International Journal of Computer and Communication Technology

Volume 2 | Issue 3 Article 5

July 2011

Predictive Data Mining: Promising Future and Applications

Debahuti Mishra
Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India, debahuti@iter.ac.in

Asit Kumar Das
Department of Computer Applications Institute of Technical Education & Research Siksha O Anusandhan University Bhubaneswar, Odisha, India, asitdasbbsr@gmail.com

Mausumi Mausumi
Department of Information Technology Institute of Technical Education & Research Siksha O Anusandhan University Bhubaneswar, Odisha, India, mausumi.singh@gmail.com

Sashikala Mishra
Department of Computer Sc. & Engineering Institute of Technical Education & Research Siksha O Anusandhan University Bhubaneswar, Odisha, India, Sashi.iter@gmail.com

Predictive Data Mining: Promising Future and Applications

Debahuti Mishra *

Department of Computer Sc. & Engineering
Institute of Technical Education & Research
Siksha O Anusandhan University
Bhubaneswar, Odisha, India
debahuti@iter.ac.in
* Corresponding author

Asit Kumar Das

Department of Computer Applications
Institute of Technical Education & Research
Siksha O Anusandhan University
Bhubaneswar, Odisha, India
asitdasbbsr@gmail.com

Mausumi

Department of Information Technology
Institute of Technical Education & Research
Siksha O Anusandhan University
Bhubaneswar, Odisha, India
mausumi.singh@gmail.com

Sashikala Mishra

Department of Computer Sc. & Engineering
Institute of Technical Education & Research
Siksha O Anusandhan University
Bhubaneswar, Odisha, India
Sashi.iter@gmail.com

Abstract: Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. For example, an insurance company is likely to take into account potential driving safety predictors such as age, gender, and driving record when issuing car insurance policies. Multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to forecast future probabilities with an acceptable level of reliability. In predictive modeling, data is collected, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. Predictive analytics are applied to many research areas, including meteorology, security, genetics, economics, and marketing. In this paper, we have done an extensive study on various predictive techniques with all its future directions and applications in various areas are being explained

Key words: Data Mining; Prediction; Predictive Analytics; Multiple Predictors; Predictive Modeling

1. Introduction

Predictive Analytics [1] [2] is the process of dealing with variety of data and apply various mathematical formulas to discover the best decision for a given situation. Predictive analytics gives company a competitive edge. It is the decision science that removes guesswork out of the decision-making process and applies proven scientific guidelines to find right solution in the shortest time possible. Predictive analytics is a solution used by many businesses today to gain more value out of

large amounts of raw data by applying techniques that are used to predict future behaviors within an organization, its customer base, its products and services. Predictive analytics encompasses a variety of techniques from data mining, statistics and game theory that analyze current and historical facts to make predictions about future events.

The term Predictive data mining is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest [8]. For example, a credit card company may want to engage in predictive data mining, to derive a (trained) model or set of models (e.g., neural networks, meta-learner) that can quickly identify transactions which have a high probability of being fraudulent. Other types of data mining projects may be more exploratory in nature (e.g., to identify cluster or segments of customers), in which case drill-down descriptive and exploratory methods would be applied. Data reduction is another possible objective for data mining. Business metrics do a great job summarizing the past [5]. But if you want to predict how customers will respond in the future, there is one place to turn - *predictive analytics*. By learning from your abundant historical data, predictive analytics provides the marketer something beyond standard business reports and sales forecasts: actionable predictions for each customer. These predictions encompass all channels, both online and off, foreseeing which customers will buy, click, respond, convert or cancel. The customer predictions generated by predictive analytics deliver more relevant content to each customer, improving response rates, click rates, buying behavior, retention and overall profit [6][8]. For online applications such as e-marketing and customer care recommendations, predictive analytics acts in real-time, dynamically selecting the ad, web content .or cross-sell product.

The layout of rest of this paper is as follows: section 2 gives preliminary concepts of predictive analytics techniques, related work on predictive mining is given in section 3 finally section 4 gives the conclusion and future work.

