International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

March 2020 | Issue 42 | Volume 12 | Number 1 | Pages 25-30

PARTICULAR CASE OF BIG DATA FOR WIND POWER FORECASTING: RANDOM FOREST

J. Bilbao E. Bravo O. Garcia C. Varela C. Rebollar

Applied Mathematics Department, University of the Basque Country, UPV/EHU, Alda. Urkijo, Bilbao, Spain
javier.bilbao@ehu.es, eugenio.bravo@ehu.es, olatz.garcia@ehu.es, concepcion.varela@ehu.es, carolina.rebollar@ehu.es

Abstract- The irruption of Big Data in all kind of technical projects is increasing exponentially. The electrical market is one of these areas where the use of huge quantities of data is being applied. The electricity system has always depended on the supply of generated energy, which in turn depended on the demand for electricity coming from the market. Nowadays, the expansion of renewable energy and distributed generation, now with parts as smart grids, needs the use of data from several sources, including demand on the one hand, and external conditions in the other hand. Moreover, this amount of data is possible to store, transmit and analyze due to the advances in computer science. These advances have significantly improved the forecasting of possible energy generation by renewable producers. This paper analyzes the irruption of Big Data, focusing briefly the study on one of the methods that can be used for the generation of wind power: Random Forest.

Keywords: Energy Generation, Energy Forecasting, Big Data, Wind Power Forecasting, Random Forest.

1. INTRODUCTION

The electricity system has remained very stable over the last few decades in terms of its dependence on new technologies in energy generation. However, since a few years ago, the increase of renewable energies is remarkable. In addition, some governments are trying to promote this type of energy generation through certain economic policies of aid for its development and implementation [1].

On the other hand, although information has always been generated, the capacity to store and analyze it is growing enormously in recent years. At the beginning of the century the rise of relational databases, public web access, wireless and other technologies made the study and management of data in volume a real and current challenge that needed a name.

When we talk about Big Data we understand the recent phenomenon of generating large amounts of data that represent a challenge for their storage, processing and transformation into useful knowledge for a purpose. On the other hand, Machine Learning can be considered a branch of Artificial Intelligence that aims to provide machines with algorithms that can "learn" from experience and generalize behavior, in the same way that a human brain can.

Big Data would be useless without being able to extract valuable information from the data. On the other hand, in order for Machine Learning systems to learn to recognize patterns and predict them, "training" is required, using large amounts of data.

Therefore, the link between these two concepts is the need to learn from data in an automatic way, taking advantage of statistical and computational intelligence for the navigation of large amounts of information with a minimum, or even without, human supervision.

To get to this situation, we have developed many devices and applications that allow to measure, structure, process and analyze large volumes of data by means of new technological infrastructures designed for supporting petabytes of information [2]. Constantly new and refined technologies are appearing that are making possible several aspects related to all fields of the society:

- Improve the efficiency of the organizations;
- Specify tastes and needs of customers and consumers;
- Connect governments and citizens;
- Strengthen security and intelligence of cities;
- Make predictions of all kinds and generate behaviour patterns useful for companies get to know their public better and to make people aware of their own habits;
- Even giving recommendations on how to change them to improve aspects of his life like his health.

All of this is just the beginning, since we are just beginning to tap the potential that offers massive data analysis. Big data has started a more informed society and more efficient.

2. BIG DATA

One of the objectives pursued by companies is to discover an interesting perspective as to the projection pursued by the company itself, the objectives it wants to achieve and knowledge from obtaining data. Discoveries that are frequently expressed as "models", and that we often describe Data Mining as the process of building models [3].

A model, which can be used to help our understanding of the world or created to solve a particular problem, but it can also be used to make short or long-term predictions. The use of Big Data is applied in most fields of activity, such as business, government, financial services, biology, medicine, science and finally engineering.

But it is not all about "only" collecting data. There are also cases where, for example, among the data collected by the electricity companies, there may be hidden clues as to the fraudulent activity of certain customers. This new way of proceeding turns data into information, being the basis for the identification of new opportunities that lead to the discovery of new knowledge, which is the key and fundamental piece of society.

A possible definition of Big Data is the procedure by which large amounts of data are collected and analyzed in a structured and unstructured manner, from various sources, with aim of extracting relevant information of great strategic and tactical value for objectives of business or company [4]. Big data encompasses a four dimensional model that is very important in business sector (Gartner's Three V Model plus one more) [5, 6], which are as follows: volume, velocity, variety and veracity.

Volume: This dimension refers to the size of the sets of data we generate today. Companies are covered by an ever-increasing amount of data of all kinds, thus easily accumulating terabytes, even petabytes of information.

