International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 4, July- August 2017 ISSN 2278-6856

Review of Predictive Analytic Modeling Techniques

Razeef Mohmmad[1], Muheet Ahmed Butt[2], Majid Zaman[3]

[1]Ph.D Scholar, PG Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India

[2] Scientist, PG Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India

[3] Scientist, Information Technology and Support Systems University of Kashmir, Srinagar, J&K, India

Abstract

Data Mining techniques and tools has become important and useful for extracting and manipulating historical data and for establishing relationships and patterns among data in order to generate useful information for decision-making. The proposed research work presents various predictive analytic techniques that are useful for effective decision making and prediction. These techniques are used to improve prediction and decision making in many fields such as: weather prediction, marketing, business, medicine etc. In this work we have discussed four main predictive analytic techniques viz., Classification, Regression, Clustering and Association Analysis and their applications in various domains.

Keywords: Data Mining, Predictive Analytics, Regression, Classification, Clustering and Association Analysis.

1. INTRODUCTION

Nowadays data storage and collection abilities have allowed the accumulation of enormous amounts of data. To handle this huge accumulation of data, tools and techniques has been developed and used that can analyze and manage this huge amount of data into meaningful information [35][36][37]. Data Mining is the technique that retrieves useful information from large amounts of archived data. Data mining is defined as a process involving the extraction of useful and interesting information from the underlying data [1][38][39].

Data mining techniques [29][30][31]has evolved with the time, giving many different analytical solutions in different fields. Data mining techniques are used in prediction and decision making, cost effectiveness and identify new business opportunities. Data mining aids in predictive analysis by providing a record of the past that can be analyzed and used to predict the future events and behavior. Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. It is used to identify the relationship between independent variables and relationship between dependent and independent variables [2]. In search of extracting useful and relevant information from large data sets, predictive analytics derives computational techniques and methods from various fields like statistics, machine learning, database theories, artificial intelligence and pattern recognition. Algorithms actually implemented in data mining originated from these fields [3]. Most of the algorithms used in data mining are borrowed from the fields of artificial intelligence and machine learning. However, algorithms based on Bayesian probabilistic theories and regression analysis, originated a century ago.

The main objective of data mining techniques is to find patterns in data set based on the relationship between data themselves. Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining [32][33]34][that analyze current and historical facts to make predictions about future, or otherwise unknown, events[1][4]. Prediction analysis has vast domain. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. In predictive analytics, various techniques are used to create a predictive model of future behavior. It is used in software engineering and development, weather prediction, actuarial science, marketing and financial services, insurance, telecommunications, retail, travel, healthcare pharmaceuticals, capacity planning and various other fields.

2. LITERATURE REVIEW

Zhenyu Huang et al.,[5], proposed a decision support system that enables continuous monitoring of network performance and turns large amounts data into actionable information. Actual power grid data is used to demonstrate the capability of a proposed decision support system. In this work predictive analytic techniques are used to establish a decision support system for complex network operation management and help operators predict potential network failures and adapt the network in response to adverse situations.

Velmurgun T et al., [6] attempted to analyze performance of K-means and Fuzzy C-means clustering techniques in the field of data mining. The performance compared on the basis of clustering result quality.

Kavitha P., T. Sasipraba, [7], evaluated the performance of distributed data mining framework on Java platform. Association rule mining was used for discovering interesting patterns from a large amount of data.

Yujie Zheng [8] proposed a methodology for clustering in data mining to improve the standard of higher education used to find data segmentation and pattern information.

M.Sukanya et al., [9] used classification and clustering algorithms of data mining for the performance improvement in education sector. By using these algorithms an educational institute could predict the number of enrolled students.

Hu [10] started the implementation of ANN in weather forecasting. He used an adaptive system called Adaline for pattern classification. In 1991 Cook and Wolfe presented a neural network to predict average air temperatures. They used back-propagation learning algorithm for this purpose and got satisfactory result.

