"Mircea cel Batran" Naval Academy Scientific Bulletin, Volume XIX – 2016 – Issue 1
Published by "Mircea cel Batran" Naval Academy Press, Constanta, Romania // The journal is indexed in: PROQUEST / DOAJ / DRJI / JOURNAL INDEX / I2OR / SCIENCE LIBRARY INDEX / Google Scholar / Crossref / Academic Keys / ROAD Open Access / OAJI / Academic Resources / Scientific Indexing Services / SCIPIO

MACHINE LEARNING TECHNIQUES USED IN BIG DATA

Stefania Loredana NITA[1]
Laurentiu DUMITRU[2]
Adrian BETERINGHE[3]
[1]Integrated Systems Department, Institute for Computers
[2]Military Technical Academy
[3]Department of IT&C, LUMINA – The University of South East Europe

Abstract: *The classical tools used in data analysis are not enough in order to benefit of all advantages of big data. The amount of information is too large for a complete investigation, and the possible connections and relations between data could be missed, because it is difficult or even impossible to verify all assumption over the information. Machine learning is a great solution in order to find concealed correlations or relationships between data, because it runs at scale machine and works very well with large data sets. The more data we have, the more the machine learning algorithm is useful, because it "learns" from the existing data and applies the found rules on new entries. In this paper, we present some machine learning algorithms and techniques used in big data.*

Keywords: *Big Data, machine learning, supervised learning, unsupervised learning.*

Introduction

The 2012 year could be seen as the Big Data technologies year, while the 2013 could represent the year of data analysis. Some of the most important things are data collecting and data management, but a challenge represents the extraction of helpful information from these datasets. Big Data has a major impact over comprehensions extraction and explanation, changing the instruments used in predictive analytics.In the traditional way, the domination in data science was represented by the trials and error analysis, but these techniques are less useful when the data collections are enormous and diverse. Moreover, the more data we have, it becomes almost impossible to construct a predictive model, because, there is a small number of tools which process large amount of data in an acceptable time. Additionally, usually, classical statistical methods are concentrated on static analytics which is restricted to the observations of representatives that are iced in time and often leads to exceed and untrustworthy conclusions.

Because of these issues, the researchers have made efforts in order to find alternative techniques for analysis of large amount of data. Thereby, the interest for machine learning has rapidly growth.As the amount of information has increased, the scientists are trying to find ways in order to process these data in real time, such that to be obtained precise predictions or descriptions of different types, which could be integrated in powerful applications in various domains. A few examples are recommendation systems and division of customers in different groups, cheat detection or many others.The machine learning methods are fit very well with these requirements, and use a suite of general techniques different from the classical statistical ones.

In this paper we present the main machine learning techniques used in big data. The paper contains five sections – the first section is the *Introduction*, next the section *Big Data*, where we present some characteristics of big data, then next the *Machine Learning* section, in which we present some existing definitions of machine learning, we give some examples of daily life machine learning applications. The next section is *Machine Learning Techniquesused in Big Data*, where we present the most important techniques and give example of uses for each technique, and finally, next the *Conclusions* section, where we present the conclusions of the paper.

Big Data

Big Data is a term which refers to large amounts of information (organized or unorganized), collected from any domain of activity. The important thing is not the data itself, but the way the data is manipulated and analyzed. In other words, big data represents the data which overtakes the processing capabilities of traditional database systems. These data are too large, move too rapidly, or do not match with the database architectures. So, there is the need of others tools which process data.

In [1], Doug Laney have defined Big Data using the three *V*s (Figure 1). These come from *Volume*, *Velocity*, and *Variety*.

Volume. The data is gathered from many different resources, such as transaction processes, community media, and many others. In the last

DOI: 10.21279/1454-864X-16-I1-078

few years, it have developed new technologies which ease data manipulation, like Hadoop.

Velocity. This characteristic refers to how fast the information is operated and analyzed. The flowing of information is intensive and uninterrupted. The data is received in real time and it could help the researchers and the businesses to have more valuable and accurate decision process for providing strategical competing advantages.

Variety. This characteristic refers to the variety of the data, which could be structured or unstructured. The classical storage for the data are spreadsheets or databases, but these are facing with difficulties when comes from nowadays data. The nowadays information comes from e-mails, pictures, recordings, tracking devices and many others, which represent unstructured data.

