Neuro-Fuzzy Data Mining Mexico's Economic Data

Gustavo Becerra-Gaviño and Liliana Ibeth Barbosa-Santillán

Universidad de Guadalajara, Av Enrique Daz de León Sur,
Americana, Guadalajara, Jalisco, Mexico

Abstract. Given the increase of data being collected, there is a need to explore the use of tools to automate the recognition and extraction of patterns within some targeted data. The present work explores the use of a neuro-fuzzy classifier for the multi-factor productivity from the manufacturing sector in the Mexican economy. The chosen data set contains the time series for the variables: Sale Value of products, Wages, Work Force, Days Worked, and Hours Worked. The data is taken from the Banco de Información Económica at the Instituto Nacional de Estadística y Geografía. The neuro-fuzzy system is implemented on top of the Neuroph library extending on the ideas behind the Neuro-Fuzzy Reasoner. A sample run tends to assign the same values given by a visual inspection.

Keywords: Neuro-Fuzzy, Data Mining, Multi-Factor Productivity, Supervised Learning, Multi-Layer Perceptron, Fuzzy Logic.

1 Introduction

With the advent of the Internet also came the opportunity to share information and make it more readily accessible. In Mexico the Instituto Nacional de Estadística y Geografía (INEGI) is in charge of gathering information about the country. The INEGI maintains the Banco de Informacion Económica (BIE). The BIE is accessible through an interface available at the INEGI website [4]. Given that the information is readily available the opportunity presents itself to contribute to its understanding by providing tools to analyze it.

The INEGI BIE web interface readily provides, as a simple analytical tool, a graphing utility to visualize the data. The following chart is generated using the utility provided in the BIE web interface for selected time series.

In addition to the graphing utility provided in the BIE web interface, there are other analytic tools provided in the Analysis Lab. The INEGI provides access to analytic tools such as Excel, STATA, and SPSS. IBM SPSS provides the package SPSS Neural Networks [17] as an extension to the main SPSS statistics software package. However, the documentation for these products doesn't mention any implementation of a neuro-fuzzy system geared for analytics. The public information about the tools used to analyze data at the INEGI tends to indicate that neuro-fuzzy systems are not being used in such institution. Therefore, the

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 645–657, 2014.

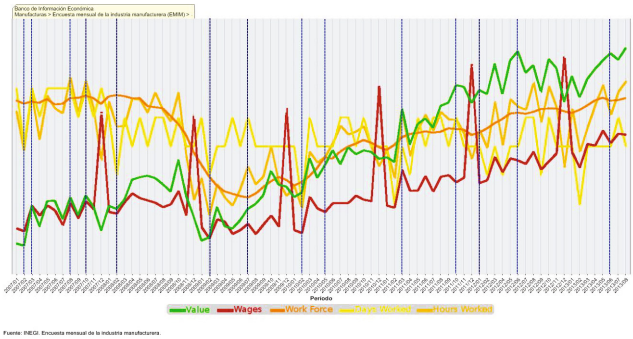

Fig. 1. Input Data

present work will exhibit how a neuro-fuzzy system facilitates the classification of productivity in the manufacturing sector for a given month using the data provided at the INEGI. The reasons why it is important to measure productivity are that it is used for tracing technology, identifying the efficiency of a given production system, and indexing the standard of living among others [22]. The productivity discussed in the present work refers to Mexico's productivity in the manufacturing sector. However, productivity measurement for smaller economies, for example a company, is similarly used for strategic planning in operations management [24].

2 Giants' Work

Even in Greek mythology the existence of intelligent mechanical beings captured human imagination as far as to describe a mythical bronze being, Talos. In this current age, the efforts to understand how intelligence exists has provided us with useful tools that help us make better sense of the phenomena around us. The introduction of a mathematical model for the biological neuron gave way to having artificial neural networks capable of learning from the data being processed. Fuzzy logic expresses the linguistic values for variables. The combination of both neural networks and fuzzy logic provides us with tools that learn and express values in a more human-like language. The use of neuro-fuzzy systems in data mining automates the analysis of data.