Radhika and Shashi [11] have applied Support Vector Machines (SVM's) for weather prediction. Mean square error was taken as a performance measure. It was observed that irrespective of the order, SVM performed better than MLP as root mean square error in SVM was much smaller. It was concluded that parameter selection in all cases of SVM has a significant effect on the performance of model.
Recently ANN technique has been used for rainfall forecasting in Alexandria, Egypt (Shafie et.al., 2011) [12]. Comparison with linear regression models was made on the basis of mean error (BIAS), mean absolute error (MAE), root mean square error (RMSE), and the correlation coefficient (CC) and it was observed that both the models had the ability to predict the extreme values, maximum and minimum values. However, the ability to predict the mid-range values is better for the ANN model.
In commercial fields, business and organizations are deploying sophisticated analytic techniques to evaluate rich data sources, identify patterns within the data and exploit these patterns in decision making [13]. These techniques combine strategic planning procedures with informational technology instruments, summarized under the term "Business Intelligence" [14]. These techniques constitute a well-established process that allows for synthesizing "vast amount of data into powerful decision making capabilities".

2. PREDICTIVE ANALYTIC TECHNIQUES

Today, the term data mining has been watered down so much that vendors and consultants now embrace the term "predictive analytics" or "advanced analytics" or just "analytics" to describe the nature of the tools or services they offer. But even here the terminology can get fuzzy. Not all analytics are predictive [15]. In fact, there are two major types of predictive analytics: supervised learning and unsupervised learning.

In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classifications, Decision Tree, Bayesian Classification, Neural Networks, and Association Rule Mining etc. are common examples of supervised learning [16]. In unsupervised learning, we do not create a model or hypothesis prior to the analysis [17]. We just apply the algorithm directly to the dataset and observe the results. Then a model can be created on the basis of the obtained results. Clustering is one of the examples of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbour have been used for knowledge discovery from large data sets [3]. Whether we use supervised or unsupervised learning, the result is an analytic model. Analysts build models using a variety of techniques, some of which we have already mentioned. Each type of model can be implemented using a variety of algorithms with unique characteristics that are suited to different types of data and problems. Part of the skill in creating effective analytic models is to know which models and algorithms to use. Fortunately, we can now automatically apply multiple models and algorithms to a problem to find the combination that works best. This advance alone has made it possible for non-specialists to create fairly effective analytical models using today's workbenches [15]. Figure 1 shows the hierarchical classification of data mining techniques.

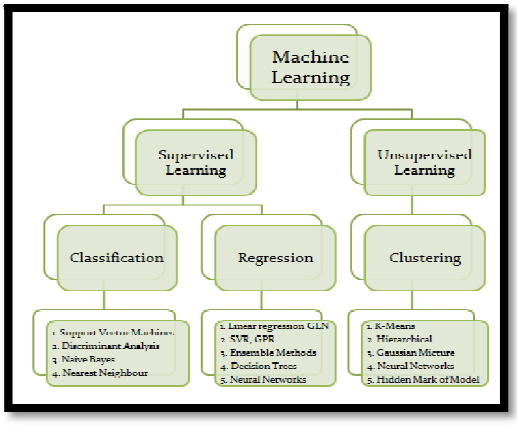

Figure 6 Predictive Analytic Techniques

2.1. Classification

Classification also knows as class prediction is the most commonly used data mining technique to predict the future. In this technique the value of the target or class variable is predicted based on the input variables or predictors. The information from the predictors is used to classify the data samples into two or more distinct classes. The main objective of classification technique is to build a classification model known as classifier (model). The

model or classifier is built from the previously known data sets with predefined class or target values. The developed model is then applied to predict the target or class variable for the input data set. The model does the prediction process by learning the generalized relationship between the predicted target variable with all other input attributes from a known data set [3]. Classification has many applications; some of them are in Credit risk evaluation, Drug discovery, Computer vision, Geostatistics, Speech recognition, Handwriting recognition, Biometric identification, Biometric identification, Biological classification, Statistical natural language processing, Document classification, Internet search engines, Credit scoring, Pattern recognition and weather prediction. For Example, in rainfall prediction, the information from the predictors or independent variables is used to categorize the data samples into two or more distinct classes. Based on various atmospheric variables, the prediction model is developed which will predict the rainfall with appreciable accuracy. The model predicts the Rainfall to class label YES or NO. The various classification algorithms that are widely implemented for prediction include Support Vector Machines (SVM), Decision trees, neural networks, Bayesian models, induction rules and k-nearest neighbors.