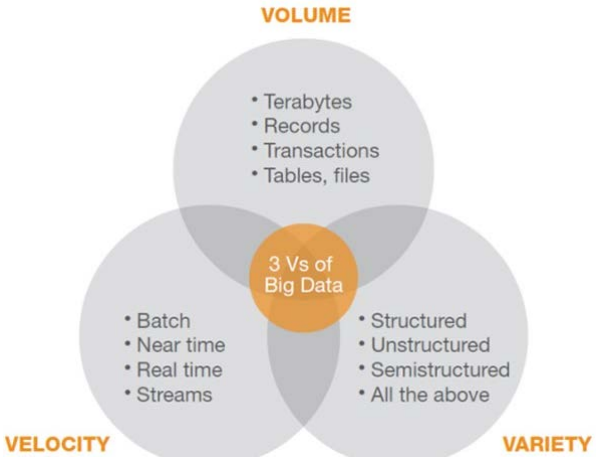

Figure 1. The 3Vs which characterize Big Data [2]

However, in other opinions, besides the 3 *V*s, the Big Data is characterized by another two *V*s, which come from *variability* and *complexity*[3] or *validity* and *volatility*.

Machine Learning

Machine learning is a subdomain of computer science used in order to analyze the data, which automates the construction of analytical models. The purpose of machine learning algorithms is to learn from the existing data "*without being explicitly programmed",* as defined by Arthur Samuel in 1959 [3].An important aspect regarding machine learning is that, when the models are applied on new data sets they are adapting independently, characteristic which comes from the iterative feature of machine learning. These models are learning from preceding calculations for producing certain and replicable decisions and outcomes.As the technology advances, the old machine learning techniques do not fit well with the large amount of data. In the last years, the researchers have created new machine learning techniques, which perfectly match with big data. Below, are some examples of daily life machine learning applications:

- Google's autonomous car. These cars contains sensors which are detecting the object from a large area (approx. the dimension of two football fields) in all directions. Among the detected objects are persons on foot, bicyclists and other another vehicles, but also flying shopping sack of plastic or rogue birds. The car is equipped with a software which analyzes all received data and process them for a safety navigation on the road [4], [5]. This example of machine learning application represents the quintessence of machine learning.
- The recommendation systems of Amazon or Netflix. For example, in 2014, the Netflix recommendation system was based on restricted Bolzman machine and a kind of matrix factorization [6], [7]. These are examples of daily life machine learning application.
- Fraud detection. This is one of the most important use case of machine learning nowadays. Both, e-business and e-commerce represent a challenging duty because it is difficult to distinguish a limit between systems for fraud detection and systems for network intrusion detection [8].

The interest for machine learning has increased because it works well with large amount and manifold of data. Also, the computation processing is more inexpensive and strong.As a result, the models for analyzing the large and more complex data and for faster delivering and more accurate outcomes, are produced quick and automatically. The use of these models lead to very precise predictions through which are taken better decisions and intelligent actions on real time in the absence of human interference.

There are two types of machine learning techniques: *supervised learning* and *unsupervised learning*.

Supervised learning is often used in problems in which the data should be classified, and is based on adefined classification model from which the computer should learn. Specifically, the classification learning is useful in all problems based on the deduction of a classification. It is possible to not be need of pre-defined classification rules. Supervised learning is the most used method used to train the neural networks or decision trees, about which we will

DOI: 10.21279/1454-864X-16-I1-078

discuss in section four. These two techniques are dependent of the pre-determined classification rules.

Unsupervised learning is more complicated because the computer should learn how to accomplish a task without having any instructions. This type of learning is nearly of real world, generalizing it.

Even if the unsupervised learning is stronger than supervised learning, in practice is used more supervised learning, than unsupervised.

Machine Learning Techniques in Big Data

The steps in data analysis are:

- Defining the problem;
- Identification of data source;
- Collecting and selection of data;
- Preparing the data;
- Construction of the model;
- Evaluation of the model;
- Integration of the model.

A few of the most popular machine learning techniques are: artificial neural networks, genetic algorithms, cluster analysis.