2.1 Neural Networks

In 1943 Warren S. McCulloch and Walter H. Pitts [11] introduced the mathematical model for the artificial neuron. This is a very simplified model visualized in Figure 2. It basically consists of a set of weighted inputs. Those inputs are aggregated. The aggregated value is then passed through the activation function and an output is produced. This model by itself does not have much usefulness. However, it is the building block for more complex systems like the multilayer perceptron represented in Figure 3.

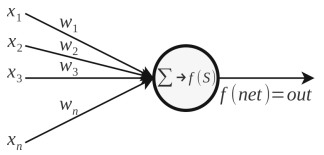

Fig. 2. Artificial Neuron

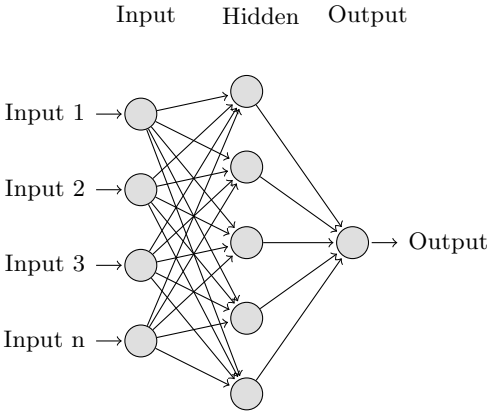

Fig. 3. Muiltilayer Perceptron

In 1958 F. Rosenblatt published the mathematical model for the Perceptron [20, 21]. He took the idea of an artificial neuron a bit further by including excitatory inputs, inhibitory inputs, and feedback signals [21]. One of the ways a neural network stores information (learns) is through adjusting the input weights based on the feedback signals. Later in 1969 Minsky and Papert [12] proved that the Perceptron could not learn the XOR function [10]. This problem slowed down the advancement in artificial intelligence until the multilayer perceptron was used to find a solution.

2.2 Fuzzy Logic

The notion of fuzzy sets was introduce by Zadeh [28] in 1965. In fuzzy sets the values are expressed as a degree of membership to the elements of the set. Consider the following membership function in Figure 4 for the variable Value:

In this membership function, also called characteristic function, the fuzzy set is ***{Low,Medium,High}***. The fuzzy value of ***Medium*** has a degree of membership or truth of 0 at 305 increasing to 1 at 327; from 327 to 375 it has a degree of 1; then it decreases from 1 at 375 to 0 at 397. The fuzzy value of ***High*** has a value of 0 at 375 increasing to 1 at 397 and staying at 1 from there on. This example illustrates the fact that a given crisp value for a variable can be a member of two fuzzy sets when the variable goes through the membership functions and gets fuzzified.

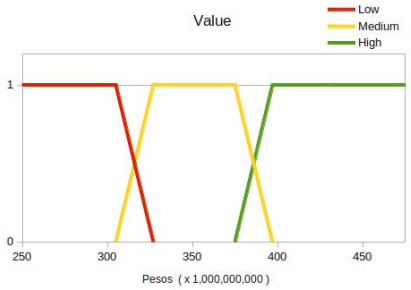

Fig. 4. MF: Value

```
1 IF value IS high AND wages IS high THEN productivity IS low;
2 IF value IS high AND wages IS medium THEN productivity IS medium;
3 IF value IS high AND wages IS low THEN productivity IS high;
```

Fig. 5. Fuzzy Rules Example

At the same time that a variable looses information through the process of fuzzification, it gains on flexibility, tolerance, and expressiveness [9]. Fuzzy logic uses IF-THEN constructs to express the relations between fuzzy variables. For example:

2.3 Neuro-Fuzzy Systems

As early as 1985 Keller, Hunt, and Douglas [7] researched the idea of combining the fuzzy logic and the perceptron. Their efforts aimed at alleviating the problem that the crisp perceptrons had on converging in the case where the classes in a hyperplane were not linearly separable. Later came Sankar and Sushimita [18] who introduced the use of fuzzy membership functions along with a supervised learning perceptron for classification. The fuzzy perceptron is a 3-layered network intended to include knowledge defined in the rules it is implementing. In machine learning, supervised learning refers to the process of feeding knowledge previously defined into the system as opposed to unsupervised learning where the system discovers hidden information as it processes data.