Velocity (and Frequency): Sometimes one minute is too late. In processes where time counts (such as discovering trends, economic losses, etc.), large volumes of data should be used as an added value in the company to maximize value. Data remain outdated and lose their value quickly. About all bearing in mind that data is generated every second, including transactions, photos and video.

Variety: Large volumes of data include any type of data, structured and unstructured, such as text, sensor data, audio, video, click streams or log files, among others. By analyzing this data together, new information is found that is beneficial to the company, and with this data obtained, the business strategy can be effectively prepared.

Veracity: Most of the data comes in dirty form, with fields that are missing or that are incorrect. This can be more complex if several providers use different formats and if data come from different countries, and it can be drastically different in function of local habits and traditions. Cleaning these data can be the most creative activity to generate value.

3. USE CASES OF BIG DATA IN THE ELECTRICAL SUPPLY

Due to the evolution of this technology, more and more agents are involved in the electricity supply. Users: Demand for new and improved services, in addition to enabling the possibility of being able to connect individual energy generation to the grid, in order to be able to sell the energy surplus generated. Another improvement will be real time tariffing and the release to choose energy suppliers.

Electricity grid companies: Network owners and operators will be responsible for responding to user requests in an efficient and cost-effective manner. On the other hand, the energy service companies will be responsible for making tangible the savings obtained thanks to the improvements implemented in the network, highlighting also the savings obtained thanks to changes in people's energy consumption habits.

Researchers and Developers: It will be necessary to make a strong investment in research applied to demand and generation, as well as in the technologies needed to implement the telecommunications network that supports the transfer of data needed for monitoring and control of the network.

Operators: Customers will be able to benefit from the opportunity to choose the energy provider that best suits their needs.

Generators: Power grids are complex integrated systems, with significant interaction between generators, grids, and demand. Therefore, it will be important to encourage the participation of agents that can bring energy to the network, facilitating access to both technological and regulatory level.

Regulators: The European energy market and related services must be supported by a clear and stable regulatory framework.

Government Agents: Governments will have to prepare new legislation to govern all aspects and procedures. It is expected that the increase in competition will result in a reduction in tariffs on the part of users, although on the other hand the integration of renewable energies into the network will require a strong initial investment.

4. USE OF BIG DATA IN WIND POWER GENERATION PROJECTS

All utilities and installations related to renewable energy generation, such as solar photovoltaic energy, thermal, wind, hydraulic, geothermal, tidal, wave, etc., work with a large amount of data. Wind farms or solar farms, for example, have now the ability to collect more and more information, and meteorological agencies are able to predict more and more variables more accurately. At the same time, the system operator has also the ability to collect more and more data in an increasingly connected world.

In order to manage and extract information and knowledge from the available data, it is necessary to use Big Data techniques (including Data Mining, Cloud Computing, etc.). With these tools we can achieve, among others, the following goals:

- Analyze in real time the operation variables of the equipment;
- Analyzing meteorological data in real time;
- Extract behavioral patterns from the installation and from predictive maintenance, reduction of downtime, etc.
- Make predictions that have a direct impact on the efficiency and costs of the installation;
- Extraction of consumption patterns.

In the particular case of wind energy, making short- and medium-term wind predictions has a direct impact on the operation and maintenance of wind farms. Likewise, in order to go to the electricity market, it is necessary to have hourly production forecasts one day in advance (in the daily market).

Making predictions in the wind sector depends on a multitude of factors. One of these factors is the wind and its variable character, which makes necessary to control a multitude of data. In order to extract the knowledge and necessary information from this large volume of data that affect generation, management, distribution, it is necessary to apply Big Data techniques, such as Machine Learning.

Some of the models that are currently being studied more in relation to renewable energies are, on the one hand, the models of Random Forests or Gradient Boosting [7], and on the other hand, Deep Belief Networks, which have been demonstrated their effectiveness [8].

5. WIND POWER FORECASTING

A good prediction is a direct economic reward for a wind farm, because in the energy markets, normally, they must offer the energy they are going to generate and, in case of deviations and of a non-beneficial production to the system, these farms will be penalized.

Good predictions are not only important from a market point of view, but also when we have to perform operational tasks and maintenance of the wind farms. Both types of work have a direct impact on the economy of the farm, and therefore in its electrical company, so if we get a reliable and accurate model, we will have a tremendously high level service. Moreover, it will be very useful and quickly amortizable.