Artificial neural networks. A special type of networks are the artificial neural networks (ANN), in which the nodes represent artificial neurons. The first creators of the artificial neuron are McCulloh and Pitts in 1943. The inspiration for the artificial neuron (Figure 2) is the biological neuron. The last one works as follows: the dendrites or neuron's membrane contains synapses through which the neuron receives different signals. If the received signal has a certain intensity, in other words, if the signal is passing over a specific threshold, the neuron becomes activated and sends a signal to its axon. The sent signal could be sent further to another synapse or could activate another neurons. The natural neuron is abstracted in order to make a model which represent the artificial neuron. This is characterized by the following: inputs (analogous of the synapses), weights (analogous with the intensity of the signal), and a mathematical function which determines whether the neuron will activated or not, and another function that calculates the output (the function could be the identity and could depend of a threshold) [10]. If the weights are negative, the signal is inhibited. If the weights are good (depending on specific criteria) will be obtained the desired outputs for the specific inputs. In order to adjust the weights for the artificial neurons, are used learning or training algorithms. The ANNs are used in different domains:

- Classification, which includes pattern identification and succession identification, newness detection and sequential decision taking;
- Control, which includes computer numerical control;
- Data processing, which include techniques like filtering or clustering;
- Function approximation, regression;
- Robotics.

An example of neural network application could be found in [11], where the authors have used the artificial neural network in sentiment analysis over twitter in order to find the users opinions regarding a specific brand. Another example of application is [12], where the authors have studied if the ANNs are useful in predicting solar radiation. An example of study which shows clearly the link between big data and machine learning is [13], where the authors have used neural network in order to create a semi-supervised model, which has as inputs the spatial characteristics. The study [13] analyses the urban air quality.

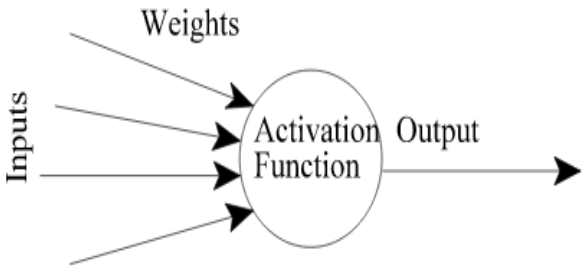

Figure 2. The artificial neural network [10]

Genetic algorithms. The genetic algorithm (GA) represents a search technique which imitates the natural selection [14]. This heuristic technique is used to engender solutions for problems regarding optimization and search. The creator of genetic algorithms is John Holland, in 1960. In order to solve a problem, genetic algorithms are simulating the outliving of the fittest through individuals on successive generations. Every generation represents a population consisting in series of characters similar to the chromosomes from out DNA. The individuals are points in a searching field and potential solutions, and are the subject of the evolution process. The inspiration for genetic algorithms are the chromosomes, which are abstracted such that to become a theoretical model. The ideas behind GAs are [15]:

- In a population there is a competition for resources and pair.
- The individuals which successfully handle the competition are producing more descendants than those individual which poorly handles.
- It is expected that a descendant to be better than the both parents, because, it inherits the

DOI: 10.21279/1454-864X-16-I1-078

best genes from them. So, every generation is better than the preceding generation.

The first step is to generate an arbitrary population, and then, to apply the genetic operators: selection (which determines the survivals form the fittest), crossover (which represents the pairing of the individuals) or mutation (which adds random changes) [15].

An example of genetic algorithms application is [16], in which the authorspresents an effective hybrid genetic search used in a big category of problems regarding vehicle routing.

Cluster analysis. Cluster analysis splits the data in different groups, called clusters, which are relevant, helpful, or both. If the relevant groups are the purpose, then the clusters should catch the natural structure of the information. Sometimes, cluster analysis represents a beginning point for another scopes, like summarization. Cluster analysis was one of the most important techniques, applied in different areas of study, like social sciences, natural sciences, medicine or businesses [17]. There are two main types of clustering: *partitional clustering* (Figure 3) and *hierarchical clustering*. The partitional clustering just divides the dataset objects into distinct subsets (clusters), as every object stands in only one subset. Thus, every collecting of clusters represents a partitional clustering.If it is allowed the clusters to have sub-clusters, then we are talking about hierarchical clustering that represent a group of nested clusters, represented as a tree. Every node, excepting the leaves, is the unification of its children nodes, so the root represents the initial set of objects[17]. Sometimes, the leaves represent sets of a single object. Cluster analysis does not represent itself an algorithm, but represent the general task, which should be accomplished. Examples of clustering algorithms are: k-means - for centroid models, density-based spatial clustering of applications with noise (DBSCAN) orordering points to identify the clustering structure (OPTICS) – for density models, biclustering or co-clustering – for subspace models, etc. [18], [19].

