ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
March 2017,
Vol. 10, No. (1):
Pgs. 187-193

Predictive Analytics - The Cognitive Analysis

T.VENKAT NARAYANA RAO, SOHAIL ALI SHAIK and S. MANMINDER KAUR

Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology Yamnampet, Ghatkesar, Rangareddy,India.

http://dx.doi.org/10.13005/ojcst/10.01.25

(Received: March 07, 2017; Accepted: May 16, 2017)

ABSTRACT

Predictive analytics plays an important role in the decision-making process and intuitive business decisions, by overthrowing the traditional instinct process. Predictive analytics utilizes data-mining techniques in order to predict the future outcomes with a high level of certainty. This advanced branch of data engineering is composed of various analytical and statistical methods which are used to develop models that predict the future occurrences. This paper examines the concepts of predictive analytics and various mining methods to achieve the prior. In conclusion, paper discusses process and issues involved in the knowledge discovery process.

Keywords: BigData; Predictive Analytics; Predictive Modelling; Data Mining; Prediction.

INTRODUCTION

Predictive Analytics is a division of data mining that deals with the analysis of existing data to perceive the concealed trends, patterns and the relationship between these to predict the future probabilities and patterns[1]. It incorporates concepts of statistics, data mining and machine learning that analyses the historic and existing data to predict future events[2]. The notable step in Predictive analytics is predictive modelling, where the data is collected, and formulation of the statistical model takes place and the predictions are made. Furthermore, the model is ratified and revised as additional data is made available[3]. Predictive data mining uses the training dataset and automatically generates the classification model from it and implements this model to predict the unknown constraints of unclassified datasets[4].

Statistical models created, learn from the derived patterns and apply the knowledge attained to future datasets[5]. Machine learning methods build multi-variate models from extant data and consequently develop solutions to unknown data[6]. Machine learning model is generally categorized as the supervised classification approach. The process of creation of the model is preceded by data pre -processing to clear the noise to ensure that true data is analysed[7]. Predictive analytics automatically analyses the huge amount of data with various constraints, but the primary variable to consider is

'predictor', the element that aids an entity to predict and measure its future behavior[8].

Predictive analytics is the combination of various statistical techniques ranging from data mining, game theory, modelling and machine learning. These approaches analyse the historic and present data to make predictions about the future[9].Decision models portray the relationship between all the constraints of a decision i.e.; the known data (results of predictive models inclusive) and the decision. These models maximize certain outcomes while minimizing others, and finally augment the optimization of end result[8].

Predictive Analytics and BigData

The profits and capability of predictive analytics have recently been appreciated by numerous researchers due to the contemporary technology BigData and the compact relationship between them[10]. In this modern world we cannot rely on the blind beliefs to support the business decisions, thus a novelway to make this scientific is predictive analytics, but this is possible only with thehuge amount of data and the solution is 'BigData'.

This explosive accretion of data has diverse sources and different datasets ranging from the public domain and comprising of enterprise data, sensor data, data from transactions and social media (Fig. 2 and 3). Of the above data - 85% is unstructured or not metric data (SAS Institute,2012) colossal and complex in volume, velocity, variety, veracity and variability[11].

Such huge amount of data is significantly ahead of typical data processing and analytics tools, thus predictive analytics has the capability to deal with raw, large-scale datasets and complex models[11].

Predictive analytics is very integral to uphold and maintain BigData, this technology not only makes it possible to tackle the capability of BigData but also organize the datasets. Predictive analytics mutates the bulk unstructured data into meaningful, profitable business information[11]. The extent of benefit by employing predictive analytics is widespread to various departments. Although this will be applicable to government operations it is very extensively used by corporate sectors[12].