Making these predictions is not easy, because the wind energy presents many fluctuations very changeable and unpredictable for a large period of time, both at a global and a local level, and, especially, if we focus on the place of the farm. This fact is closely related to the changeable character of the wind, which can present strong bursts at a given time and during the next hour show calm. However, it is not the only fact that can have influence for changing the curve of wind energy.

In general, to make these predictions, we need to know meteorological data, normally offered by meteorological agencies such as the Spanish State Meteorological Agency (AEMET), the European Centre for Medium-range Weather Forecasts (ECMWF), the NOAA Global Forecast System (GFS) or the Weather Company (IBM), in order to use them as input data in the statistical models.

These agencies usually provide predictions of certain variables such as the solar radiation, the dew point, humidity or the components of the wind, given in certain points of a rectangular grid covering the area above the one we want to predict them. This grid is the result of a smoothing of the orographic model for a certain resolution (currently, provided by data at a resolution of, at least, 0.125º in the majority of agencies for most of the models).

The number of variables offered by these agencies is huge; for example, GFS offers 145 different variables, at 26 pressure levels for next 16 days with four daily updates, and this by each point of the chosen grid. Obviously, we will never use all of these variables at once, and what we select will depend on the problem that we are managing. And this will determine its dimension. But it is clear that we are going to find a high-dimensional problem, typical of Big Data solutions.

All these features and data related to wind prediction make the tools provided by the automatic learning are the most appropriate to deal with this problem. Traditionally, the following have been used as models: multilayer perceptrons [9] or support vector machines (SVM) [10], where these SVM have been the most used in literature in recent years.

However, due to its transcendence, the prediction of clean and renewable energies, such as the wind energy, continues to be a recurring theme in the research [11, 12], and in order to improve these predictions, new statistical models have been tested.

These include, on the one hand, the models of Random Forests or Gradient Boosting [13], which have produced very good results [14], and which present the advantage of being simpler and more efficient models, and also more easily parameterized than the SVMs.

On the other hand, the effectiveness of deep networks is currently being demonstrated [8], so it seems to be the research line for the next future in this area. In fact, some work on this line has been presented for prediction of wind speed [15] and for testing prediction of wind energy [16].

6. RANDOM FOREST

Random Forest is a combination of decision trees in which each tree depends on the values of a random vector tested independently and with the same distribution for each of them.

In recent years this technique has been very successful due to a number of characteristics:

- With this technique it is possible to perform both classification and regression;
- You can work with supervised and unsupervised learning problems;
- Great precision is obtained in the results thanks to generalization, in which many slightly different trees provide information to obtain a more reliable measure;
- Both training and test can be implemented in parallel.

The Random Forest algorithm is based on a set of decision trees. In other words, a sample enters from above the tree and is subjected to a series of binary tests on each node (split) until it reaches a leaf, in which the answer is found. Therefore, this technique can be thought of as a technique to divide a complex problem into a set of simple problems [17, 18].

A Random Forest is an ensemble of decision trees combined with bagging. By using the ensemble learning technique of bagging, we are causing different trees to see different portions of the data. In this way, no one tree sees all of the training data. This makes each tree train with different samples of data for the same problem. Therefore, by combining their results, some errors will be compensated with others and we will have a prediction that will generalize better.

An ensemble is a set of Machine Learning models. Each model produces a different prediction. The predictions of the different models are combined to obtain a single prediction.

The advantage we get from combining different models is that since each model works differently, its errors tend to be compensated. This results in a better generalization error. There are several ways to build these ensembles [19-21]:

- majority voting
- bagging
- boosting
- stacking

When we use bagging, we also combine several models of Machine Learning. Unlike majority voting, the way to get errors to compensate for each other is that each model is trained with subsets of the training set. These subsets are formed by randomly choosing samples (with repetition) from the training set [22].

The results are combined, for ranking problems with soft voting for models that give probabilities. For regression problems, the arithmetic mean is normally used.

In the training phase the algorithm tries to optimize the parameters of the split functions from the training samples:

$$\theta_j = \arg\max_{\theta_j \in \tau_j} I_j \tag{1}$$

For this we use following information gain function:

$$I_j = H(S_j) - \sum_{i \in 1,2} \frac{\left|S_j^i\right|}{\left|S_j\right|} H(S_j^i) \tag{2}$$

where, S represents the set of samples that are in the node to divide, and S^i are the two sets that are created from the split. The $H(S)$ function measures the entropy of the set, and depends on the type of problem we are dealing with. In the case of regression, we use continuous probability distribution functions, arriving at the expression:

$$I_j = \sum_{v \in S_j} \log\left(\left|A_y(v)\right|\right) - \sum_{i \in 1,2} \sum_{v \in S_j^i} \log\left(\left|A_y(v)\right|\right) \tag{3}$$

where, A_y is the matrix of conditional covariances.

