeds to use datamining tool (EMMA).
The experiment ends when the subject reaches the predictive model based on data mining, Camtasia recording is cut and all files are stored in specific directories for each subject and so, and then the results are analyzed.
Subjects had not been informed about the objectives of the experiment until the experiment was completed.

4.9. Results analysis and interpretation

According to the results of predictive models based on econometric analysis (See Table:4) and dataminig based

[2] H1 and H2 are the hypothesis, when equals (=) is the null hypothesis that may be rejected

TABLE 4. Econometrics Modeling Results

Subjects	Case	**Model RAE (%)**	**Time (min)**	**RAE/Time (min)**
1	Chicken	18.68%	09:45	0.0000252
2	Wires	148.64%	15:00	0.0000149
3	Wires	148.64%	02:46	0.0000027
4	Chicken	94.23%	08:20	0.0000043
5	Chicken	18.68%	39:00	0.0001007
6	Wires	N/R	N/R	N/R
7	Wires	N/R	N/R	N/R
8	Chicken	N/R	N/R	N/R

TABLE 5. Datamining modeling results

Subjects	Case	**Model RAE (%)**	**Time (min)**	**RAE/Time (min)**
1	Chicken	18.02%	00:54	0.2205
2	Wires	108.32%	00:40	1.4379
3	Wires	108.32%	00:50	1.1503
4	Chicken	18.02%	00:45	0.2646
5	Chicken	18.02%	01:54	0.1044
6	Wires	108.32%	01:02	0.9277
7	Wires	108.32%	00:52	1.1061
8	Chicken	18.02%	01:10	0.1701

(See Table:5), the tables 6 and 7 and shows the corresponding descriptive statistics results of the two measurements; Effectiveness and Efficiency. It Presents fields like N, represents the number of subjects, the median, the mean and standart desviation. Analyzing it, it can be noticed there is better performance of the predictive model based on data mining.

In econometrics analysis, comparing the subjects experience and knowledge (See Table:3), a strong relationship to the subject's experience regarding the performance of the predictive model is observed, however, in the KDD process does not considering students could perform the test with data mining and could not with econometrics. [3]

TABLE 6. Efficiency descriptive Statistics

		Efficiency		
Treatments	N	Mean	Median	DS
E.A	5	0.000029	0.000014	0.00004
D.M	8	0.67274	0.59619	0.53624847

TABLE 7. Effectiveness descriptive Statistics

		Effectiveness		
Treatments	N	Mean	Median	DS
E.A	5	85.78%	94.23%	0.65
D.M	8	63.17%	63.17%	0.48

[3] E.A acronym of Econometric Analysis, D.M Datamining

5. Conclusion.

The goals in statistics are to use data to predict and to get information about the underlying data mechanism. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data. To make myposition clear, I am not against data models per se. In some situations they are the most appropriate wayto solve the problem. But the emphasis needs to be on the problem and on the data. [6]

According to the empirical economic analysis [1], note the dominant role of economic theory and the modest role of the sample data in this procedure (develop a predictive model). Furthermore the process is very dependent on the experience of economic analyst and level of technical training and the model variables are adjusted based on the economic theory that the analyst knows.

A data based search for a good model specification is rejected by the traditional approach in Econometrics.
One starts with the favoured model, which is usually a relatively simple theoretical model, next repair and extend it to uphold the favoured hypothesis if any data problems are encountered. Researchers would start from their favoured theoretical model and "patch" the model, for example by including additional variables, if the data didn't agree.
[7] In datamining the data is used extensively to search for a good model and the models are "atheoretical" in the sense that received economic theory plays a minor role in the analysis. Spanos [8] uses the vivid analogy of shooting at a blank wall and then drawing a bull's eye around the bullet hole: the probability of the shot being in the bull's eye is equal to one. The proper way according to the classical view is to specify the model (i.e. drawing the target) before looking at the data (seeing where the bullet hole is).

Just a centralized data approach is proposed, using data mining techniques to analyze them and try putting aside the traditional approach in Econometrics, isolating the high dependence on knowledge and experience required in ecometrics for a good predictive model.

References

[1] J. Wooldridge, *Introductory econometrics: A modern approach.* Cengage Learning, 2012.

[2] J. Han, M. Kamber, and J. Pei, *Data mining, southeast asia edition: Concepts and techniques.* Morgan kaufmann, 2006.

[3] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[4] A. I. R. L. Azevedo, "Kdd, semma and crisp-dm: a parallel overview," 2008.

[5] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering.* Springer Science & Business Media, 2012.

[6] L. Breiman, "Statistical modeling: The two cultures," *Quality control and applied statistics*, vol. 48, no. 1, pp. 81–82, 2003.

[7] A. Feelders, "Data mining in economic science," *Dealing with the data flood*, pp. 166–175, 2002.

[8] A. Spanos, "Revisiting data mining: hunting with or without a license," *Journal of Economic Methodology*, vol. 7, no. 2, pp. 231–264, 2000.