A REVIEW ON PREDICTIVE ANALYTICS IN DATA MINING

Kavya.V[1], Arumugam.S[2]

[1]M.E.Scholar, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India

[2]Professor, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India

ABSTRACT

The data mining its main process is to collect, extract and store the valuable information and now-a-days it's done by many enterprises actively. In advanced analytics, Predictive analytics is the one of the branch which is mainly used to make predictions about future events which are unknown. Predictive analytics which uses various techniques from machine learning, statistics, data mining, modeling, and artificial intelligence for analyzing the current data and to make predictions about future. The two main objectives of predictive analytics are Regression and Classification. It is composed of various analytical and statistical techniques used for developing models which predicts the future occurrence, probabilities or events. Predictive analytics deals with both continuous changes and discontinuous changes. It provides a predictive score for each individual (healthcare patient, product SKU, customer, component, machine, or other organizational unit, etc.) to determine, or influence the organizational processes which pertain across huge numbers of individuals, like in fraud detection, manufacturing, credit risk assessment, marketing, and government operations including law enforcement.

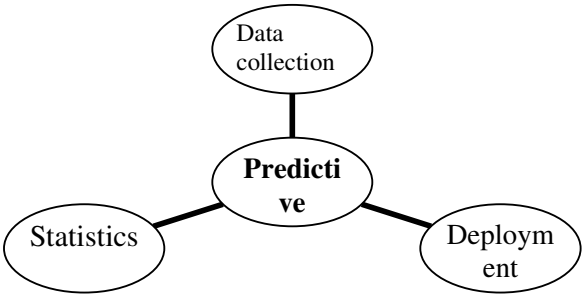
KEYWORDS

Predictive analytics, Credit history, forecasting, Regression techniques

1. INTRODUCTION

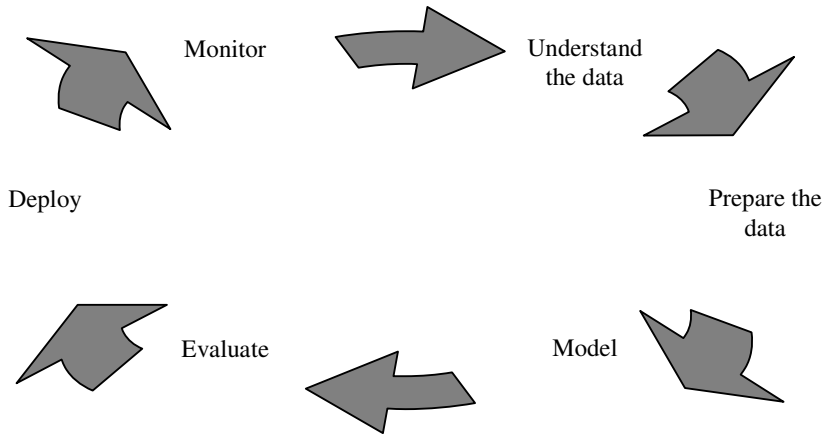
Large amount of data available in information databases becomes waste until the useful information is extracted. Predictive analytics is the roof of advanced analytics which is to predict the future events. Predictive analytics is capsuled with the data collection and modelling, statistics and deployment. The figure 1 gives the basis of predictive analysis. Based on the availability of high quality data and effective sharing, the success of data mining relies.

DOI:10.5121/ijccms.2016.5301



The figure 1 gives the basis of predictive analysis.

The business users allow discovering the predictive intelligence by uncovering patterns and relationships in both the structured and unstructured data through the data mining and text analytics along with statistics. The structured data like gender, age, income .etc. Unstructured data are social media and they are extracted data used in the model building process. The figure 2 explains the cycle of predictive analytics.



The figure 2 explains the cycle of predictive analytics.

2. OVERVIEW

PREDICTIVE ANALYTICS

Predictive analytics encapsulated with statistical techniques from predictive modelling, data mining and machine learning which are used to analyse the current and historical facts to found the predictions about future [15]. In business, predictive analytics are used to identify risks and opportunities. Predictive analytics are used in various fields such as marketing, finance, travel and health care [10].

