A Survey on Predictive Techniques in Datamining

C. Ganesh[1], Dr. E. Kesavulu Reddy[2]
[1]*Research Scholar, Computer Science – SVU College of CM&CS, Tirupati*
[2]*Assistant Professor, Computer Science – SVU College of CM&CS, Tirupati*

Abstract - Knowledge Discovery in Data (KDD) aims to extract non obvious information using careful and detailed analysis and interpretation. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions. Cloud computing is a versatile technology that can support a wide range of applications. These data are unique, having a combination of the following characteristics: few predictor variables, many predictor variables, highly collinear variables, very redundant variables and presence of outliers. In this paper we explain the various predictive data-mining techniques used to accomplish the goals and the methods of comparing the performance of each of the techniques.

Keywords: *Data Mining, Data Management, Data Prediction techniques.*

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviour, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Today, one of the biggest challenges that institutions/organizations face is the explosive growth of data and to use this data to improve the quality of managerial decisions. These networks can be expanded to enhance the access to the information resources using Data Mining and Cloud Computing technology [1]. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1].

II. PREDICTIVE DATAMINING

The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [2]. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc. The various application areas of data mining are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, Competitive Intelligence, Teaching Support, Climate modeling, Astronomy, and Behavioural Ecology etc.

According to Fayyad [3] data mining can be divided into two tasks: predictive tasks and descriptive tasks. The ultimate aim of data mining is prediction; therefore, predictive data mining is the most common type of data mining and is the one that has the most application to businesses or life concerns. Figure 2.1 shows a more complete picture of all the aspects of data mining.

DM starts with the collection and storage of data in the data warehouse. Data collection and warehousing is a whole topic of its own, consisting of identifying relevant features in a business and setting a storage file to document them. It also involves cleaning and securing the data to avoid its corruption. According to Kimball, a data ware house is a copy of transactional or non-transactional data specifically structured for querying, analyzing, and reporting [4]. Data exploration, which follows, may include the preliminary analysis done to data to get it prepared for mining. The next step involves feature selection and or reduction. Mining or model building for prediction is the third main stage, and finally come the data post-processing, interpretation, and/or deployment.

2.1. DATA ACQUISITION

The use of General Purpose Instrumentation Bus (GPIB) interface boards allows instruments to transfer data in a parallel format and gives each instrument an identity among a network of instruments [5],[6],[7]. Another way to measure signals and transfer the data into a computer is by using a Data Acquisition board (DAQ). A typical commercial DAQ card contains an analog-to-digital converter (ADC) and a digital-to-analog Converter (DAC) that allows input and output to analog and digital signals in addition to digital input/output channels [5],[6],[7].The process involves a set-up in which physical parameters are measured with some sort of transducers that convert the physical parameter to voltage (electrical signal) [8]. The signal is conditioned (filtered and amplified) and sent to a piece of hardware that converts the signal from analog to digital, and through software, the data are acquired, displayed, and stored. The topic of data acquisition is an extensive one and is not really the subject of this thesis. More details of the processes can be found in many texts like the ones quoted above.