One type of characteristic that we can use to find the best node split are simple binary linear classifiers. For each node j.

$$h(v, \theta_j) \in j = 1, 2, ... \tag{4}$$

where, v is a vector that represents the input sample and θ_j are the parameters to optimize in the node j.

The question of the measure of importance of variables is a crucial and delicate point because the importance of a variable is conditioned to its interaction, possibly complex, with other variables. Random Forest calculates two measures of different importance.

The first, called MDA (Mean Decrease Accuracy), is based on the contribution of the variable to the prediction error, that is, to the percentage of misclassified.

The second measure of importance, called MDG (Mean Decrease Gini), is calculated from the Gini index. This is the criterion used for select the variable in each partition in the construction of the trees and that implies a decrease of this measure.

7. CONCLUSIONS

A Random Forest, is an increasingly popular machine learning technique. Random Forests have a very high generalization capacity for many problems, including prediction for wind farms. Random Forest is a supervised machine learning technique based on decision trees. Its main advantage is that it gets a better generalization performance for similar training performance. This improvement in generalization is achieved by compensating for the errors in the predictions of the various decision trees.

Although it is a model with a very high success rate, it does not always provide good results for all hours of the day or even for every day. It is known that a bad prediction can cause a high deviation in the generation of wind energy, which can lead to serious economic losses when it goes to market. To minimize this risk, it is useful to make combinations with the predictions made by different forecasting agents, thus improving the generalization capacity of each predictor individually.

Using a good combination of prediction methods will take advantage of the complementarity and independence of the different base predictors, by compensating in such combination the errors of opposite sign of the the different methods.

ACKNOWLEDGEMENTS

The use of big amounts of data needs computers that can process them in a reasonable interval of time. We would like to thank the Department of Applied Mathematics, University of the Basque Country, Bilbao, Spain for computer equipment they have made available to us to perform the tests on which this paper is based.

REFERENCES

[1] J. Bilbao, E. Bravo, O. Garcia, C. Varela, M. Rodriguez, P. Gonzalez, "Electric system in Spain: generation capacity, electricity production and market shares", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 9, Vol. 3, No. 4, pp. 91-96, December 2011.

[2] V. Grover, R.H.L. Chiang, T.P. Liang, D. Zhang, "Creating Strategic Business Value from Big Data Analytics: A Research Framework", Journal of Management Information Systems, Vol. 35, No. 2, pp. 388-423, 2018.

[3] W. Chen, H.R. Pourghasemi, S.A. Naghibi, "A Comparative Study of Landslide Susceptibility Maps Produced Using Support Vector Machine with Different Kernel Functions and Entropy Data Mining Models in China", Bulletin of Engineering Geology and the Environment, Vol. 77, Issue 2, pp. 647-664, May 2018.

[4] A. Gandomi, M. Haider, "Beyond the Hype: Big data Concepts, Methods, and Analytics", International Journal of Information Management, Vol. 35, Issue 2, pp. 137-144, April 2015.

[5] "Big Data Definition", Gartner, Inc., www.gartner.com/ it-glossary/big-data/.

[6] S. Sicular, "Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s",

Gartner Inc., 27 March 2013, www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/.

[7] A. Alonso, A. Torres, J.R. Dorronsoro, "Random Forests and Gradient Boosting for Wind Energy Prediction", International Conference on Hybrid Artificial Intelligence Systems, HAIS 2015, pp 26-37, 2015, http://link.springer.com/ chapter/10.1007%2F978-3-319-19644-2_3.

[8] Y. Bengio, "Learning Deep Architectures for AI", Foundations and Trends in Machine Learning", Vol. 2, No. 1, pp. 1-127, Canada, 2009, http://dx.doi.org/10.1561/2200000006.

[9] R. Duda, P. Hart, D. Stork, "Pattern Classification", 2nd Ed., Wiley, New York, USA, 2000.

[10] B. Scholkopf, A.J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond", MIT Press, Cambridge, MA, USA, 2002.

[11] R. Kerimov, Z. Ismailova, N.R. Rahmanov, "Modeling of Wind Power Producing in Caspian Sea Conditions", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 15, Vol. 5, No. 2, pp. 136-142, June 2013.