For instance consider the scenario of retail super market chains where they use this technology to understand and analyse the current and earlier data, observe and determine the customer and product trends and avail these behaviours to predict about the product/s customers are most likely to purchase. Predictive analytics is utilized even in the financial services companies and commercial banks, here the technology is used to identify the

Fig. 1: Predictive analytics (softnetsearch.com)

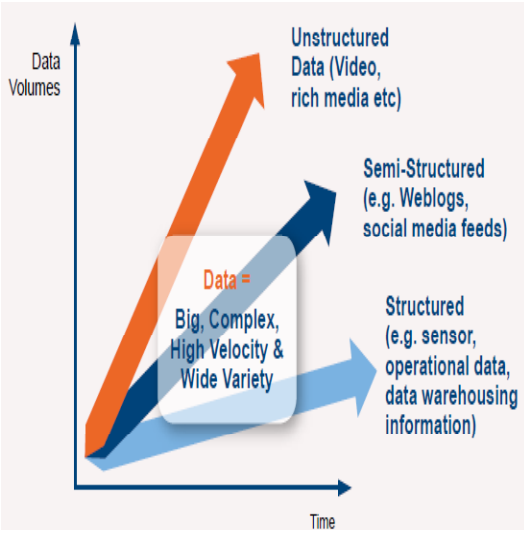

Fig. 2: Types of data—IDC (2012)

customer patterns from the existing data. Then the patterns are used to anticipate the customer's stand on taking loan or mortgage repayment and also the credit and debit card overdraft limit.

The expanse of predictive analytics in association with BigData is not just bounded to the above areas but also provides the services to Healthcare sector to predict the occurrence of noxious diseases, to Insurance sector in foreseeing the bogus and fraud claims and too many more.

Predictive Analytics Process

Predictive analytics is widely used to analyze the patterns and trends from existing data and then applying these trends to the current data and derive a solution, foresee the behaviour of unpredictable constraints. The process of analysing the data and deriving the patterns is not a single step process, but it involves multiple degrees in order to achieve the final result.

Define Project

It identifies datasets which needs to be imported to implement the analysis along with defining the project scope, outcomes, requirements and business objectives.

Data Collection

Data collection is the process where the data from multiple sources, which are required for analysis, are collected. In this step the process of data pre - processing is also initiated. This gives the total perspective of the customer's transactions [13].

Data Analysis

Data Analysis is the phase where cleaning, scrutiny and organizing, transformation of data is carried in order to lay the groundwork for generation of statistical model. This stage is the initial stride to the process of decoding the patterns from the data.

Statistics

Statistical analysis is one of the prominent phase in the process of predictive analytics, here the data in ordered format that has been analyzed in the earlier stage is represented statistically and helps to verify or corroborate the assumptions. This also helps to investigate with help of statistical models[13].

Modeling

The most imperative phase of predictive analytics process is modelling. In this phase the data and models from preceding phases are gathered and the predictive model is designed from them, which thus accurately predicts the outcome of the future events and eventually the results are attained[13].

Deployment

The process where the developed model is applied to the new data in which constraints are

Fig. 3: Sources of Big Data—Huijbers (2012)[15]

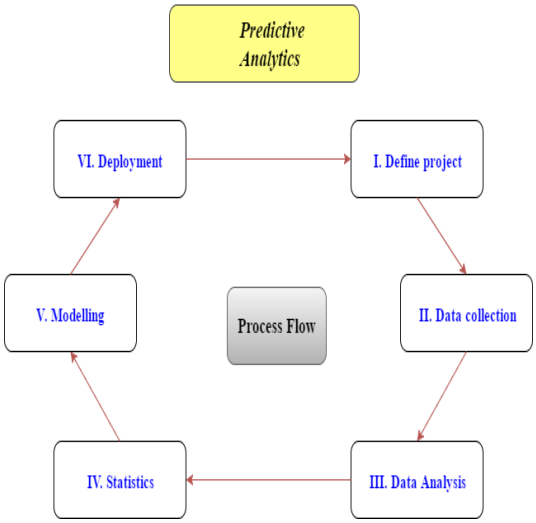

Fig. 4: Predictive Analytics process

missing is known as Deployment. Deploying this model helps us to enhance the decision making process by the application of analytical results and reporting them.

Predictive Analytics Techniques

Predictive Data Mining techniques are favourable to reach a conclusive decision based on the missing constraints and the historical data. The main idea for application of these techniques is to derive the patterns from the existing data and thus concluding the decision. Several predictive rules are cluster analysis, classification, association rule mining, comparison, and characterization.