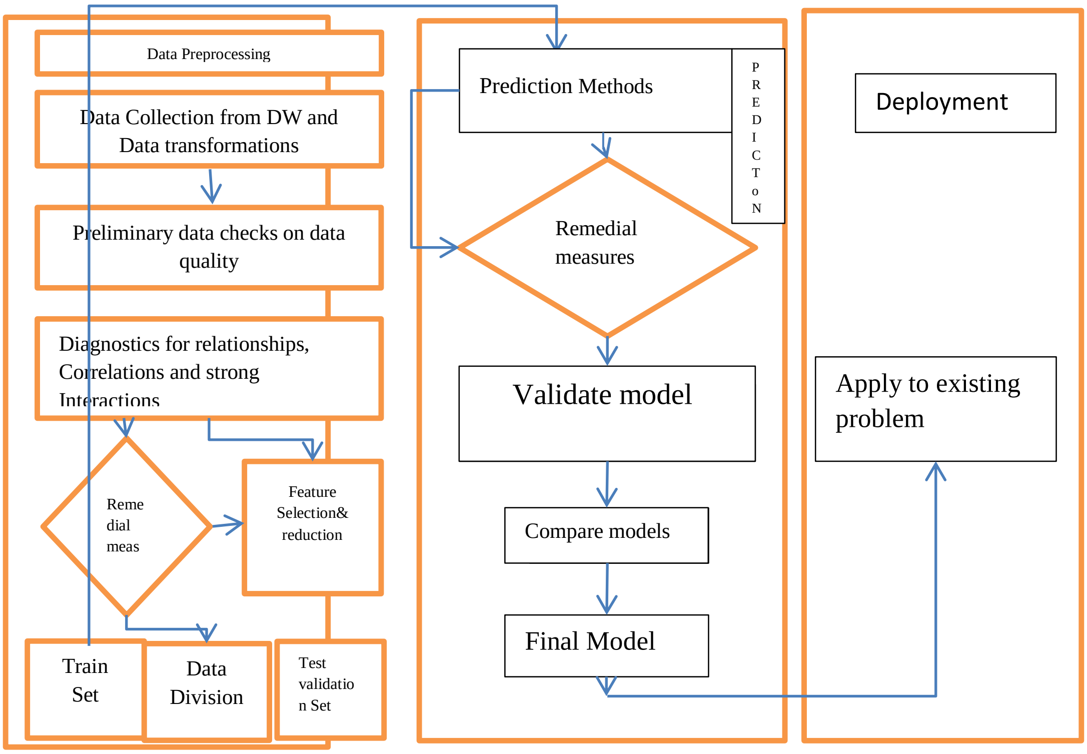
Data Preprocessing
Data Collection from DW and Data transformations
Preliminary data checks on data quality
Diagnostics for relationships, Correlations and strong Interactions
Reme dial meas
Feature Selection& reduction
Train Set
Data Division
Test validation Set
Prediction Methods
PREDICToN
Remedial measures
Validate model
Compare models
Final Model
Deployment
Apply to existing problem

Figure 1. The Stages of Predictive Datamining

2.1. DATA PREPARATION

Data preparation is very important because different predictive data-mining techniques behave differently depending on the preprocessing and transformational methods. There are many techniques for data preparation that can be used to achieve different data-mining goals.

2.2. Data Filtering and Smoothing

A filter is a device that selectively passes some data values and holds some back depending on the modeler's restrictions 10]. There are several means of filtering data.

2.2.1. Moving Average: This method is used for general-purpose filtering for both high and low frequencies [10], [11], [12]. It involves picking a particular sample point in the series, say the third point, starting at this third point and moving onward through the series, using the average of that point plus the previous two positions instead of the actual value. With this technique, the variance of the series is reduced. It has some drawbacks in that it forces all the sample points in the window averaged to have equal weightings.

2.2.2. Median Filtering: This technique is usually used for time-series data sets in order to remove outliers or bad data points. It is a nonlinear filtering method and tends to preserve the features of the data [12],[13]. It is used in signal enhancement for the smoothing of signals, the suppression of impulse noise, and the preserving of edges.

2.2.3. Peak-Valley Mean (PVM): This is another method of removing noise. It takes the mean of the last peak and valley as an estimate of the underlying waveform. The peak is the value higher than the previous and next values and the valley is the value lower than the last and the next one in the series [10],[12].

2.2.4. Normalization/Standardization: This is a method of changing the instance values in specific and clearly defined ways to expose information content within the data and the data set [10],[12]. Most models work well with normalized data sets. The measured values can be scaled to a range from -1 to +1. This method includes both the decimal and standard deviation normalization techniques.

2.2.5. Fixing missing and empty values: In data preparation, a problem arises when there are missing or empty values. A missing value in a variable is one in which a real value exists but was omitted in the course of data entering, and an empty value in a variable is one for which no real-world value exists or can be supposed [10],[12]. These values are expected to be fixed before mining proceeds. This is important because most Datamining modeling tools find it difficult to digest such values. Some data-mining modeling tools ignore missing and empty values, while some automatically determine suitable values to substitute for the missing values.

2.2.6. Categorical Data: Data-mining models are most often done using quantitative variables (variables that are measured on a numerical scale or number line), but occasions do arise where qualitative variables are involved. The variables are called categorical or indicator variables [14].

2.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [18] is an unsupervised parametric method that reduces and classifies the number of variables by extracting those with a higher percentage of variance in the data (called principal components, PCs) without significant loss of information [15],[16]. PCA transforms a set of correlated variables into a new set of uncorrelated variables. If the original variables are already nearly uncorrelated, then nothing can be gained by carrying out a PCA. PCA allows the analyst to use a reduced number of variables in ensuing analyses and can be used to eliminate the number of variables, though with some loss of information. However, the elimination of some of the original variables should not be a primary objective when using PCA.