The ANFIS [5], Adaptive-Network-based Fuzzy Inference System, is a multilayer fuzzy perceptron that uses predefined human knowledge provided in the fuzzy IF-THEN rules. It also combines adaptive neurons which are neurons with specific parameters that are updated to achieve a desired input-output mapping as the training set is processed.

NEFLASS [14–16] is a 3-layered feedforward fuzzy perceptron. It is intended to determine the correct class for a given set of values from the input variables. The output neurons represent the fuzzy set for the variable being classified. The NEFCLASS was used as the inspiration for the Neuro-Fuzzy Reasoner.

The neuro-fuzzy reasoner [23] is based on the NEFCLASS. It is a 4-layered feedforward fuzzy perceptron. However, it differs from NEFCLASS in that the membership functions are not modified throughout its execution. It was originally designed to classify how good a class was based on the score a student had on an exam and how quickly the student could answer the given exam. The present work takes on the main ideas from this model and applies them for the classification of Productivity based on five variables: Value, Wages, Workforce, Days Worked, and Hours Worked.

2.4 Data Mining

We live in an age when information about our activities is being collected constantly. The amount of data is so vast and varied that the traditional tools and ways of analyzing such data are rapidly being surpassed in their capacity. It has simply become unfeasible to meticulously look for patterns hidden within the mountains of data using traditional statistics and specialized personnel [27]. There comes the need to devise artifacts and systems to automate the extraction of hidden patterns within all the information there is available to us.

With all the tools available for data mining, sometimes data mining may just be connecting the output of one model to another using graphical tools [8]. Even so, the idea remains the same. Data mining is about discovering new information hidden within the data. It is the analysis phase within the process of Knowledge Discovery in Databases (KDD). To that end, computer scientists have devised and the computer industry has implemented various tools geared to ease the effort in analyzing data [6, 27]. In the present work the focus is placed on neuro-fuzzy systems applied to data mining.

2.5 Neuro-Fuzzy Systems in Data Mining

The main uses of neuro-fuzzy systems in analytics are for clustering, regression and classification [3, 13]. In clustering, the data is arranged in groups of similar items as it is being processed. Therefore clustering is used to discover and learn unsuspected associations in the data. Clustering uses mainly unsupervised learning. The goal of a regression is to approximate a relation between two sets X and Y by mapping items between X and Y. Regression uses generally supervised learning. Classification intends to place items in a data set within a predefined class based on an assessment of its features. Classification uses supervised learning [13, 25]. The present work implements a neuro-fuzzy classifier for the variable "Productivity".

3 Methodology

The present work requires an experiment in system design. The idea is to explore the feasibility of using a neuro-fuzzy system to classify the productivity from the manufacturing sector in Mexico's economy to facilitate its interpretation.

The system will be implemented on top of the Neuroph [1] framework using the computer language Java. The concrete implementation will consist of three classes: NeuroFuzzyClassifier, NFCFactory, and MXMiner. The NueroFuzzyClassifier class will encapsulate a flexible implementation for a neuro-fuzzy classifier. The NFCFactory class will be used to produce an instance of the NeuroFuzzyClassifier based on a properties file holding the features defined for the neuro-fuzzy classifier. The MXMiner class will perform the loading of the training set, training the system, and imputing the data into the system.

The experimental evaluation will be performed with the data retrived from the BIE [4]. The results obtained will tell us how good productivity was for a given month. The execution of the neuro-fuzzy system will conclude the first cycle in experimentation in system design.

4 Neuro-Fuzzy Classifying Productivity

The most basic definition of Productivity [19] in a production system is:

$$Productivity = \frac{Output}{Input} \tag{1}$$

This is a simplified definition. Multifactor productivity ivolves one output and many inputs [22, 24]. Still, for the present work the information that is necessary to understand about productivity is that by definition, productivity is directly proportional to the output and inversely proportional to the inputs in a production system. Therefore a high output tends to improve productivity and a high input tends to decrease productivity.