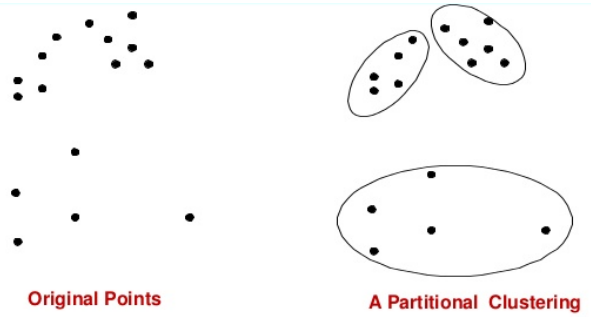

Figure 3. Partitional Clustering [20]

Decision trees. Decision trees are a specific type of graphs, which use branches to highlight all possible results of a decision. They are used for simplifying the complex strategical provocations and for evaluating the cost-effectiveness of made decisions. An application of decision trees is [21], where the authors have used them in order to detect possible financial frauds, or [22], where the authors have used decision tree in order to determine a specific firm performance.

Support vector machine.The basic concept behind Support Vector Machine (SVM) is decision planes, which lead to decision limits. The decision plane that splits a set of object in various class memberships[23]. In Figure 4 is an example of SVM, in which, the set of object was separated in more classes. In practice, the classification process is not simple, and, when it is based on the separation using different lines it is known as hyperplane classifiers. SVM is used in classification or regression and could work with different types of variables. There are more types of SVM[23]:

- SVM Type 1 for classification process (or C-SVM classification);
- SVM Type 2 for classification process (or nu-SVM classification);
- SVM Type 1 for regression process (or epsilon-SVM regression);
- SVM Type 2 for regression process (or nu-SVM regression).

An example of SVM application is [24], where the authors have used SVM in order to create a new hybrid classifier system for deciding if a magnetic resonance image of a brain is normal or not.In [25], the authors prove that the SVM could be used in quantum computers.

DOI: 10.21279/1454-864X-16-I1-078

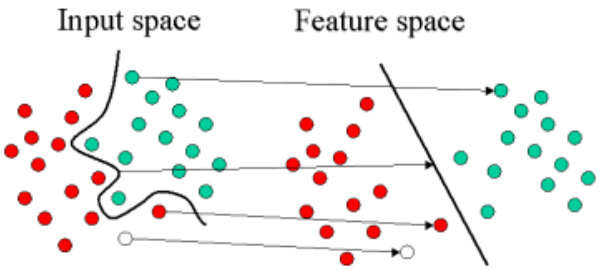

Figure 4. Planes of SVM [23]

Reinforcement learning.Reinforcement learning represents a type of learning in which, an agent should learn from the real world in order to maximize its recompense. Learner has no instructions of decisions, but if the task is successfully accomplished then it is rewarded, otherwise it is punished. When the agent will have a similar situation, it will make decisions based on past experience. The reward is a numerical value received as a signal, and codes the successfulness rate of an action. The main task is to learn to take decisions such that the reward to be always improved, so this is a process of trial and error learning. In [26], the authors discuss about the impact of big data and techniques like reinforcement learning in psychology

The above presented methods are the most used techniques. Of course, depending of the model of the data, could be used another machine learning techniques,like: deep learning, inductive logic programming, representation learning, Bayesian networks, and many others.

CONCLUSIONS

In this paper we have discussed about big data, and have presented some characteristics of machine learning. We have seen that machine learning has many application in daily life. Next, we have presented the most used machine learning techniques in big data, giving examples of applications for every described technique. We have seen that machine learning techniques have many applications domains, such as medicine, nature sciences, finance, and many other.