PCA is appropriate only in those cases where all of the variables arise "on an equal footing." This means that the variables must be measured in the same units or at least in comparable units, and they should have variances that are roughly similar in size. In case the variables are not measured in comparable units, they should be standardized or normalized before a PCA analysis can be done.

The PCA is computed using singular value decomposition (SVD) [18], which is a method that decomposes the *X* matrix into a unitary matrix *U*, and a diagonal matrix S that have the same size as *X*, and another square matrix *V* which has the size of the number of columns of *X*.

$$X = U \bullet S \bullet V\ T$$

U = Orthonormal (MxM) matrix of
S = Diagonal (MxN) matrix

where n is the rank of X and the diagonals are known as the singular values and decrease monotonically. When these singular values are squared, they represent the eigenvalues.

V = Orthonormal matrix (NxN) of the eigenvectors, called the loadings vectors or the Principal Components:

$$z = U \bullet S$$
or
$$z = X \bullet V.$$

Where **Z** is an M X N matrix called the score matrix, **X** is an M X N matrix of original data, and **V** is an N X N transformation matrix called the loading matrix. M is the dimensionality of original space, N is the dimensionality of the reduced PC space, and M is the number of observations in either space.

This whole process is one of projecting the data matrix ***X*** onto the new coordinate
System ***V***, resulting in scores ***Z***. ***X*** can be represented as a linear combination of *M*
Orthonormal vectors ***Vi***

$$X = z_1 v_1^T + z_2 v_2^T + \ldots.. + z_m v_m^T$$

Vectors ***vi*** are the columns of the transformation matrix ***V***. Each feature ***zi*** is a linear
Combination of the data ***x***:

$$Z_i = x_1 v_{1i} + x_2 v_{2i} + \ldots\ldots\ldots + x_n v_{ni} = \sum_{j=1}^{n} x_j v_{ji}$$

It is possible to get the original vector ***x*** back without loss of information by transforming the feature vector ***z.*** This is possible only if the number of features equals the dimension of the original space, *n*. If k<*n* is chosen, then some information is lost. The objective is to choose a small *n* that does not lose much information or variability in the data. Many times there is variability in the data from random noise source; this variability is usually of no concern, and by transforming to a lower dimensionality space this noise can sometimes be removed. The transformation back to the original space can be represented
by important features ***zi*** and unimportant or rejected features ***ri***

$$x = \sum_{i=1}^{n} x_i v_i + \sum_{i=n+1}^{u} r_i v_i$$

In the above equation, there are n important features and ***u - n*** unimportant features. The transformation is selected so that the first summation contains the useful information, and the second summation contains noise [18].The vectors ***vi*** form an orthogonal (actually orthonormal) basis in the PC space.The basis vectors are chosen to minimize the sum of squared errors between the estimate and the original data set

$$\text{error} = x - \sum_{i=1}^{n} z_i v_i$$

As shown above, the optimal choice of basis vectors satisfies the following relationship [17].

$$\sum v_i = \lambda_i v_i$$

Again, we recognize this as an eigenvalue problem where λ***i*** and ***vi*** are the
Eigen values and eigenvectors of the covariance matrix Σ respectively. The eigenvectors ***vi*** are termed the principal components (PCs) or loadings.

2.4. Correlation Coefficient Analysis (CCA)

Correlation coefficient analysis (CCA) [19] assesses the linear dependence between two random variables. CCA is equal to the covariance divided by the largest possible covariance and has a range from -1 to +1. A negative correlation coefficient means the relationship is an indirect one, or, as one goes up, the other tends to go down. A positive correlation coefficient shows a direct proportional relationship: as one goes up, the other goes up also [8]. The correlation coefficient can be shown with an equation of the covariance relationship. If the covariance matrix is given by

$$\text{COV(x,y)=} \begin{bmatrix} \sigma_x{}^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y{}^2 \end{bmatrix}$$

The correlation coefficient is:

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The correlation coefficient function returns a matrix of the following form:

$$\text{Corrcoef =} \begin{bmatrix} 1 & p_{xy} \\ P_{xy} & 1 \end{bmatrix}$$

The correlation coefficient ≤ 0.3 shows very little or no correlation (= 0). A correlation coefficient >3 but <0.7 is said to be fairly correlated. A correlation coefficient ≥ 0.7 shows a high or strong linear relationship. The correlation coefficient of any constant signal (even with noise) with another signal is usually small. To get a good estimate of the correlation coefficient, especially for data sets with varying magnitudes, the data should first be scaled or normalized, or it will give more importance to inputs of larger magnitude.