2. Preliminaries

2.1 Predictive Analytics Techniques

Linear regression [40] is the first kind of regression analysis to be studied and applied in various practical applications like epidemiology, environmental science and finance. In a very lucid term, linear regression otherwise called as straight line regression analysis is a regression to estimate the unknown effect of changing one variable over another. Specifically, it models Y as a linear function of X i.e. how much Y changes when X changes one unit. So it is expressed as a straight line equation:

$$Y = b + wX \tag{1}$$

Here b and w are the regression coefficients where b is the Y intercept and w is the slope of the line. In cases, the coefficients can be assumed to be weights, where:

$$Y = w_0 + w_1X \tag{2}$$

This can be solved for the coefficients by method of least squares so as to minimize the error between the actual data and the estimated data. A training data set D, consisting of several predictor variable X and response variable Y,

$$|D| = \{ (X_1 , Y_1), (X_2 , Y_2)....(X_{|D|} , Y_{|D|}) \} \tag{3}$$

The estimated regression coefficient is given as:

$$w_1 = \sum_i (X_i - Mean_X)(Y_i - Mean_Y) / \sum_i (X_i - Mean_X)^2 \quad where\ i = 1\ to\ |D|, \tag{4}$$
$$w_0 = Mean_Y - w_1.Mean_x \tag{5}$$

Linear regression model identifies the relationship between a single predictor variable X_i and the response variable Y when all other predictor variables in the model are "held fixed". This is called as the *unique effect* X_i on Y.

Multiple linear regression (MLR) [30] is a mathematical technique that uses a number of variables to predict some unknown variable. It is a study on the relationship between a single dependent variable and one or more independent variables. This model describes a dependent variable Y by independent variable $X_1, X_2 ... X_p\ (p>1)$ is expressed by the equation as ,

$$Y = \alpha + \sum_k \beta_k X_k + \epsilon \tag{6}$$

Where $\alpha, \beta_k\ (k = 1,2...p)$ are the parameters and ϵ is the error term. MLR combines the idea of correlation and linear regression. Basically it is a multivariate statistical technique for examining the linear correlations between two or more independent (IVs) and a single dependent variable (DV). Generalized Linear Regression (GLR) [31] is a class of linear model that postulates the linear regression model in a view that, it allows data to come from a distribution that is a member of the exponential family of distribution. This approach utilizes a set of continuous explanatory variables to model an exponential family response. This regression model unifies other modeling techniques to estimate the model parameters. The mean μ of an exponential distribution is given by,

$$\mu = E(Y) = g^{-1}(X\beta) \tag{7}$$

Where Y being the outcome i.e. response variable, X as the predictor and β being an unknown parameter vector. $E(Y)$ is the expected value of Y and $X\beta$ is the linear predictor, g is a monotone, twice differentiable link function. A generalized linear regression is stated:

$$g(\mu) = \sum_i X_i \beta_i \tag{8}$$

β is estimated by maximum likelihood. In GLR the variance of the response variable Y, is a function of the mean value of Y. Unlike linear regression, the variance of Y is constant. The common types of generalized linear regression include **Logistic regression** and **Poission regression**.

Logistic regression [20] [24] is a type of predictive model that can be used when the target variable is categorical variable that has exactly two categories like, win game/doesn't win, live/die. Technically it can be said as logistic regression is used for binomial regression. Simultaneously it also applies to continuous target variable that models the probability of some event occurring as a linear function of a set of predictor variable. Due to this it has extensively applied in the field of medical sciences, marketing application and social sciences. Mathematically,

$$f(Z) = e^Z / (e^Z + 1) = 1 / (1+e^{-Z}) \tag{9}$$

Z is called as the *logit*, exposure to some set of independent variable $f(Z)$ is probability of a particular outcome. Logistic regression takes Z as input and outputs $f(Z)$ i.e. it can take input as any value from negative to positive infinity and give output between 0 and 1.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k \tag{10}$$

Here β_0 is the intercept and $\beta_1, \beta_2 ... \beta_k$ are the regression coefficients.

Poission regression [41] is yet another form of regression where the dependent variable is a *count* where it limits the logistic regression in the terms of binomial distribution. This means the number of

trials becomes very large while the expectation remains stable i.e. the probability of success is comparatively small. Poission regression is referred as log linear model if it assumes the response variable *Y* that has a poission distribution and takes the logarithm of its expected value that can be modeled by a linear combination of unknown parameters. The logarithm is taken as the linear function when poission regression models contingency tables.

$$Log\ (E\ (Y \mid x)) = a + bx \tag{11}$$

The value of *a* and *b* can be estimated by maximum likelihood.