REGRESSION TECHNIQUES

A data mining function is regression which predicts a number. Regression techniques are used to predict the age, weight, distance, and temperature [7]. In regression task starts with a dataset in that

the target values are known. Common applications of regression are trend analysis, biomedical and financial forecasting [4]. There were various regression algorithms which are generalized linear models and support vector machines [13].

FORECASTING

Forecasting focus towards the predictions of the future based past and present data [3, 14]. The two terms are focussed on the forecasting are risk and uncertainty.

3. LITERATURE SURVEY

Carlos Márquez-Vera et al [1] A genetic programming algorithm and different data mining approaches are proposed for solving challenge due to the high number of factors that can affect the low performance of students and the imbalanced nature. ***Methodologies*** used to resolve are Data Gathering, Pre-Processing, Data Mining, and Interpretation. Interpretable Classification Rule Mining (ICRM) and SMOTE (Synthetic Minority Over-sampling Technique) algorithms are used. Student's data set from where data's are collected.WEKA tool is used. Accuracy, True positive rate, True negative rate and Geometric mean are the parameters for performance measurement. It results in the accurate and comprehensible classification rules and it achieved the best predictions of student failure (98.7 %).

Branko G. Celler et al [2] The telemonitoring of vital signs from the home for the management of patients with chronic conditions. New measurement modalities and signal processing techniques are proposed for increasing the ions about future.quality and value of vital signs monitoring. Automated risk stratification algorithm is used. QRS height and width (ECG), PR (Pressure Rate), RR (Respiratory Rate), QT (Abnormal heart rate) and Body temperature are the parameters. jBoss tool is used. Data's are taken from Department of Human Services Medicare databases and Telemonitoring system data. It results in identifying the key technical performance characteristics of at-home telemonitoring systems.

Hao-Tsung Yang et al [3] A neural network model to predict the value of playing style information in predicting match quality. A mix of Sternberg's thinking style theory and individual histories are used to categorize *League of Legend (LoL)* players. The data's are collected from LoLBase website. The algorithms used are Feed forward/ feedback propagation algorithm and Match making algorithm. Win rate, Match duration, Group number (ELO) are the parameters for performance measurement. PYTHON is used. It finds that the presence of global-liberal (G-L) style players is positively correlated with match enjoyment.

Jacob R. Scanlon et al [4] To forecast the daily level of cyber-recruitment activity of VE (violent extremist) groups. LDA-based Topics as predictors within time series models reduce forecast error. Latent Dirichlet allocation (LDA) algorithm is used in forecasting. Western jihadist discussion forum is used as dataset. Number of posts and % of recruitment post (per day) are the parameters for measurement. *RTextTools* and *tm* text mining packages in R are used. It achieves the automatic forecast of VE cyber recruitment using natural language processing, supervised machine learning, and time series analysis.

Quanzeng You, Liangliang Cao et al [5] To build a reliable forecasting system for the elections and its modelled to figure out the inter relationship between social multimedia as image-centric and real-world entities. Competitive Vector Auto Regression (CVAR) algorithm is used. By the competition mechanism, CVAR compares the popularity among multiple competing candidates. Data's are taken from Flickr. Number of images, Number of users, images uploaded per day (IPD) and users uploading images per day (UPD) are the parameters for measurement. OpenCV Tool is used. As a

result CVAR is able to take prior knowledge which leads to better performance in terms of prediction accuracy.

Sean M. Arietta et al [6] A method which found and evaluate automatically to figure out predictive relationships between the optic presence of a city and its non-visual attributes. Scalable distributed processing framework is implemented that speeds up the main computational barrier by an order of magnitude. Algorithms used are hard negative mining Algorithm, classification and recognition algorithm and Dijkstra's shortest path algorithm. Datasets are from 10,000 Google StreetView panorama projections (2,000 positives, 8,000 negatives) Training dataset. The parameters for performance measurements are Number of Detections, % of Detections from Positive Set. MATLAB is used. As a result it's used to define the visual boundary of city neighbourhoods, generate walking directions that prohibit or find exposure to city attributes and validate user-specified visual elements for prediction.

