Are statistics and machine learning enough to make predictions and forecasts?*

Antonio Lorenzo[1,2] [0000-0003-0752-6980] and José A. Olivas[2] [0000-0003-4172-4729]

[1] Coordinator of the Department of Business Intelligence, Castilla La Mancha Government, Toledo, Spain.
alorenzo@jccm.es

[2] SMILe (**S**oft **M**anagement of **I**nternet and **L**earning). Information Technologies and Systems Institute, University of Castilla La Mancha, Ciudad Real, Spain.
JoseAngel.Olivas@uclm.es

Abstract. Currently the techniques used to predict the future are statistical and machine learning techniques. The first continues the trend of historical data. The second learns from previous cases training. Both use historical information but do not take in mind key factors that can make the final result change. A knowledge-based framework is presented that allows predictions of some kind of events to be made using artificial intelligence techniques. This requires an expert to enter the key factors that can change the trend of historical data into the system. The current framework has been applied prior to happening to two use cases, obtaining good preliminary results, in the framework of the developing of a PhD Thesis.

Keywords: Forecast, Trend, Prediction, Statistics, Machine Learning, Expert Knowledge.

1 Introduction.

Throughout the history of humanity, knowing what will happen in the future has been a constant. Knowing what the weather will be like tomorrow, how the stock market will behave or if your football team will win the next game, are questions that are asked daily. Statistical techniques have traditionally been used to predict the future. Lately machine learning techniques are being used. But these are not always enough because they do not take in mind some key factors that influence the final result. This doctoral thesis is born with the motivation to establish a framework to predict some types of events.

* This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET: TIN2016-76843-C4-2-R (AEI/FEDER, UE).

2 Prediction: Breaking the trend…

Reviewing the literature to know what terms are used to know future events, we have observed that two terms are commonly used to know the future: forecast and prediction. In these papers, both terms are used interchangeably, but they are not exactly the same. After this revision, the conclusion obtained is that both terms are used ambiguously, but in general, "forecast" refers to the analysis of data from a time series following the trend, and "prediction", as the forecast plus other factors that can change the trend (Selvin et al. [1], Minh et al [2], Sezer et al. [3], Stuparu et al. [4]). In general, all forecasts are predictions, but not all predictions are forecasts.

The next step was to search the literature to find what techniques were used in forecasting and/or prediction. Traditional techniques use statistics with analysis of times series and regression Atsalakis et al. [5]. The newest techniques use artificial intelligence like machine learning (Garcés Ruiz et al. [6], Vaidehi, V. et at [7], Atsalakis et al. [8], Nojek, S. et al. [9]). After analyzing some papers, it is concluded that statistical techniques work well when the trend of historical data is maintained, and machine learning algorithms work correctly when the model has been previously trained with the type of case to be predicted. When the above conditions are not met, current prediction techniques are not enough.

We try to define a methodology and apply it to two case studies, based on knowledge, which allows predictions to be made in some cases because we want to improve the results of current techniques. The methodology starts from the analysis with the data of the current forecasting and/or prediction techniques, adding expert knowledge that indicates which elements, ideas or aspects may have a determining role in the result. For this, we are looking for an expert in the field been able to identify the key elements that have influenced the result. It is about looking for previous cases that are similar to the future event, establishing an analogy between both events. If previous events, some premises produced some results, in future events, we can establish that if they are part of the premises, they will also be part of the results.

For this, artificial intelligence techniques such as heuristics, rule-based systems, learning by analogy and case-based reasoning (CBR) are used. Analysis of historical and current data can only determine "what has happened" and "why it has happened". If you want to determine "what will happen", additional descriptive knowledge based on heuristics should be applied to the descriptive analysis of the data.

The methodology is not applicable in all scenarios. There are four scenarios to determine the future: the first is certain (practically 100% of the information is available). The second is forecasting (there is a linear relationship between the historical data and the results establishing a projection), the third is random (the results do not depend only on the historical data). The fourth scenario is prediction (much of the information is unknown and there is no linear relationship between the historical data and the outcome). The methodology is developed in this last scenario and is defined to predict the result of the event, not when the event will happen