4.1 Variables

The target is a set of five time series belonging to the monthly survey for the manufacturing sector. The data for these variables is available at the BIE [4] referencing their Spanish description.

1. Valor de ventas de los productos elaborados
2. Remuneraciones totales
3. Personal ocupado total
4. Días trabajados
5. Total de horas trabajadas

For ease of reference, the following corresponding variables will be used for the rest of this writing.

1. Value
2. Wages
3. Workforce
4. Days Worked
5. Hours Worked

The downloaded series have a range of January 2007 to June 2013 and contain the data for the variables ***Value***, ***Wages***, ***Work Force***, ***Days Worked***, and ***Hours Worked***. The present work uses a neuro-fuzzy system to tell (classify) if productivity is low, medium, or high for a given month based on these variables. In manufacturing, these variables can be classified as follows:

$$\text{Inputs} \begin{cases} \text{Wages} \\ \text{Work Force} \\ \text{Days Worked} \\ \text{Hours Worked} \end{cases} \qquad \text{Outputs} \mid \text{Value}$$

Fig. 6. Variables used as Inputs and Outputs for productivity

Figure 1 presents the visualization of the inputs and output for the manufacturing sector in the Mexican economy. The question about productivity in such system is then: When is productivity low, medium, or high? The values low, medium, and high in this question are the values for the fuzzy set for Productivity. The fuzzy sets for the other variables are defined similarly. Thus the fuzzy sets for the system are:

$$\begin{aligned} \text{Productivity} &= \{\text{low}, \text{medium}, \text{high}\} \\ \text{Wages} &= \{\text{low}, \text{medium}, \text{high}\} \\ \text{Days Worked} &= \{\text{low}, \text{medium}, \text{high}\} \end{aligned} \qquad \begin{aligned} \text{Value} &= \{\text{low}, \text{medium}, \text{high}\} \\ \text{Work Force} &= \{\text{low}, \text{medium}, \text{high}\} \\ \text{Hours Worked} &= \{\text{low}, \text{medium}, \text{high}\} \end{aligned}$$

Fig. 7. Fuzzy Sets

Given that the delimiters for the fuzzy sets for the fuzzy functions are defined by the available human knowledge, they are roughly based on the statistical quartiles for the time series. Figures 8 through 12 present the membership functions for the system.

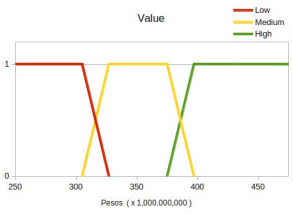

Fig. 8. MF: Value

Fig. 9. MF: Wages

The trapezoid function is used for ease of implementation only four points are needed as the delimiters for the functions. The data for the variable ***Wages*** for

Fig. 10. MF: Workforce

Fig. 11. MF: Days Worked

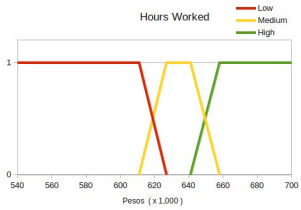

Fig. 12. MF: Hours Worked

each month varies very slightly with the exception of the month of December. Thus the delimiters for the membership functions for ***Wages*** in Figure 10 are closer in shape compared to the rest with the exception of the membership functions for days worked. As is evident, the membership functions for the number of days worked in Figure 11 take up a rectangular shape. This is due to the fact that the days worked for a given month is an integer. The range for ***Days Worked*** in the data set is between 23 and 28 with most of the months having 26 days worked.

4.2 IF-THEN Constructs

As already established in equation 1, productivity is directly proportional to the outputs and inversely proportional to the inputs in a production system. Therefore, based on that knowledge the antecedent parts of the fuzzy rules are constructed as shown in Figure 13. The consequent part of the rules will be learned by the neuro-fuzzy classifier based on the desired output in the training set.