2.2. Regression

Sir Francis Galton (Galton, 1888), a relative of Charles Darwin, originally came up with the concept of "regressing toward the mean" while systematically comparing children's heights against their parent's heights by implementing Regression [3]. Regression is also known as function fitting or fitting data with functions. It is a predictive analytic technique in which an equation is used to fit a data set and is mainly used for prediction and forecasting. The simplest form of regression is linear regression [17] used for numeric prediction and logistic regression for classification. In Regression we try to fit a curve/ line to the data points in such a way that the difference distances of data points from the curve or line are minimized [18].

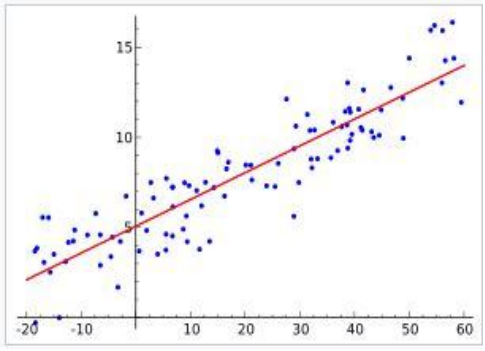

Figure 7 Illustration of linear regression on a data set [28].

The most basic type of regression model is the linear regression model, which can be expressed as:

$$Y = \beta_0 + \beta_1 X + u \qquad [19]$$

The equation shows the relationship between independent variable (X) known as target variable and dependent variable (Y) know as predictor variable. β_1 is the slope, β_0 is intercept and u is error term responsible for the variation not predicted by slope and intercept. Linear regression uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x [20] [21]. Basically a model or a function is used to explain and predict the value of the target variable when the values of the predictor variables are given. The prediction is based on learning from the known data set. Regression technique is found to have many applications some of them are in the field of marketing, finance, banking, weather prediction, sociology, biology, psychology, pharmacology, and engineering.

2.3. Clustering

Clustering or cluster analysis is a data mining technique in which data points/ objects with more similar properties are grouped together into smaller groups known as clusters by applying appropriate clustering algorithms. In other words, clustering arrange data points into subsets in such a way that more similar data points are grouped together, while different data points belong to different groups. Clustering can formally be expressed as a set of subsets $C = C_1, C_2, C_3, ..., C_k$ of S such that: $S = \bigcup_{i=1}^{k} C_i$ and $C_i \cap C_j = \varnothing$ for $i \neq j$. Therefore, any data point in S belongs to exactly one and only one subset [22]. Appropriate clustering algorithms are used for clustering the data objects into similar groups. Partitional clustering and Hierarchical clustering are two basic types of clustering techniques. In Partitional clustering, for a given 'n' number of data points/ data set, clustering algorithms is applied which constructs 'k' partitions of the data set, where each cluster optimizes a clustering criterion, such as the minimization of the sum of squared distance from the mean within each cluster. The main drawback of Partitional clustering algorithms is that they are highly complex and exhaustive in nature. In Hierarchical clustering algorithms, a hierarchy of data points is created either by applying bottom-up approach (agglomerative) or top-down approach (divisive) [23].

Other than partitional and hierarchical clustering algorithms, more clustering techniques have emerged, depending upon the specific problem or data sets available. These techniques include: Density-Based Clustering, Grid-Based Clustering, Model-Based Clustering and Categorical Data Clustering [23]. Though there are many application of clustering but primarily clustering technique is mainly used in finance, genetics, micro-economics, marketing, engineering, weather prediction, document classification, pattern recognition, spatial data analysis, Image processing and more.

2.4. Association Analysis

A data mining technique that measures the strength of co-occurrence between one item and another is known as Association Analysis [3]. It is a rule-based machine learning method for discovering interesting relations between variables in large databases [24]. Unlike Classification and regression which predicts the occurrence of an event, association analysis is concerned with searching of usable patterns in the occurrence of the items. Association analysis is an unsupervised learning approach which is responsible for discovering hidden patterns in archived data, in the form of easily recognizable rules [3]. In other words Association analysis identifies relationships within a set of items/records based on transaction data.