[12] A. Hajebrahimi, R. Moradi, M. Rashidinejad, A. Abdollahi, "Impact of Wind Energy Penetration to Connect the Large-Scale Distant Wind Farm into the Grid in Probabilistic Multi Objective Transmission Expansion Planning", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 18, Vol. 6, No. 1, pp. 82-88, March 2014.

[13] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning", 2nd Ed., Springer, 2009.

[14] A. Alonso, A. Torres, J.R. Dorronsoro, "Random Forests and Gradient Boosting for Wind Energy Prediction", LNCS, Lecture Notes in Computer Science, 9121, pp. 26-37, Springer, 2015.

[15] Q. Hu, R. Zhang, Y. Zhou, "Transfer Learning for Short-Term Wind Speed Prediction with Deep Neural Networks", Renewable Energy, Vol. 85, pp. 83-95, 2016.

[16] D. Diaz, A. Torres, J.R. Dorronsoro, "Deep Neural Networks for Wind Energy Prediction", LNCS, Lecture Notes in Computer Science, 9094, pp. 430-443, 2015.

[17] L. Breiman, "Bagging predictors", Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996.

[18] L. Breiman, "Random forests", Machine Learning, Vol. 45, No 1, pp. 5-32, 2001.

[19] M. Akour, I. Alsmadi, I. Alazzam, "Software Fault Proneness Prediction: A Comparative Study between Bagging, Boosting, and Stacking Ensemble and Base Learner Methods", International Journal of Data Analysis Techniques and Strategies, Vol. 9, No. 1, 2017.

[20] S. Agarwal, C.R. Chowdary, "A-Stacking and A-Bagging: Adaptive Versions of Ensemble Learning Algorithms for Spoof Fingerprint Detection", Expert Systems with Applications, Available online 24 December 2019.

[21] R. Campos, S. Canuto, T. Salles, C. C. A. de Sa, M. A. Gonçalves, "Stacking Bagged and Boosted Forests for Effective Automated Classification", SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information, pp. 105-114, 2017, https://doi.org/10.1145/3077136.3080815.

[22] D.R. Nayak, R. Dash, B. Majhi, "Brain MR Image Classification Using Two-Dimensional Discrete Wavelet Transform and Adaboost with Random Forests", Neurocomputing, Vol. 177, pp. 188-197, 2016. doi: 10.1016/j.neucom.2015.11.034.

BIOGRAPHIES

Javier Bilbao obtained the degree in Electrical Engineering from University of the Basque Country, Spain, in 1991. At present he is Ph.D. in Applied Mathematics and Professor at the department of Applied Mathematics of that university. He has been General Chairman of some conferences of WSEAS organization. He is the General Chairman of the International Conference on Engineering and Mathematics (ENMA) and member of the Organizing, Executive and Scientific Committees of the International Conference on Technical and Physical Problems of Engineering (ICTPE). His current and previous research interests are distribution overhead electrical lines compensation, Optimization of series capacitor batteries in electrical lines, modelization of a leakage flux transformer, losses in the electric distribution networks, artificial neural networks, modelization of fishing trawls, e-learning, noise of electrical wind turbines, light pollution, health risk of radiofrequencies.

Eugenio Bravo obtained the degree in Electrical Engineering from University of the Basque Country, Spain, in 1991. At present he is Ph.D. in Applied Mathematics and Professor at the department of Applied Mathematics of that university. His current and previous research interests are distribution overhead electrical lines compensation, optimization of series capacitor batteries in electrical lines, modelization of a leakage flux transformer, losses in the electric distribution networks, artificial neural networks, modelization of fishing trawls, e-learning, noise of electrical wind turbines.

Olatz Garcia obtained the degree in Mathematics from University of the Basque Country, Spain, in 1989. At present she is Ph.D. in Applied Mathematics and Professor at the department of Applied Mathematics of that university. Her current and previous research interests are e-learning, optimization of series capacitor batteries in electrical lines, noise of electrical wind turbines.

Concepcion Varela obtained the degree in Mathematics from UNED, Spain, in 1986. At present she is Ph.D. in Economics and Statistics and Professor at the department of Applied Mathematics of the University of the Basque Country. Her current and previous research interests are e-learning, optimization of series capacitor batteries in electrical lines, noise of electrical wind turbines.

Carolina Rebollar obtained the degree in Mathematics from University of the Basque Country, Spain, in 1986. At present she is Ph.D. in Applied Mathematics and Professor at the Department of Applied Mathematics of that university. She is the Academic Secretary at the Faculty of Engineering Bilbao (University of the Basque Country) and Director of the "Aula Company ZIV", where projects are developed annually for the electric company "ZIV I+D Smart Energy Networks", participating teachers and students.