The prediction is complex because the variables that form it are unknown, as well as the relationships between them. Making a prediction is difficult. The work done to date has consisted of defining a ten-step prediction methodology and it will be successful if

the methodology improves the results of current forecasting and/or prediction techniques. To simplify complexity and to be able to work with the problem of "prediction", knowledge must be represented and uncertainty must be managed. It is necessary to represent knowledge to identify what concepts and strategies have been used successfully in previous use cases to be formulated at a higher level of abstraction and can be used in other analogue use cases. Knowledge has an apparent simplicity for humans, but it is very complex to manage it artificially. All representation is an imperfect approximation of reality. There is no way of representing knowledge as rich as natural language. Knowledge in humans is not structured; instead the representation of knowledge in machines needs to be structured. To represent knowledge, logic, rules or semantic networks can be used [10]. None of them is complete. The representation of knowledge must allow identifying, model, representing and using that knowledge. Selecting the way of representing conditions, focusing on some aspects of reality and forgetting others. In the management of uncertainty, imprecision is something innate to the human being, both in his way of thinking and in his way of speaking. In the real world there are numerous sources of uncertainty. The information may be imprecise, incomplete and erroneous. Statements like “Luis is much older than Ana” are difficult to represent with predicates of bievaluated logic. Fuzzy logic manages imprecise quantifiers "quite", "often", "sometimes" ... Sometimes information is true but the defined model is imprecise. To manage uncertainty, certainty factors and fuzzy logic are used. The certainty factors expresses the reliability with which we can accept the hypothesis in the case of having the evidence. Fuzzy logic affirms that statements are more or less true in certain contexts and more or less false in a different one. To do this, it manages imprecision by indicating the degree of membership of its members to a set.

3 Cases studies.

The methodology has been applied to two case studies Lorenzo, A. et al. [11]. The aim in both cases was to predict the number of Deputies that each political party will obtain. The artificial intelligence technique of “Rules Based System” is applied. The methodology was applied prior to the celebration of both events and had different results. In April 2019 Spanish General Elections, the expert did not fully appreciate the keys factors. Surveys got a best result. In the General Elections of November 2019, the methodology improved the results predicted by the surveys. The expert fully agreed with the key factors that influenced the results.

4 Conclusions.

We begin with reviewing the literature to find out what techniques are used to make forecasts and predictions. The usual are statistical techniques and machine learning ones. The first case works well when the trend continues. The second case works well when the model has been previously trained with the type of case to predict. The main problem of prediction is complexity because, a priori, there are many variables and the relationship between them is unknown. To reduce complexity, artificial intelligence

techniques are used: heuristics, case-based reasoning, learning by analogy, and rule-based systems. The objective of the thesis is to propose a framework, based on knowledge, which allows predicting some events to improve the effectiveness of current techniques. We have applied it to two use cases before they occurred. In the first case, predicting the electoral results for the general elections in Spain in April 2019, did not work well because the expert did not correctly determine the key factors. In the second case, predicting electoral results in Spain in the general elections of November 2019, worked better than traditional methods because the expert determined the key factors well. The framework is not yet complete because to systematically work on the prediction problem, we must be able to generalize the problem and predict outcomes for new events. Therefore, the next steps we are working on are representing knowledge and managing uncertainty.

5 References.

1. Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) pp. 1643-1647. IEEE.
2. Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access, 6, pp. 55392-55404.
3. Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90, p. 106181.
4. Stuparu, D., Bachmann, D., Bogaard, T., Twigt, D., Verkade, J., de Bruijn, K., & de Leeuw, A. (2017). Case studies of extended model-based flood forecasting: prediction of dike strength and flood impacts. In EGU General Assembly Conference Abstracts Vol. 19, p. 17173.
5. Atsalakis, G., & Valavanis, K. P. (2010). Surveying stock market forecasting techniques-Part I: Conventional methods. Journal of Computational Optimization in Economics and Finance, 2(1), pp. 45-92.
6. Garcés Ruiz, A., Molina Cabrera, A., Ocampo, T., & Mirledy, E. (2006), Stock market forecasting using intelligent techniques. Tecnura 9 (18), pp. 57-66.
7. Vaidehi, V., Monica, S., Mohamed Sheik Safeer, S., Deepika, M., & Sangeetha, S. (2008). A prediction system based on fuzzy logic. Proc. of the World Congress on Engineering and Computer Science, WCECS 2008.
8. Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. Expert Systems with Applications, 36 (3), pp.5932-5941.
9. Nojek, S., Britos, P., Rossi, B., & García Martínez, R. (2003). Sales Forecast: Comparison of Neural Network Based Forecast versus Statistical Method. Technical Reports in Software Engineering, 5(1), pp.1-12. (In Spanish).
10. Brachman, R. J., & Levesque, H. J. (1985). Readings in knowledge representation. Morgan Kaufmann Publishers Inc.
11. Lorenzo, A., & Olivas, J. A. (2019). A Case Study of Forecasting Elections Results: Beyond Prediction based on Business Intelligence. Journal of Computer Science & Technology, 19 (2) pp. 143-152.