4.3 Neuro-Fuzzy System Design

The Mexican manufacturing production system uses the input variables of Wages, Work Force, Days Worked, and Hours Worked; the output is the variable Value. The neuro-fuzzy classifier is a feed forward neural network. The first layer comprises five input neurons one for each variable in the production system. The neurons in the second layer represent the corresponding variable's fuzzy sets.

```
IF value IS high AND wages IS medium;\
IF value IS high AND wages IS low;\
IF value IS medium AND wages IS high;\
IF value IS medium AND wages IS medium;\
IF value IS medium AND wages IS low;\
IF value IS low AND wages IS high;\
IF value IS low AND wages IS medium;\
IF value IS low AND wages IS low;\
```

Fig. 13. Fuzzy Rules Antecedents

The rule antecedents are implemented through the connections from the fuzzification neurons to the neurons in the third layer. The fourth layer is the output layer representing the fuzzy sets for Productivity. The neural network output is the classification of productivity based on the neural network input variables. Figure 14 presents a graphical representation of the neuro-fuzzy classifier design used in the current work.

5 MXMiner Neuro-Fuzzy Classifier

5.1 Implementation

The current work researches the implementation of a neuro-fuzzy classifier for productivity expanding on the idea exposed in the Neuro-fuzzy Reasoner [23]. However, since the Neuro-Fuzzy Reasoner constructed in [23] is particular to using two variables, Score and Time, it is necessary to implement a more flexible neuro-fuzzy classifier using the artificial intelligence Java library Neuroph [1].

5.2 Execution

The neuro-fuzzy classifier is trained using a set with chosen input and output values. The blue lines in figure 1 intersect the data points used in the training set. The training set includes the desired output for a given training input set. The network is set to use backpropagation [2] with a SigmoidDeltaRule [26] learning function for correcting the output error as it processes the learning set. The input data to be analyzed is fed through a CSV file.

5.3 Results

An execution of the Neuro-Fuzzy Classifier for Productivity based on the variables of Value, Wages, Workforce, Days Worked, Hours Worked yields the output summarized in the following figure:

In the chart of figure 15 the rows of numbers represent how strong productivity belongs to each of the elements in the productivity fuzzy set. From top to bottom, the first row represents how strongly the productivity belongs to the fuzzy value of ***High*** the second row is for ***Medium*** and the bottom row is for ***Low***. Therefore, the output suggests that productivity was between medium and high for the months of 2007/01 to 2008/02. In contrast, the months of 2013/03