Decision Trees

A decision tree is usually represented as flowchart like tree structure, which consists of internal nodes, branches and Leaf nodes. The purpose of a decision tree is to classify an unknown sample, which is testing the attribute values of the sample against the decision tree. The internal nodes denote tests on an attribute, branches represent the outcome of the test and leaf node represents class labels or class distributions. There are two different phases involved in decision tree generation:

Tree construction

In this phase, initially all the training examples are at the root. Then the partition of attributes occur recursively based on the set of selected attributes.

Tree Pruning

This is the phase where we identify and remove the branches that reflect noise and are outliers. The techniques used here are pre-pruning and post-pruning, where the "best-pruned tree" is established with respect to data other than the training data.

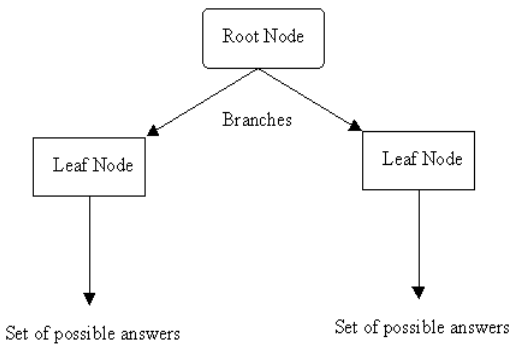

Fig. 5: Decision Tree structure

The composition of a decision tree is top-down recurrent divide and conquer manner. Initially, all the training examples are at the root and the attributes are categorical i.e.; if continuous-valued they are discretized in advance. Examples are segregated periodically established on the selected attributes, test attributes are preferred on the basis of a statistical measure (e.g.; information gain).

The alternatives of decision tree algorithms comprise of CART, ID3, C4.5, SLIQ and SPRINT the measure for attribute selection is Information gain, where all the attributes are assumed to be categorical and can be modified for continuous valued attributes. The decision of selection of the attribute to initiate the partition is taken based on the information gain. Attribute with the higher information gain is the one to start the partition of the decision tree.

Firstly, assume that there are two classes, P and N. Let the set of examples S contains p elements of class P and n elements of class N. The amount of information required to determine if an irrational example in S belongs to P or N is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Then assume that considering attribute A set S will be partitioned into sets $\{S_1, S_2 \ldots S_v\}$. If S_i contains p_i examples of P and n_i examples of N, the entropy, or the predicted information required to segregate objects in all sub trees S_i is

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

The information that would be gained when the branching would be based on A is

$$Gain(A) = I(p,n) - E(A)$$

Thus calculate gain for all the attributes at each step of the decision tree and determine the optimal attribute.

Artificial Neural Networks

Neural network is biologically motivated approach to machine learning. It is a set of connected input & output units, where each connection has a weight associated with it. Neural network learning is also known as connectionist learning due to the connections between the units. The considerable plus point regarding neural networks is it learns by adjusting the weights so as to correctly classify the training data and thus after the testing phase classify the unknown data. Neural networks also have a high tolerance to noisy and incomplete data, but it needs longer time for training.

The deeply driven formula adapted in neural networks is the classification of data, the inputdata set contains classification attribute. As in the case of usual classification problem the input data is divided into training data and testing data. The imperative step to be considered for all the data is normalization. All values of attributes in the database are changed to contain values in the internal as [0,1] or [-1,1]. Neural networks can work in the range of (0,1) or (-1,1). Max-Min normalization and Decimal scaling normalization are the two most basic normalization techniques followed.

Max-Min Normalization

The frequently used normalized technique to scale the values between fixed ranges is referred to as the Max-Min normalization technique. The formula applied here is

$$v' = \frac{v - \min A}{\max A - \min A}(new\_\max A - new\_\min A) + new\_\min A$$

min A - Minimum value of attribute A
max A - Maximum value of attribute A

Normalization formula above maps a value v of A to v' in the range {new_minA, new_maxA}

If we wish to normalize the data to a range of the interval [0,1] we put:

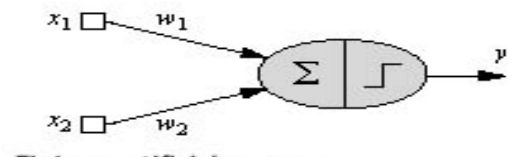

Fig. 6: Artificial Neuron

new_max A= 1new_minA =0

Decimal scaling normalization

The technique of normalization by decimal scaling scales the ranges by shifting the decimal point of values of attribute A.

$$v' = \frac{v}{10^j}$$

Here j is the smallest integer such that max|v'|<1.