Non linear regression [42] is characterized by the fact that the prediction equation depends non-linearly on one or more unknown parameter unlike linear regression that is used to build a purely empirical model. While adapting linear least squares in linear regression method we usually fit a straight line or a flat plane to a bunch of data points. Sometimes the relationship that is to be modeled is more of a curved one rather than flat. For something growing exponentially, the relation between *X* and *Y* is a curve. So to fit into the case, non linear regression creates new variables by applying transformations to the variables. It is whenever possible desired to convert the non linear model into linear one that can then be solved by the method of least squares. Usually before applying regression it is common to perform attribute subset selection to eliminate attributes that are unlikely to be good predictors.

One of the very common forms of linear regression is,

$$Y_i = f(X_i, \theta) + \epsilon_i \ ; i = 1....n \tag{12}$$

Where Y_i are responses, f is known function of the covariate vector $X_i = (X_{i1}, ...,X_{ik})^T$ and the parameter $\theta = (\theta_1... \theta_p)^T$ and ϵ_i are random errors.

Regression tree [26] [27] is a component of the CART system. At times regression trees are considered as a variant of decision trees since it is capable of approximating real valued functions. It takes as input a mixture of continuous and categorical variables and outputs a numerical variable. The tress is basically built by a process of recursive partitioning that is, it splits the data iteratively into partitions and again splits them further at each of the branch. Initially the training set are together and then the algorithm chooses a split point that partitions the data and then tries breaking the training set in such a way that its minimizes the sum of the squared deviation from the mean in the separate parts. The splitting or partitioning process continues until each node becomes terminal node i.e. the sum of the squared deviations from the mean for a node becomes zero. Regression tress make prediction fast in the view that it does not involve any complicated calculations and even easy to understand. Just by looking at the tree we can infer which of the variable are making the prediction. In some cases if there are some missing data then we might not be able to go further down the tree to a leaf but we can still predict by averaging all the leaves that are reachable in the sub tree.

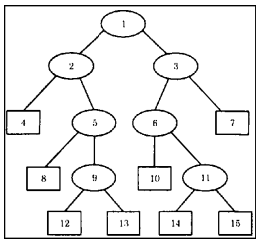

FIGURE 1: The figure shows an outline of a regression tree where the splitting occurs at the node 1 and goes on up till the terminal nodes 12, 13, 14, 15.

Factor Analysis [13] [14] is a method that is used to determine which variables are combined to generate a given factor. This is basically a multivariate statistical technique in which the whole set of

independent relationship is examined. This is a method for investigating whether a number of variables of interest $Y_1, Y_2, Y_3, \ldots\ldots, Y_i$ are linearly related to a smaller number of unobservable factors $F_1, F_2, F_3, \ldots\ldots, F_k$.

2.2 More Advanced Predictive Analytics Techniques

- **Time series forecasting** [33] predicts the future value of a measure based on past values. Time series forecasting uses a model to forecast future events based on known past events. Examples include stock pries and sales revenue.

- **Data profiling and transformation** [23] uses functions that analyze row and column attributes and dependencies, change data formats, merge fields, aggregate records, and join rows and columns.

- **Bayesian analytics** [26] [35] capture the concepts used in probability forecasting. It is a statistical procedure which estimate parameters of an underlying distribution based on the observed distribution.
- **Regression analysis** [40] is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another-the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate.

- **Classification** [12] used attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest. Examples include credit risk analysis, likelihood to purchase. Examples include acquisition, cross-sell, attrition, credit scoring and collections.

- **Clustering or segmentation** [14] separates data into homogeneous subgroups based on attributes. Clustering assigns a set of observations into subsets (clusters) so that observations in the same cluster are similar. An example is customer demographic segmentation.
- **Dependency or association analysis** [27] [12] describes significant associations between data items. An example is market basket analysis. Market basket analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

- **Simulation models** [18] a system structure to estimate the impact of management decisions or changes. Simulation model behavior will change in each simulation according to the set of initial parameters assumed for the environment. Examples include inventory reorder policies, currency hedging, and military training.

- **Optimization models** [16] a system structure in terms of constraints to find the best possible solution. Optimization models form part of a larger system which people use to help them make decisions. The user is able to influence the solutions which the model produces and reviews them before making a final decision as to what to do. Examples include scheduling of shift workers, routing of train cargo, and pricing airline seats.