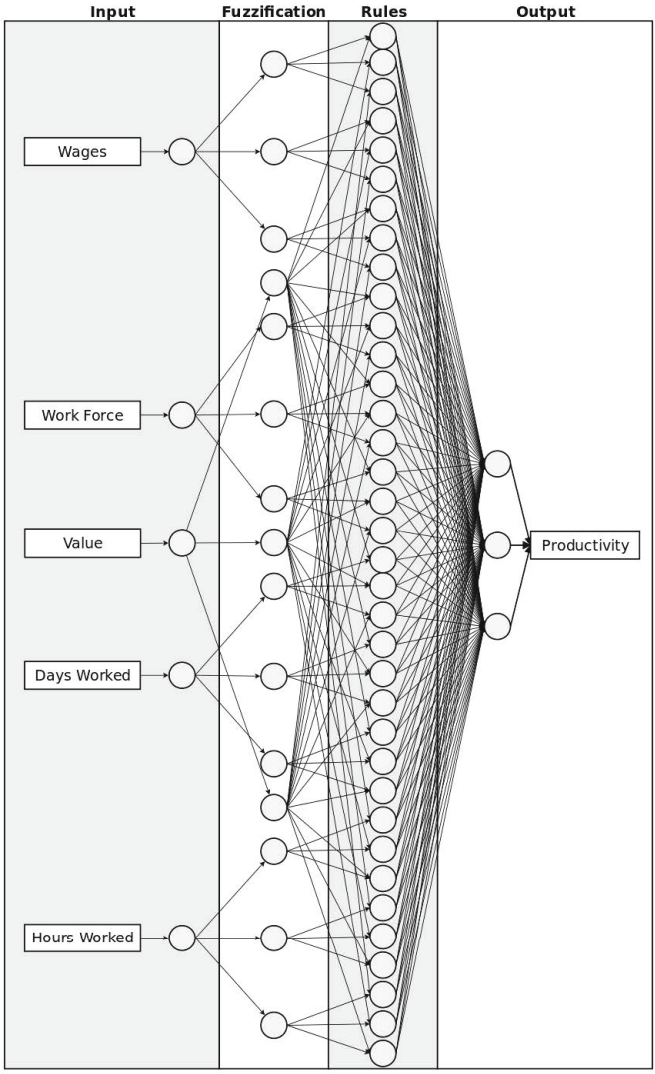

Fig. 14. Productivity Neuro-Fuzzy Classifier Design

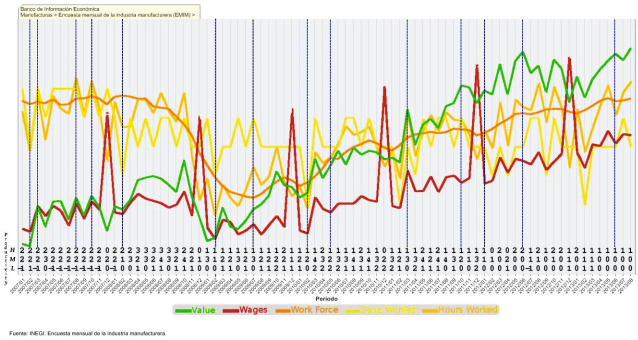

Fig. 15. Sample Run Results

to 2013/08 were high. Noteworthy is the output for the months of 2012/05 and 2012/06 that suggests a strongly high productivity for that period. This system output is consistent with the high ratio of outputs over inputs for the manufacturing sector for the same period. Overall the output of the neuro-fuzzy classifier constructed for the present work is consistent with the expected results based on a visual inspection on the graphed data. Thus it facilitates the classification of productivity for a given month However, the system can be improved in several ways detailed in section 6.2.

6 Conclusions

In the field of information technology there is always a problem to solve and many ways to solve it. In the present work the implementation for a neuro-fuzzy classifier for ***Productivity*** to classify it as ***High***, ***Medium***, and ***Low***is presented.

6.1 Achievements

The two main contributions in the present work are the NeuroFuzzyClassifier and the fuzzy classification of manufacturing productivity. The NeuroFuzzyClassifier is an implementation of a feedforward neuro-fuzzy classifier expanding on the work presented in the neuro-fuzzy reasoner work [23]. Compared to the neuro-fuzzy reasoner the NeuroFuzzyClassifier increases flexibility to construct nuero-fuzzy networks and allows for a wider range of classification applications. The proposed classification of productivity based on many imputs is a subject to be researched further. With that note, a neuro-fuzzy system is presented as an approach to classify productivity in complex multi-factor systems.

6.2 Further Work

One of the ways to infuse pre-existing knowledge into a neuro-fuzzy system is through the definition of the fuzzy rules. Better constructed fuzzy rules will

yield better results. To that end, an individual with vast understanding of the intricacies of the system to be automated will define better fuzzy rules and contribute to its accuracy. On the same page as with fuzzy rules, the chosen training set is another way of embedding knowledge into the neuro-fuzzy system. The desired output will be used to learn the consequent part of the fuzzy rules. Thus if the training set contains precise information, the system will be able to better cope with the data it analyzes. The available data set has only 81 items. A subset of that data of 14 items were used for training. If more data is made available a bigger training set can be selected and thus better train the neuro-fuzzy classifier.

In addition, in its present state, the neuro-fuzzy classifier outputs numbers less than 0 and greater than 1. However the degree of membership to a fuzzy element is a value between 0 and 1 inclusive [28]. Therefore it is necessary to find a way to normalize the output.