Here x1 and x2 are normalized attribute value of data. y is the output of the neuron, which is the class label. x1 and x2 are multiplied by weight values w1 and w2 are input to the neuron x. The value of x1 is multiplied by a weight w1 and value of x2 is multiplied by a weight w2. The inputs are fed simultaneously into the input layer and the weighted outputs of these units are fed into a hidden layer. The weighted outputs of the last hidden layer are inputs to units making up the output layer.P(X) is constant.

Naïve Bayesian

Bayesian classification is probabilistic learning, where we calculate the explicit probabilities for hypothesis and is among the most practical approaches to certain types of learning problems. Prior knowledge and observed data can be merged which in turn helps to calculate explicit probability. Bayesian provides a useful perspective for understanding many learning algorithms and It is robust to noise in input data. Generally the models are built in forward casual direction, but baye's rule allows us to work backward using the output of the forward model to infer causes or inputs.

Let X be a data sample whose class label is ambiguous and let H be some hypothesis that X belongs to a class C. For the classification, we need to determine P(H/X). P(H/X) is the probability that H holds given the observed data sample X. Here P(H/X) is a posterior probability.

P(X), P(H) and P(H/X) may be estimated from given data.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Steps involved

1. Each data sample is of the type

$X=(x_i)$ i =1 to n, where x_i is the values of X for attribute A_i

2. Suppose there are m classes C_i, i=1 (1 to m).

$X \in C_i$ if

$P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$

The class for which $P(C_i|X)$is maximized is called the maximum posterior hypothesis.

From Baye's theorem

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

3. P(X) is constant. Only this needs be maximized.

$P(X|C_i)P(C_i)$

- If class prior probabilities not known, then assume all classes to be equally likely
- Otherwise maximize $P(X|C_i)P(C_i)$

$P(C_i) = S_i/S$

4. Naïve assumption: Attribute dependence

$P(X|C_i) = P(x_1, \ldots x_n|C) = PP(x_k|C)$

5. In order to classify an unknown sample X,

Evaluate$P(X|C_i)P(C_i)$ for each class C_i. Sample X is assigned to the class C_i if

$P(X|C_i)P(C_i) > P(X|C_j)\ P(C_j)$ for $1 \leq j \leq m, j \neq i$.

Issues in Predictive Analytics

Though the usage of predictive analytics achieved great results in transforming BigData into business related information and earnings based insights, but the project will not be able to produce desired results unless the challenges faced are adequately met and resolved[14].

Data Quality

The primary factor based on which the entire process of predictive analytics is judged is the "quality of data". The pivotal element in the process is data and it is essential to have a deeper and clear understanding of the quality of data to apply it in the predictive analytics technology[15]. Clearly, the quality of the predictive analytics solution directly depends on the quality of the data. Thus to assure and maintain the virtue of data there is a immediate necessity for the organization to have a system or model. The process for ensuring data quality involves data preparation, cleaning and formatting, which in turn helps data mining[17].

Communicate Outcome

Another issue to consider is the approach to communicate the outcome of the analysis performed. Data scientists and Analysts feel hysterical regarding the insight from the data but when converting the analysis into values they fail to play their part. Hence the organization must take into account the services offered by people with skills and expertise in not only analyzing the data but also exhibiting the information in coercive manner[10].

Privacy and Ownership of data

Data has been playing the vital role in the data mining process from pre-processing to prediction, thus there has always been a conflict between the producers and users of the data on the privacy of the data. Many organizations believe that data should be open, thus the scope for interoperability is expands[18].

Analysis of User data

The primary spotlight in the analysis of user data is to predict the user's motive. This is the major area where the most of the online advertising focuses on, but the user's decision is completely inclusive of fluctuation[18].