- **Predictive Analytics Services Achievement** [17] [38] dramatically increasing the response rates of offers, mailings, and promotions, thus cross-selling and up selling your way to greater wallet share.
    - Recognizing the triggers and drivers that maximize customer satisfaction and actualize unrealized potential value.
    - Strengthening customer loyalty.
    - Acquiring the most profitable customers.
    - Optimizing product features and pricing.
    - Optimizing customer service and touch-point processes.

2.3 Applications of Predictive Analytics

Generally Predictive Analytics can be put to use in many applications, some of them are as follows:

- **Analytical customer relationship management (CRM)** [20], in which Predictive Analysis is applied to customer data to pursue CRM objectives.

- **Clinical decision support systems** [1],[32] most frequently use Predictive Analysis in health care to determine patient severity in certain conditions like heart disease, cancer, diabates, and other life time illnesses.
- **Insurance** [25], in which risk assessment is at the core of the business, where actuarial statistics have been the traditional tools to model various aspects of risk such as accident, health claims, or disaster rates, and the severity of these claims. The claim frequency is rare and stochastic in nature. For instance, the auto accident rate of an insured driver is never a clear no-accident class vs. accident class problem. In general, different kinds of insurance can use different statistical models depending on the fundamental nature of the claims process, requiring a predictive model that can be optimized for different underlying distributions.
- **Cross-sell** [5] **,** collect and maintain abundant data (e.g. customer records, sale transactions) and exploiting hidden relationships in the data can provide a competitive advantage to the organization. For an organization that offers multiple products, an analysis of existing customer behavior can lead to efficient cross sell of products. This directly leads to higher profitability per customer and strengthening of the customer relationship. Predictive analytics can help analyze customers' spending, usage and other behavior, and help cross-sell the right product at the right time.

- **Fraud detection** [28] **,** fraudulent insurance claims and credit card transactions alone cost tens of billions of dollars a year. In the case of credit card fraud, artificial neural-networks have been widely-used by many banks. The pattern of fraudulent transactions varies with time, requiring relatively frequent and rapid generation of new models..

- **Direct marketing** [34], are the amount of competing services available, businesses need to focus efforts on maintaining continuous consumer satisfaction. In such a competitive scenario, consumer loyalty needs to be rewarded and customer attrition needs to be minimized Predictive analytics can also predict this behavior accurately and before it occurs, so that the company can take proper actions to increase customer activity.

- **Product or economy level prediction** [33] in which**,** focus of analysis is not the consumer but the product, portfolio, firm, industry or even the economy. For example a retailer might be interested in predicting store level demand for inventory management purposes. Or the Federal Reserve Board might be interested in predicting the unemployment rate for the next year. These types of problems can be addressed by predictive analytics using Time Series techniques.

- **Text mining** [29] text fields in databases are a large percentage of the data stored in centralized data warehouses. Text mining is the search for valuable patterns in stored text. When stored documents have correct labels, such as the topics of the documents, then that form of text mining is called text categorization.

3. Related Work On Predictive Analytics

K. S. Kannan et al.[38], proposed an algorithm that was able to predict if the following day's closing price would increase or decrease better than chance (50%) with a high level of significance. The algorithm performed well on the technical analysis of the stocks and is used as a buying or selling signal that can give confidence to a trader's prediction of stock prices. The algorithm generated both increase and decrease predictions, but the predictions did not come very often. Patricia E.N. Lutu et al.[2] in his work focused on decision rule-based feature selection. Feature subset search algorithm uses decision rules to guide the search where the decisions are compared when mathematical functions are used. Selection of feature provides a high level of predictive classification performance. Tzung-Pei Hong et al. [3], Combined the advantages of the Apriori and the (*n*,*p*) algorithm in finding large item sets. As the (*n*, *p*) algorithm does, the proposed algorithm reduced the number of scanning datasets for finding *p* levels of large item sets. A new parameter was included to improve the computational efficiency. Jaaman Saiful Hafizah et al. [39], identified factors that influenced the gold price to develop a forecasting model using MLR that was able to predict the future gold price. For this two models were considered. The first model considered all possible independent variables and appeared to be useful for predicting gold price with 85.2% of sample variations where as the second model that took only four independent variables namely Commodity research bureau future index, Euro foreign exchange rate, Inflation rate, Money supply, achieved high rate of predictive analysis. S. H. Jaaman et al.[39], studied to mine profitable trading rules using rough set approach for Kuala Lumpur Composite Index and eighty individual firms listed in Bursa Malaysia. According to him the need for an extended study in using rough set methodology was for the validation process because the future is never exactly like the past. So it required the knowledge of artificial intelligence and other statistical tools. R. Bellazzi et al.[1], discussed the extent and role of predictive data mining and proposed a framework to cope with the problems of constructing, assessing and exploiting data mining models in clinical medicine. This work also reviewed the relevant work published in the area of clinical medicine, highlighted critical issues and summarized the approaches. The goal was to derive model that can use patient specific information to predict the outcome of interest and to thereby support clinical decision-making. Se June Hong et al.[4], emphasized on techniques for data mining for extracting massive collections of data, such as the historical record. Major enhancements of computing, storage, and networking capabilities were introduced.