BIBLIOGRAPHY

[1] Doug Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, META Group Research Note, 2001, 6: 70.
[2] Michael Walker, *Data Veracity*, 2012, http://www.datasciencecentral.com/profiles/blogs/data-veracity
[3] *Big Data History and Current Considerations*, http://www.sas.com/en_us/insights/big-data/what-is-big data.html
[4] *Machine Learning*, https://en.wikipedia.org/wiki/Machine_learning
[5] Sharon L.Poczter, and Luka M Jankovic. *The Google Car: Driving Toward A Better Future?*, Journal of Business Case Studies (Online), 2014, 10.1: 7.
[6] *Google Self-Driving Car Project*, https://www.google.com/selfdrivingcar/
[7] Stéphane Gaıffas, and Olga Klopp. *High Dimensional Matrix Estimation with Unknown Variance of the Noise*, 2015, http://www.cmap.polytechnique.fr/~gaiffas/gaiffas/files/papers/variance_jan_2015.pdf
[8] Xavier Amatrian. *How does the Netflix movie recommendation algorithm work?*, 2014, https://www.quora.com/How-does-the-Netflix-movie-recommendation-algorithm-work
[9] Clifton Phua, Vincent Lee, Kate Smith and Ross Gayler. *A Comprehensive Survey of Data Mining-Based Fraud Detection Research*, arXiv preprint arXiv:1009.6119, 2010.
[10] Carlos Gershenson, *Artificial Neural Networks for Beginners*, 2003, https://arxiv.org/ftp/cs/papers/0308/0308031.pdf
[11] M. Ghiassi, J. Skinner, D. Zimbra. *Twitter Brand Sentiment Analysis: A Hybrid System using N-Gram Analysis and Dynamic Artificial Neural Network*, Expert Systems with applications, 2013, 40.16: 6266-6282.
[12] S. Shanmuga Priya, Mohammad Hashif Iqbal. *Solar Radiation Prediction using Artificial Neural Network*, International Journal Of Computer Applications, 2015, 116.16:0975-8887
[13] Yu Zheng, Furui Liu, Hsun-Ping Hsieh. *U-Air: When Urban Air Quality Inference Meets Big Data*, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013. p. 1436-1444.
[14] *Genetic Algorithm*, https://en.wikipedia.org/wiki/Genetic_algorithm
[15] Kumara Sastry, David E. Goldberg, Graham Kendall. *Genetic Algorithms*, Search methodologies, Springer US, 2014. pp. 93-117.
[16] Thibaut Vidal, Teodor G. Crainic, Michel Gendreau, Christian Prins. *A Hybrid Genetic Algorithm with Adaptive Diversity Management for a Large Class of Vehicle Routing Problems with Time-windows*, Computers & Operations Research, 2013, 40.1: 475-489.

DOI: 10.21279/1454-864X-16-I1-078

[17] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Data mining cluster analysis: Basic concepts and algorithms*, 2013.
[18] *Cluster analysis*, https://en.wikipedia.org/wiki/Cluster_analysis
[19] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, SebtiFoufou, Abdelaziz Bouras, *A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis*, Emerging Topics in Computing, IEEE Transactions on, 2014, 2.3: 267-279.
[20] Mona M. Suliman, *Data Clustering Using Swarm Intelligence Algorithms An Overview*, http://www.slideshare.net/AboulEllaHassanien/data-clustering-using-swarm-intelligence-algorithms-an-overview
[21] Yusuf Sahin, Serol Bulkan, Ekrem Duman. *A cost-sensitive decision tree approach for fraud detection*, Expert Systems with Applications, 2013, 40.15: 5916-5923.
[22] Dursun Delen, Cemil Kuzey, Ali Uyar. *Measuring firm performance using financial ratios: A decision tree approach*, Expert Systems with Applications, 2013, 40.10: 3970-3983.
[23] *Support Vector Machines (SVM) Introductory Overview*, http://www.statsoft.com/Textbook/Support Vector-Machines
[24] Yudong Zhang, Shuihua Wang, Genlin Ji, Zhengchao Dong. *An MR brain images classifier system via particle swarm optimization and kernel support vector machine*, The Scientific World Journal, 2013, 2013.
[25] Patrick Rebentrost, Masoud Mohseni,Seth Lloyd*. Quantum support vector machine for big data classification*, Physical review letters, 2014, 113.13: 130503.
[26] Alex Gomez-Marin, Joseph J. Paton, Adam R. Kampff, Rui M. Costa, Zachary F. Mainen. *Big behavioral data: psychology, ethology and the foundations of neuroscience*, Nature neuroscience, 2014, 17.11: 1455-1462.

DOI: 10.21279/1454-864X-16-I1-078