Abish Malik, Ross Maciejewski et Al [7] A visual analytics approach that provides decision makers with a proactive and predictive environment which helps them in making effective resource allocation and deployment decisions. Analysts are provided with a suite of natural scale templates and methods that enable them to focus and drill down to appropriate geospatial and temporal resolution levels. Prediction algorithm is used. Accuracy, Average, Count and Time are the parameters for performance measurement. This Methodology is applied to Criminal, Traffic and Civil (CTC) incident datasets. It provides users with a suite of natural scale templates that support analysis at multiple spatiotemporal granularity levels.

Ronaldo C. Prati et al [8] Various graphical performance evaluation methods are increasingly drawing the consideration of data mining. Ability to depict the trade-offs between evaluation aspects in a multidimensional area. Graphical evaluation methods are applicable for binary classification problems. The predictive models deployed on Classification, Ranking, Probability estimation. The parameters for performance measurements are True positive and false positive rate, precision and recall. The Insurance Company (TIC).Benchmark data set includes 86- variables are used and the tool used is WEKA- 3.5.8 version. ROC graph, ROC curve, cost lines, cost curve Precision-recall curves, lift curve, Reliability diagram, ROI curve, Discrimination diagram and attribute diagram are different graphical evaluation methods. It helps on deciding the methods which is well fitted for the situations.

Gang Fang, Gaurav Pandey et al [9]Discriminative patterns can provide valuable insights into data sets with class labels. Low-support patterns that can be discovered using SupMaxPair. Per pattern precision, Density, dimension, count and Frequency are the parameters for performance measurement. Frequent pattern mining algorithm and discriminative pattern mining algorithms are used. Synthetic and cancer gene expression data sets are used for prediction. This result in exploring discriminative patterns by speculating patterns with relatively low support from dense and high-dimensional data sets comparably the other approaches fall to explore within desired amount of time.

Nanlin Jin, Peter Flach et al [10] Data mining methods for exploring incredible consumption patterns and their associated descriptive models from smart electricity meter data. Target concept, Target type, Double regression, Coverage and Strategy type are the parameters for performance measurement. Subgroup discovery algorithms and S-Transform algorithms are used. Data used were collected by the Energy Demand Research Project (EDRP). Cortana tool is used. This approach outperforms more conventional data mining methods in terms of their predictive power and classification accuracy, while consuming similar computational resource.

4. COMPARISONS ON DIFFERENT PREDICTIVE ANALYTIC TECHNIQUES

Title	Techniques And Algorithms	Datasets	Parameter	Conclusion
Predicting Student Failure At School Using Genetic Programming And Different Data Mining Approaches With High Dimensional And Imbalanced Data	Data Gathering, Pre-Processing, Data Mining, And Interpretation. Interpretable Classification Rule Mining (ICRM) And SMOTE (Synthetic Minority Over-Sampling Technique) Algorithms	Student's Data Set	Accuracy, True Positive Rate, True Negative Rate And Geometric Mean	Accurate And Comprehensible Classification Rules
Home Telemonitoring Of Vital Signs Technical Challenges And Future Directions	Automated Risk Stratification Algorithm	Department Of Human Services Medicare Databases And Telemonitoring System Data	QRS Height And Width (ECG), PR (Pressure Rate), RR (Respiratory Rate), QT (Abnormal Heart Rate) And Body Temperature	Identifying The Key Technical Performance Characteristics Of At-Home Telemonitoring Systems.
Thinking Style And Team Competition Game Performance And Enjoyment	Feed Forward/ Feedback Propagation Algorithm And Match Making Algorithm.	Lolbase Website	Win Rate, Match Duration, Group Number (ELO)	Finds That The Presence Of Global-Liberal (G-L) Style Players Is Positively Correlated With Match Enjoyment
Forecasting Violent Extremist Cyber Recruitment	Latent Dirichlet Allocation (LDA) Algorithm	Western Jihadist Discussion Forum	Number Of Posts And % Of Recruitment Post (Per Day)	Achieves The Automatic Forecast Of VE Cyber Recruitment Using Natural Language Processing, Supervised Machine Learning, And Time Series Analysis
A Multifaceted Approach To Social	Competitive Vector Auto Regression (CVAR) Algorithm	Flickr	Number Of Images, Number Of Users, Images	Better Performance In Terms Of