4. Conclusion And Future Work

This study mainly intends to focus on the mining techniques using predictive analytics. Since predictive analytics is a major area of interest to almost all communities and organization, the application of it has provided a very high level of predictive performance. At the same time the widespread availability of several new computational methods and tools for predictive modelling assists the researchers and the practitioners to select the most appropriate strategy. We have presented an overview of some of the notable techniques for prediction. All analytical tools enable greater transparency and can find and analyze past trends to predict the probable future outcome of an event or

its likelihood to occur, as well as to discover the hidden nature of data. We can use the above techniques hybridized with few soft computing techniques to predict the future trends.

References

[1] Riccardo Bellazzi , Blaz Zupan , “Predictive data mining in clinical medicine: Current issues and guidelines”, International Journal of Medical Informatics 77 (2008) 81-97.

[2] Patricia E.N. Lutu , Andries P. Engelbrecht , “A decision rule-based method for feature selection in predictive data mining”,Expert Systems with Applications 37 (2010) 602-609.

[3] Tzung-Pei Hong , Chyan-Yuan Horng , Chih-Hung Wu , Shyue-Liang Wang, “An improved data mining approach using predictive itemsets”, Expert Systems with Applications 36 (2009) 72-80.

[4] Se June Hong , Sholom M. Weiss, “Advances in predictive models for data mining” Pattern Recognition Letters 22 (2001) 55 – 61.

[5] John R. Davies, Stephen V. Coggeshall, Roger D. Jones, and Daniel Schutzer, "Intelligent Security Systems," in Freedman, Roy S., Flein, Robert A., and Lederman, Jess, Editors (1995). Artificial Intelligence in the Capital Markets. Chicago: Irwin. ISBN 1-55738-811-3.

[6] Higgins J, (2005), The Radical Statistician, Prentice Hall Publishing.

[7] Jonathan T, (2009), Introduction to Applied Statistics, Introduction to applied statistics, Department of statistics, Stanford University, Statistics 191, pp. 1-38.

[8] Stockwell I, (2008), Introduction to Correlation and Regression analysis, SAS Global Forum, Paper-364, pp. 1-8.

[9] Marx B. D., Smith E. P.,(1990), principal component estimation for generalized linear regression, Journal of the Royal Statistical Society, vol.77, Issue.1, pp. 23-31.

[10] Ruppert D., Wand M. P., Caroll R.J., (2003), Generalized Regression Meodels, Journal of the American Statistical Association, STAT902, pp. 1-25.

[11] Smyth G. K., (2002), Non Linear Regression, Journal of American Statistical association, Encyclopaedia of Environ metrics, ISBN 0471 899976, Vol. 3, pp. 1405-1411.

[12] Ruggeri, Kennet, Faltin, (2008), Classification and Regression tree methods, In Encyclopaedia of Statistics in Quality and Reliability, Wiley, pp.315-123.

[13] Bryant and Yarnold (1994). "Principal components analysis and exploratory and confirmatory factor analysis". In: Grimm and Yarnold, Reading and understanding multivariate analysis. American Psychological Association Books. ISBN 978-1-55798-273-5.

[14] Sheppard, A. G. (1996). The sequence of factor analysis and cluster analysis: Differences in segmentation and dimensionality through the use of raw and factor scores. Tourism Analysis, 1 , 49-57.

[15] Sternberg, R.J.(1990). The geographic metaphor. In R.J. Sternberg, Metaphors of mind: Conceptions of the nature of intelligence (pp. 85–111). New York: Cambridge.

[16] Jannedy, Stefanie; Bod, Rens; Hay, Jennifer (2003). Probabilistic Linguistics. Cambridge, Massachusetts: MIT Press. ISBN 0-262-52338-8 .