Mathematically, Association Rule can be expressed of the form A → B, where A and B are disjoint item sets, i.e, $A \cap B = \varnothing$. Confidence and Support are used to measure the strength of an Association Rule. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in B appear in transactions that contain A. The formal definitions of these metrics are [25]:

$$\text{Support, } S(A \to B) = \frac{\sigma(A \cup B)}{N}$$

$$\text{Confidence, } C(A \to B) = \frac{\sigma(A \cup B)}{\sigma(A)}.$$

Association rule mining algorithms adopts a general approach by which a problem is decomposed into two main subtasks as: Frequent Item-set Generation and Rule Generation. The objective of the Frequent Item-set Generation is to find all the item-sets that minsup threshold; these item-sets are known as frequent item-sets. In Rule Generation the objective is to extract all high-confidence rules from frequent data item-sets found in the previous step, these rules are called strong rules [25]. Apriori algorithm, ECLAT (Equivalence Class Transformation) and FP (frequent pattern)-Growth algorithm are some well known algorithms that are proposed to generate association rules. These algorithms are responsible to efficiently find the frequent item sets from the pool of all possible item sets. The Apriori algorithm leverages some simple logical principles on the lattice item sets to reduce the number of item sets to be tested for the support measure (Agrawal & Srikant, 1994). According to Apriori principle "If an item set is frequent, then all its subset items will be frequent." (Tan et al, 2005). The item set is "frequent" if the support for the item set is more that support threshold [3]. Market Basket Analysis is one of the widely used techniques of Association Rule Mining. For example, a database of a departmental store consists of huge amount of transaction records collected through barcode scanner while purchasing. Each record contains list of items bought by a customer for a single transaction. Managers would be interested to know if certain groups of items are consistently purchased together. By applying Market Basket Analysis, the departmental store manager will use the huge amount of archived data to adjust store layouts and design by placing the items close to each other for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns [26]. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns [26]. Association Rule Mining is mainly used to improve decision making in a wide variety of applications such as: market basket analysis, medical diagnosis, bio-medical literature, protein sequences, census data, logistic regression, and fraud detection in web, CRM of credit card business etc [27].

3. CONCLUSION

Predictive analytic technique is an approach through which hidden and useful patterns are extracted from the huge archived data which will help in better prediction and decision making. . Predictive analytic algorithms help in the process of extraction and prediction. There are vast number of algorithms that are applied to improve decision making and prediction. In this work, we have mentioned four main categories of the predictive analytic techniques. It can be concluded that predictive analytic techniques have vast application domain.

References

[1] Nyce, Charles (2007), Predictive Analytics White Paper (PDF), American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1

[2] Stevenson, Erin "Tech Beat: Can you pronounce health care predictive analytics?", Times- Standard, December 16, 2011.

[3] Kotu, Vijay, and Bala Deshpande. Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann, 2014.

[4] Eckerson, Wayne (May 10, 2007), Extending the Value of Your Data Warehousing Investment, The Data Warehouse Institute.

[5] Zhenyu Huang, Pak Chung Wong, Patrick Mackey, Yousu Chen, Jian Ma, Kevin Schneider, and Frank L. Greitzer, "Managing Complex Network Operation with Predictive Analytics," in Proceedings AAAI Spring Symposium on Techno social Predictive Analytics, pp. 59–65, 2009.

[6] Velmurugan T. et al., "Performance Evaluation of K-Means & fuzzy C-means Clustering Algorithm for Statistical Distribution of Input Data Points", European Journal of Scientific Research, Vol. 46, 2010.

[7] Kavitha P., T. Sasipraba, "Performance Evaluation of Algorithms using a Distributed Data Mining Framework based on Association Rule Mining", International Journal on Computer Science & Engineering (IJCSE), 2011.