Scaling of Algorithms

Data quantity is directly proportional to the efficiency of the predicted result i.e. Having more data is always favourable to data based companies. Due to the recent prominence of the social media huge database repository is being created. "The notable issue associated with scaling of the algorithms is that either the communication or the synchronization overheads go up and so a lot of efficiency can be lost, especially where the estimation doesn't fit greatly into a map/reduce model" [14].

CONCLUSION

Predictive Analytics is the amalgamation of human expertise and proficiency with technology

- people, tools and algorithms are the core of the predictive analytics. Learning the patterns from the historical and current data and the application of algorithms not only to analyse the trends but also to predict the future outcomes is possible because of the above factors[10].

The recent upraise in the field of predictive analytics is mainly due to the BigData, huge volume and abundant data available for research and its application irrespective of the field. But an organization should be well versed in case of why they would require predictive analytics. Imminent step after this is to explain the business requirement and the sort of questions the organizations need to find the answer. Establishment of the technology is the initial step and then comes the important part of testing the applied constraints so that they meet the confined requirements. Another vital motive is to deal with all the challenges and fulfil them, thus extending the output to next scale of advancement.

REFERENCES

1. C. Nyce, "Predictive Analytics," AICPCU-IIA, Pennsylvania, 2007.
2. F. Buytendijk and L. Trepanier, "Predictive Analytics: Bringing The Tools To The Data," Oracle Corporation, Redwood Shores, CA 94065, 2010.
3. D. Mishra, A. K. Das, Mausumi and S. Mishra, "Predictive Data Mining: Promising Future and Applications," *International Journal of Computer and Communication Technology*, **2**(1), pp. 20-28, 2010.
4. T. Bharatheesh and S. Iyengar, "Predictive Data Mining for Delinquency Modeling," ESA/ VLSI, pp. 99-105, 2004.
5. R. Bellazzi, F. Ferrazzi and L. Sacchi, "Predictive data mining in clinical medicine: a focus on selected methods and applications," WIREs Data Mining Knowledge and Discovery, **1**(5), pp. 416-430, 11(2011).
6. P. B. Jensen, L. J. Jensen and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, **13**(6), pp. 395-405, June 2012.
7. M. Kinikar, H. Chawria, P. Chauhan and A. Nashte, "Data Mining in Clinical Practices Guidelines," Global Journal of Computer Science and Technology (GJCST-C), **12**(12)-C, pp. 4-8, 2012.
8. http://www.articlesbase.com/ strategicplanning-articles/ predictiveanalytics-1704860.html.
9. http://en.wikipedia.org/wiki/Predictive_ analytics
10. Ogunleye, J. (2013b) 'Before everyone goes predictive analytics ballistic', available: http:// jamesogunleye.blogspot.co.uk/2013/05/ before-everyone-goes-predictive.html.
11. James ogunleye 'The concepts of predictive analytics', *International journal of knowledge, innovation and entrepreneurship*, **2**(2) pp. 82-90, 2014.
12. Abbott, (2014) Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, New Jersey: John Wiley & Sons.
13. Purvashi mahajan, Abhisekh sharma, "Predictive Analysis of diseases : An overview" , *International journal for research in applied science & engineering technology* (IJRASET), **4**(VI), 2016.
14. http://www.quora.com/Predictive-Analytics/ What-are-the-mostsignificant-challenges-and-opportunities-in-predictive-analytics.
15. Huijbers, C. (2012) What is BigData? [Online] http://clinthuijbers.wordpress.com/2012/08/ 24/ what-is-big-data/; accessed: 26, 2014.
16. McCue, C. (2007) *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*, Butterworth-Heinemann: Oxford, UK.
17. Han, J., Kamber, M., and Pei, J. (2011) *Data Mining Concepts and Techniques* (Third ed). ElsevierInc.: p.6 and 8.
18. Nischol Mishra, Dr. Sanjay silakari, "Predictive Analytics : A Survey, Trends, Applications, Opportunities & Challenges", *International Journal of Computer Science and Information Technologies* (IJCSIT), **3**(3), pp. 4434-4438, 2012.