Multimedia-Based Prediction Of Elections			Uploaded Per Day (IPD) And Users Uploading Images Per Day (UPD)	Prediction Accuracy.
City Forensics: Using Visual Elements To Predict Non-Visual City Attributes	Hard Negative Mining Algorithm, Classification And Recognition Algorithm And Dijkstra's Shortest Path Algorithm	Google Streetview Panorama Projections Training	Number Of Detections, % Of Detections From Positive Set	Define The Visual Boundary Of City Neighbourhoods, Generate Walking Directions That Prohibit Or Find Exposure To City Attributes And Validate User-Specified Visual Elements For Prediction
Proactive Spatiotemporal Resource Allocation And Predictive Visual Analytics For Community Policing And Law Enforcement	Prediction Algorithm	Criminal, Traffic And Civil (CTC) Incident Datasets.	Accuracy, Average, Count And Time	It Provides Users With A Suite Of Natural Scale Templates That Support Analysis At Multiple Spatiotemporal Granularity Levels.
A Survey On Graphical Methods For Classification Predictive Performance Evaluation	Graphical Evaluation Methods	Insurance Company (TIC).Benchmark Data Set	True Positive And False Positive Rate, Precision And Recall	It Helps On Deciding The Methods Which Is Well Fitted For The Situations.
Mining Low-Support Discriminative Patterns From Dense And High-Dimensional Data	Frequent Pattern Mining Algorithm And Discriminative Pattern Mining Algorithms	Synthetic And Cancer Gene Expression Data Sets	Per Pattern Precision, Density, Dimension, Count And Frequency	Exploring Discriminative Patterns By Speculating Patterns With Relatively Low Support From Dense And High-Dimensional Data Sets Comparably The Other

				Approaches Fall To Explore Within Desired Amount Of Time.
Subgroup Discovery In Smart Electricity Meter Data	Subgroup Discovery Algorithms And S-Transform Algorithms	Energy Demand Research Project (EDRP)	Target Concept, Target Type, Double Regression, Coverage And Strategy Type	Outperforms More Conventional Data Mining Methods In Terms Of Their Predictive Power And Classification Accuracy, While Consuming Similar Computational Resource.

5. CONCLUSION

Predictive analytics is the future of data mining .This study focus towards the predictive analytics, regression techniques and forecasting in knowledge discovery domain. Business intelligence is used in predictive analytics for modelling and forecasting. Predictive analytics are more efficient in choosing marketing methods and helpful in social media analytics.

REFERENCES

[1] Carlos Márquez-Vera, Alberto Cano, Cristóbal Romero, Sebastián Ventura,"Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data", Springer Science+Business Media, LLC 2012.

[2] Branko G. Celler, and Ross S. Sparks, "Home Telemonitoring of Vital Signs Technical Challenges and Future Directions", IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 1, January 2015.

[3] Hao Wang, Hao-Tsung Yang, and Chuen-Tsai Sun," Thinking Style and Team Competition Game Performance and Enjoyment", IEEE Transactions On Computational Intelligence And Ai In Games, Vol. 7, No. 3, September 2015.

[4] Jacob R. Scanlon and Matthew S. Gerber, "Forecasting Violent Extremist Cyber Recruitment", IEEE Transactions On Information Forensics And Security, Vol. 10, No. 11, November 2015.

[5] Quanzeng You, Liangliang Cao, Yang Cong, Xianchao Zhang, and Jiebo Luo" A Multifaceted Approach to Social Multimedia-Based Prediction of Elections", IEEE Transactions On Multimedia, Vol. 17, No. 12, December 2015.