[17] Gershenfeld, Neil A. (1999). The Nature of Mathematical Modeling. Cambridge, UK: Cambridge University Press. ISBN 978-0521-570954.

[18] Kingsland, Sharon E. (1995). Modeling nature: episodes in the history of population ecology. Chicago: University of Chicago Press. ISBN 0-226-43728-0.

[19] Weisstein, Eric W., "Logistic Equation" from Math World.

[20] Agresti, Alan. (2002). Categorical Data Analysis. New York: Wiley-Interscience. ISBN 0-471-36093-7.

[21] Amemiya, T. (1985). Advanced Econometrics. Harvard University Press. ISBN 0-674-00560-0.

[22] Balakrishnan, N. (1991). Handbook of the Logistic Distribution. Marcel Dekker, Inc.. ISBN 978-0824785871.

[23] Greene, William H. (2003). Econometric Analysis, fifth edition. Prentice Hall. ISBN 0-13-066189-9.

[24] Hilbe, Joseph M. (2009). Logistic Regression Models. Chapman & Hall/CRC Press. ISBN 978-1-4200-7575-5.

[25] Apte,C Apte, C., Grossman, E., Pednault, E., Rosen, B., Tipu, F.,White, B., 1999. Probabilistic estimation based data mining for discovering insurance risks. IEEE Intelligent Syst. 14 (6),49-58.

[26] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, Calif, 1993.

[27] L. Breiman, Classification and Regression Trees, Chapman & Hall, New York, London, 1993.

[28] Stolfo, S.J., Prodromidis, A., Tselepis, S., Lee, W., Fan, W.,Chan, P., 1997. JAM: Java agents for meta-learning over distributed databases. In: Proc. of KDDM97. pp. 74-81.

[29] Weiss, S., Indurkhya, N., 1998. Predictive Data Mining: A Practical Guide. Morgan Kaufmann, Los Altos, CA.

[30] Stepwise Multiple Linear Regression Analysis http: // marketing.byu.edu / htmlpages / books / pcmds / REGRESS.html

[31] Generalized Linear Models (GLZ), http://www.statsoft.com/textbook/generalized-linear-models/

[32] L. Devroye, L. Györfi, G. Lugosi (1996). A Probabilistic Theory of Pattern Recognition. New York: Springer-Verlag.

[33] Enders, Walter (2004). Applied Time Series Econometrics. Hoboken: John Wiley and Sons. ISBN 052183919X.

[34] Greene, William (2000). Econometric Analysis. London: Prentice Hall. ISBN 0-13-013297-7.

[35] Mitchell, Tom (1997). Machine Learning. New York: McGraw-Hill. ISBN 0-07-042807-7.

[36] Tukey, John (1977). Exploratory Data Analysis. New York: Addison-Wesley. ISBN 0201076160.

[37] Guidère, Mathieu; Howard N, Sh. Argamon (2009). Rich Language Analysis for Counterterrrorism. Berlin, London, New York: Springer-Verlag. ISBN 978-3-642-01140-5.

[38] Kannan K. Senthamarai, Sekar P. Sailapathi, Sathik M. Mohamed, (2010), Financial Stock Market Forecast using Data Mining Techniques, Proceedings of the International MultiConference Engineers and Computer Scientists, ISBN: 978-988-17012-8-2, Hong Kong.

[39] Jaaman Saiful Hafizah, Shamsuddin Siti Marriyam, Yusob Bariah, Ismail Munira (2009), A Predictive Model Construction Applying Rough Set Methodology for Malaysian Stock Market Returns, International Research Journal of Finance and Economics, ISSN 1450-2887, Issue (30), pp. 211-218.

[40] An Introduction to Regression Analysis, Alan O. Sykes, Chicago Working Paper in Law & Economics, http: // www.law.uchicago.edu/ files /file / 20. Sykes_Regression.pdf

[41] D Loomis, D Richardson, and L Elliott, Poisson regression analysis of ungrouped data, Occup Environ Med. 2005 May; 62(5): 325–329.

[42] Gordon K. Smyth, Nonlinear regression, Encyclopedia of Environmetrics, Edited by, Abdel H. El-Shaarawi and Walter W. Piegorsch, John Wiley & Sons, Ltd, Chichester, 2002, Volume 3, pp 1405–1411.