III. COMPARATIVE STUDY OF PREDICTIVE DATAMINING TECHNIQUES

Having discussed the data acquisition, and some data preprocessing techniques, an overview of the predictive techniques to be compared is given in this section. There are many predictive data-mining techniques (regression, neural networks, decision tree, etc.) but in this work, only the regression models (linear models) are discussed and compared. Regression is the relation between selected values of x and observed values of y from which the most probable value of y can be predicted for any value of x [20]. It is the estimation of a real value function based on finite noisy data. Linear Regression was historically the earliest predictive method and is based on the relationship between input variables and the output variable.

3.1. Linear regression

Regression analysis is 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

$$Y = b + mX \qquad (1)$$

Here b and m 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

$$Y=m0+m1X \qquad (2)$$

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,

$$|D| = \{ (X1 , Y1), (X2 , Y2)....(X|D| , Y|D|) \} \ (3)$$

The estimated regression coefficient is given as

$$w1 = \frac{\Sigma i\ (Xi – MeanX)\ (Yi – MeanY)}{\Sigma i\ (Xi – MeanX)2}$$

where i = 1 to |D|, (4)

w0=MeanY–w1-Meanx (5)

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

3.2. Multiple Linear Regression (MLR)

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 X1, X2…Xp (p>1) is expressed by the equation as ,

$$Y = \alpha + \Sigma k\ \beta k\ Xk + € \ (6)$$

Where α, βk (k = 1,2...p) are the parameters and € is the error term. MLR combines the idea of correlation and linear regression.

3 .2.1.Logistic regression

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,

$$f (Z) = eZ / (eZ + 1) = 1 / (1+e -Z) \qquad (7)$$

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.

IV. CONCLUSION

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 modeling 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. We can use the above techniques hybridized with few soft computing techniques to predict the future trends. DM will be one of the main competitive focuses of organizations.

V. REFERENCES

[1]. Lyman, P, and Hal R. Varian, "How much storage is enough?" (2003).

[2]. Way, Jay, and E. A. Smith, "Evolution of Synthetic Aperture Radar Systems and Progression to the EOS SAR," IEEE Trans. Geoscience and Remote Sensing, 29:6 (1991), pp 962-985.

[3]. Usama, M. Fayyad, et al, Advances in Knowledge Discovery and Data mining, Cambridge, Mass.: MIT Press (1996).

[4]. Ralph, Kimball, The Data Warehouse Toolkit Practical Technique for Building Dimensional Data Warehouses. New York: John Wiley (1996).

[5]. Austerlitz, Howard., Data Acquisition Techniques Using PCS. USA: Elsevier 2003.

[6]. Caristi, A. J., IEEE-488 General Purpose Instrument Bus Manual. London academic Press 1989.

[7]. Klaassen, B. Klaas., Electronic and Instrumentation. New York: Cambridge University Press (1996)

[8]. Hansen, P. C., "Analysis of discrete ill-posed problems by means of the L Curve," SIAM Reviews, 34:4 (1992), pp. 561-580

[9]. Hines, J. Wesley, Advanced Monitoring and Diagnostic Techniques, NE 579 (Summer 2005).

[10]. Pyle, Dorian, Data Preparation for Data-Mining. San Francisco, Morgan Kaufmann (1999).

[11]. Gencay, Ramazan, and F. Selcuk, An Introduction to Wavelets and other Filtering Methods in Finance and Economics. San Diego, CA: Elsevier (2002).

[12]. Olaf, Rem, and M. Trautwein, "Best Practices Report Experiences with Using the Mining Mart System." Mining Mart Techreport. No. D11.3 (2002).

[13]. Kassams Lee Yong, "Generalized Median Filtering and Related Nonlinear Techniques," IEEE Transactions on Signal Processing,1985)pp 672- 683.

[14]. Agresti, Alan, Categorical Data Analysis, 2nd edition. Hoboken, New Jersey Wiley Inter science 2002

[15]. Morison, D. F., Multivariate Statistical Methods, 2nd Edition. New York: McGraw-Hill (1976).

[16]. Seal, H., Multivariate Statistical Analysis for Biologists, London: Methuen (1964).

[17]. Bishop, C.M., Neural Networks for Pattern Recognition, Oxford, UK: Oxford University Press (1995).

[18]. Jolliffe, I.T., Principal Component Analysis, New York: Springer-Verlag 1986

[19]. Cohen, Jacob, et al., Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Mahwah, New Jersey: Lawrence Erlbaun Asso