[6] Sean M. Arietta Alexei A. Efros Ravi Ramamoorthi Maneesh Agrawala, "City Forensics: Using Visual Elements to Predict Non-Visual City Attributes", IEEE Transactions On Visualization And Computer Graphics, Vol. 20, No. 12, December 2014.

[7] Abish Malik, Ross Maciejewski, Sherry Towers, Sean McCullough, and David S. Ebert," Proactive Spatiotemporal Resource Allocation and Predictive Visual Analytics for Community Policing and Law Enforcement", IEEE Transactions On Visualization And Computer Graphics, Vol. 20, No. 12, December 2014.

[8] Ronaldo C. Prati, Gustavo E.A.P.A. Batista, and Maria Carolina Monard," A Survey on Graphical Methods for Classification Predictive Performance Evaluation", IEEE Transactions On Knowledge And

Data Engineering, Vol. 23, No. 11, November 2011.

[9] Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta, Michael Steinbach, and Vipin Kumar," Mining Low-Support Discriminative Patterns from Dense and High-Dimensional Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 2, February 2012 .

[10] Nanlin Jin, Peter Flach, Tom Wilcox, Royston Sellman, Joshua Thumim, and Arno Knobbe," Subgroup Discovery in Smart Electricity Meter Data", IEEE Transactions On Industrial Informatics, Vol. 10, No. 2, May 2014.

[11] Hao Wang, Hao-Tsung Yang, and Chuen-Tsai Sun," Thinking Style and Team Competition Game Performance and Enjoyment", IEEE Transactions On Computational Intelligence And Ai In Games, Vol. 7, No. 3, September 2015.

[12] Quanzeng You, Liangliang Cao, Yang Cong, Senior Member, IEEE, Xianchao Zhang, and Jiebo Luo, Fellow, IEEE." A Multifaceted Approach to Social Multimedia-Based Prediction of Elections", IEEE Transactions On Multimedia, Vol. 17, No. 12, December 2015.

[13] Yun Wang and Sudha Ram," Predicting Location- Based Sequential Purchasing Events by Using Spatial, Temporal, and Social Patterns", IEEE Intelligent Systems, May/June 2015.

[14] Jesse Rio Russell," Predictive analytics and child protection: Constraints and Opportunities", Child Abuse & Neglect 46 (2015) 182–189- ELSEVIER.

[15] Karel Dejaeger, Wouter Verbeke, David Martens, and Bart Baesens," Data Mining Techniques for Software Effort Estimation: A Comparative Study", IEEE Transactions On Software Engineering, Vol. 38, No. 2, March/April 2012.

[16] Leonardo Feltrin," KNIME an Open Source Solution for Predictive Analytics in the Geosciences", IEEE Geoscience and remote sensing magazine, December 2015.

[17] Josep Ll. Berral, Nicolas Poggi, David Carrera, Aaron Call, Rob Reinauer, Daron Green," ALOJA: A Framework for Benchmarking and Predictive Analytics in Big Data Deployments", IEEE Transactions on Emerging Topics in Computing • November 2015.

[18] Minghui Zhou and Audris Mockus," Who Will Stay in the FLOSS Community? Modeling Participant's Initial Behavior", IEEE Transactions On Software Engineering, Vol. 41, No. 1, January 2015 .

[19] Sean M. Arietta Alexei A. Efros Ravi Ramamoorthi Maneesh Agrawala, "City Forensics: Using Visual Elements to Predict Non-Visual City Attributes", IEEE Transactions On Visualization And Computer Graphics, Vol. 20, No. 12, December 2014.

[20] Francisco C. Pereira, Member, IEEE, Filipe Rodrigues, Evgheni Polisciuc, and Moshe Ben-Akiva", Why so many people? Explaining Nonhabitual Transport Overcrowding With Internet Data", IEEE Transactions On Intelligent Transportation Systems, Vol. 16, No. 3, June 2015.