6.3 Closing Remarks

Throughout the present work, the goal is to explore the use of the existing tools geared towards analytics. Neuro-fuzzy systems are the focus on this work because of their adaptability and learning capabilities. Neuro-fuzzy systems present a good opportunity to analyze data using the way humans express quantities and use the learning capacity of neural networks to store information and use that information to adapt to their purposes. Given how flexible neuro-fuzzy systems are, there may just be a sea of applications waiting to be discovered. For the time being, an application of the neuro-fuzzy system to facilitate the classification of productivity has been presented.

References

[1] Contributors, Neuroph. (2013), `http://neuroph.sourceforge.net/index.html`
[2] Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: International Joint Conference on Neural Networks, IJCNN, pp. 593–605. IEEE (1989)
[3] Hüllermeier, E.: Fuzzy methods in machine learning and data mining: Status and prospects. Fuzzy Sets and Systems 156(3), 387–406 (2005)
[4] INEGI, Banco de información económica (2013), `http://www.inegi.org.mx/sistemas/bie/`
[5] Jang, J.S.: Anfis: Adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics 23(3), 665–685 (1993)
[6] Kantardzic, M.: Data mining: concepts, models, methods, and algorithms. John Wiley & Sons (2011)
[7] Keller, J.M., Hunt, D.J.: Incorporating fuzzy membership functions into the perceptron algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 693–699 (1985)
[8] Khabaza, T.: Hard hats for data miners: Myths and pitfalls of data mining. Business intelligence, data warehousing and analytics editorial from DM Review (2005)
[9] Leboeuf Pasquier, J.: Programació Basada en Lógica Difusa. Amate Editorial (2006a)

[10] Leboeuf Pasquier, J.: Programació Basada en Redes Neuronales. Amate Editorial (2006b)
[11] McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics 5(4), 115–133 (1943)
[12] Minsky, M., Seymour, P.: Perceptrons. MIT Press (1969)
[13] Mitra, S., Pal, S.K., Mitra, P.: Data mining in soft computing framework: A survey. IEEE Transactions on Neural Networks 13(1), 3–14 (2002)
[14] Nauck, D., Kruse, R.: Nefclassmdash; a neuro-fuzzy approach for the classification of data. In: Proceedings of the 1995 ACM Symposium on Applied Computing, pp. 461–465. ACM (1995)
[15] Nauck, D., Nauck, U., Kruse, R.: Generating classification rules with the neuro-fuzzy system nefclass. In: 1996 Biennial Conference of the North American Fuzzy Information Processing Society, NAFIPS, pp. 466–470. IEEE (1996)
[16] Nauck, D.D.: Fuzzy data analysis with nefclass. In: Joint 9th IFSA World Congress and 20th NAFIPS International Conference, vol. 3, pp. 1413–1418. IEEE (2001)
[17] Norušis, M.: IBM SPSS Neural Networks 20. IBM (2011)
[18] Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. IEEE Transactions on Neural Networks 3(5), 683–697 (1992)
[19] Rogers, M.: The definition and measurement of productivity. Melbourne Institute of Applied Economic and Social Research (1998)
[20] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 65(6), 386–408 (1958)
[21] Rosenblatt, F.: Two theorems of statistical separability in the perceptron. United States Department of Commerce (1958b)
[22] Schreyer, P., Pilat, D.: Measuring productivity. OECD Economic Studies 33(2), 127–170 (2001)
[23] Sevarac, Z.: Neuro fuzzy reasoner for student modeling. In: Sixth International Conference on Advanced Learning Technologies, pp. 740–744. IEEE (2006)
[24] Stevenson, W.J., Hojati, M.: Operations management, vol. 8. McGraw-Hill/Irwin, Boston (2007)
[25] Vieira, J., Dias, F.M., Mota, A.: Neuro-fuzzy systems: a survey. In: 5th WSEAS NNA International Conference on Neural Networks and Applications, Udine, Italia (2004)
[26] Widrow, B., Lehr, M.A.: 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. Proceedings of the IEEE 78(9), 1415–1442 (1990)
[27] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
[28] Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)