[8] YujieZheng, "Clustering Methods in Data Mining with its Application in Higher Education", International

Conference on Education Technology and Computer, Vol. 43, 2012, IACSIT Press, Singapore.
[9] M.Sukan et al., "Data Mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm", International Conference of Computing and Control Engineering (ICCCE) 12-13 April, 2012
[10] D.F. Cook and M. L. Wolfe, "A back-propagation neural network to predict average air temperatures", AI Applications, 5, 40-46, 1991.
[11] Y. Radhika and M. Shashi (2009), Atmospheric Temperature Prediction using Support Vector Machines, International Journal of Computer Theory and Engineering, Vol. 1, 1793-8201
[12] El-Shafie, A. H. , El-Shafie, A., El Mazoghi, H. G., Shehata, A. and Taha, M. R. (2011), Artificial neural network technique for rainfall forecasting applied to Alexandria, Egypt, International Journal of the Physical Sciences 6, 1306-1316
[13] Chaudhuri, S., Dayal, U., and Narasayya, V. 2011. "An Overview of Business Intelligence Technology," Communications of the ACM (54:8), pp. 88-98.
[14] Eckerson, W. W. (2006). Performance dashboards: Measuring, monitoring, and managing your business. Hoboken, NJ: John Wiley & Sons.
[15] Eckerson Wayne. W" Extending the Value of Your Data Warehousing Investment" TDWI best practices report, 2007.
[16] GargSumit and Sharma Arvind K," Comparative Analysis of Data Mining Techniques on Educational Dataset", International Journal of Computer Applications (0975 – 8887), Volume, 74– No.5, July 2013
[17] Jiawei Han and MichelireKamber, "Data Mining Concept and Technique", Published by Morgan Kaufman, 2006.
[18] https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
[19] Campbell, Dan, and Sherlock Campbell. "Introduction to regression and data analysis," StatLap Workshop Series. 2008.
[20] Manisha rathi Regression modeling technique on data mining for prediction of CRM CCIS 101, pp.195-200,2010Springer–Verlag Heidelberg 2010.
[21] Duda, R. O. and Hart, P. E. (1973). "Pattern Classification and Scene Analysis", John Wiley & Sons.
[22] Maimon, Oded, and Lior Rokach, eds. Data mining and knowledge discovery handbook. Vol. 2, pp 332, New York: Springer, 2005.
[23] Andritsos, Periklis. "Data clustering techniques." Rapport technique, "University of Toronto, Department of Computer Science," pp 8-9, (2002).
[24] https://en.wikipedia.org/wiki/Association_rule_learning.
[25] Tan, Pang-Ning. Introduction to data mining. Pearson Education India, 2006.
[26] http://www.resample.com/xlminer/help/Assocrules/associationrules_intro.htm.
[27] Rajak, Akash, and Mahendra Kumar Gupta. "Association rule mining: applications in various areas." Proceedings of International Conference on Data Management, Ghaziabad, India. 2008.
[28] https://en.wikipedia.org/wiki/Regression_analysis
[29] Butt, Muheet Ahmed, and Majid Zaman. "Assessment Model based Data Warehouse: A Qualitative Approach." International Journal of Computer Applications 62.10 (2013).
[30] Zaman, Majid, and Muheet Ahmed Butt. "Enterprise Data Backup & Recovery: A Generic Approach." International Organization of Scientific Research Journal of Engineering (IOSRJEN) (2013): 2278-4721.
[31] Butt, Muheet Ahmed. "Implementing ICT Practices of Effective Tourism Management: A Case Study." Journal of Global Research in Computer Science 4.4 (2013): 192-194.
[32] Butt, Er Muheet Ahmed, S. M. K. Quadri, and Er Majid Zaman. "Star Schema Implementation for Automation of Examination Records." Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.
[33] Khan, Sajad Mohammad, Muheet Ahmed Butt, and Majid Zaman Baba. "ICT: Impacting Teaching and Learning." International Journal of Computer Applications 61.8 (2013).
[34] Zaman, M., S. M. K. Quadri, and Er Muheet Ahmed Butt. "Information Integration for Heterogeneous Data Sources." IOSR Journal of Engineering 2.4 (2012): 640-643.
[35] Butt, M. A., and M. Zaman. "Data quality tools for data warehousing: enterprise case study." IOSR Journal of Engineering 3.1 (2013): 75-76.
[36] Zaman, Majid, and Muheet Ahmed Butt. "Enterprise Management Information System: Design & Architecture." International Journal of Computational Engineering Research (IJCER), ISSN 2250 (2013): 3005.
[37] Butt, Muheet Ahmed. "Information extraction from pre-preprinted documents." Energy 20.8 (2012): 729-743.