Jr. of Industrial Pollution Control 33(S3)(2017) pp 1430-1436
www.icontrolpollution.com
Research Article

PREDICTIVE DATA MINING TECHNIQUES FOR MANAGEMENT OF HIGH DIMENSIONAL BIG-DATA

SONI LANKA[1], RADHA MADHAVI M[2*], BASHIR SULEMAN ABUSAHMIN[3], NAGESH PUVVADA[4] AND VEMURI LAKSHMINARAYANA[5]

[1]Department of Computer Science and Systems Engineering, Andhra University, Vizag, India

[2]Arupadai Veedu Institute of Technology, Kancheepuram, Chennai-603104, Tamil Nadu, India

[3]Department of Petroleum and Chemical Engineering, Universiti Teknologi, Brunei, Jalan Tungku Link, BE1410, Brunei Darussalam

[4]The Institution of Electronics and Telecommunication Engineers, Chennai, Tamil Nadu, India

[5]Principal, Arupadai Veedu Institute of Technology, Kancheepuram, Chennai-603104, Tamil Nadu, India

(Received 17 June, 2017; accepted 24 November, 2017)

Key words: Multiple linear regressions, Principal component regression, Partial least squares

ABSTRACT

Data mining is a technique, wherein the historical data is explored in search of a systematic relationship between variables and/or have a consistent pattern. This relationship is utilized to validate the outcomes by applying the identified patterns onto new data subsets. This paper compares three predictive data-mining techniques, namelymultiple linear regression, principal component regression and the partial least squares ona unique dataset. This data is unique, having a characteristics combination of presence of outliers, highly collinear variables,very redundant variables and predictor variables. In the initial step after pre-preparing information, negligible number of factors are chosen that can totally anticipate the reaction variable. These diverse information mining strategies, which has distinctive use techniques were actualized on the total informational index and the best strategy in every procedure was distinguished and this is utilized for worldwide examination with different systems for similar information.

INTRODUCTION

As of late, information mining has turned out to be a standout amongst the most profitable apparatuses for separating and controlling information and for building up examples keeping in mind the end goal to deliver helpful data for basic leadership. The disappointments of structures, metals, or materials (e.g. structures, oil, water or sewage channels) in a situation are frequently either an aftereffect of numbness or the failure of architects to observe past issues or study the examples of past episodes keeping in mind the end goal to settle on educated choices that can thwart future events. About all territories of life exercises show a comparative example. Regardless of whether the movement is back, saving money, advertising, retail deals, generation, populace think about, business, human relocation, wellbeing segment, observing of human or machines, training and so forth, have approaches to record known data. However, they are impeded by not having the correct devices to utilize this known data to handle the instabilities without bounds. Most recent information

***Corresponding authors email: mrmadhavi5@gmail.com**

accumulation advances to be specific standardized identification scanners in business spaces and sensors in logical and modern segments, have prompted the era of immense measures of information (Linoff and Berry, 2011; Rao, *et al*., 2009; Busahmin, *et al*., 2016; Rao, *et al*., 2010). This enormous development in information and databases has produced a squeezing requirement for new strategies and devices that can cleverly and consequently change information into helpful data and learning (Kidd, 2012). These necessities incorporate the programmed synopsis of information, the extraction of the "embodiment" of data put away, and the revelation of examples in the crude information. These can be accomplished through information investigations, which include basic questions, or components for showing information. Such information examination methods include information extraction, change, grouping, and investigation to distinguish designs with a specific end goal to make expectations.

Assessments of future estimations of business needs and suggestions, assume unmistakable part. The business desires or assessing of supply, exhibiting and budgetary decisions needs a basic information mining (Rao, *et al*., 2010; Kidd, 2012; Han, *et al*., 2011; Liao, *et al*., 2012). Information mining is the investigation of recorded information (typically vast in size) looking for a predictable example as well as a deliberate connection between factors; it is then used to approve the discoveries by applying the identified examples to new subsets of information (Liao, *et al*., 2012; Giudici, 2005; Karri, *et al*., 2014). The underlying foundations of information mining start in three ranges established insights, counterfeit consciousness (AI) and machine learning (Han, *et al*., 2011; Karri, *et al*., 2013). (Torgo and Torgo, 2011) depicted information mining as a mix of measurements, computerized reasoning, and database examine, and noticed that it was not a field important to numerous up to this point. As per (Fayyad, *et al*., 1996) information mining can be separated into two undertakings prescient assignments and elucidating errands. A conclusive purpose of data mining is expectation; along these lines, prescient information data mining is the most generally perceived kind of data mining and is the one that has the most application to associations or life concerns.

PREDICTIVE DATA MINING

Predictive data-mining (PDM) strategies used to achieve the objectives and the strategies for looking at the execution of each of the methods. PDM has three stages which are elaborated in Fig. 1, shows a mere picture of all the aspects of data mining. Information investigation, is the preparatory stage, where examination done to information to get it handled for mining. The following stride includes highlight determination or potentially decrease. Digging or model working for expectation is the third primary stage, lastly come the information post-handling, understanding, and additionally arrangement. Applications reasonable for information mining were boundless and as yet being investigated in numerous territories of business and every day necessities. According to (Fayyad, *et al*., 1996), PDM yields unexpected pockets of information that can explore business prospects, new markets, new approaches to reach customers and better ways of doing profitable business.

DATA MINING TECHNIQUES

The Different data-mining techniques like principal component regression (PCR) which is an unsupervised technique based on the principal component analysis;the partial least squares (PLS) which is an supervised technique and multiple linear regression (MLR) which is based on the ordinary least-square approach were analyzed and applied on unique data set (Karri, *et al*., 2013; Torgo and Torgo, 2011; Kantardzic, 2011; Karri and Babovic, 2017).

A. Multiple linear regression (MLR)

MLR endeavors to show the connection between at least two illustrative factors and a reaction variable by fitting a straight condition to watched information. MLR is the most widely recognized type of direct relapse examination. As a prescient examination, the numerous straight relapse is utilized to clarify the connection between a persistent ward variable from at least two autonomous factors. The free factors can be ceaseless or straight out.

B. Principal component regression

Principal component analysis (PCA) is an unsupervised parametric strategy that diminishes and orders the quantity of factors by separating those with a higher rate of fluctuation in the information (called primary components, PCs) without noteworthy loss of data (Rao, et al., 2009; Abusahmin, et al., 2017; Karri, 2011). PCA changes an arrangement of associated factors into another set of uncorrelated factors. On the off chance that the first factors are as of now almost uncorrelated, then nothing can be picked up via doing a PCA. For this situation, the real dimensionality of the information is equivalent to the quantity of factors measured, and it is impractical to look at the information in

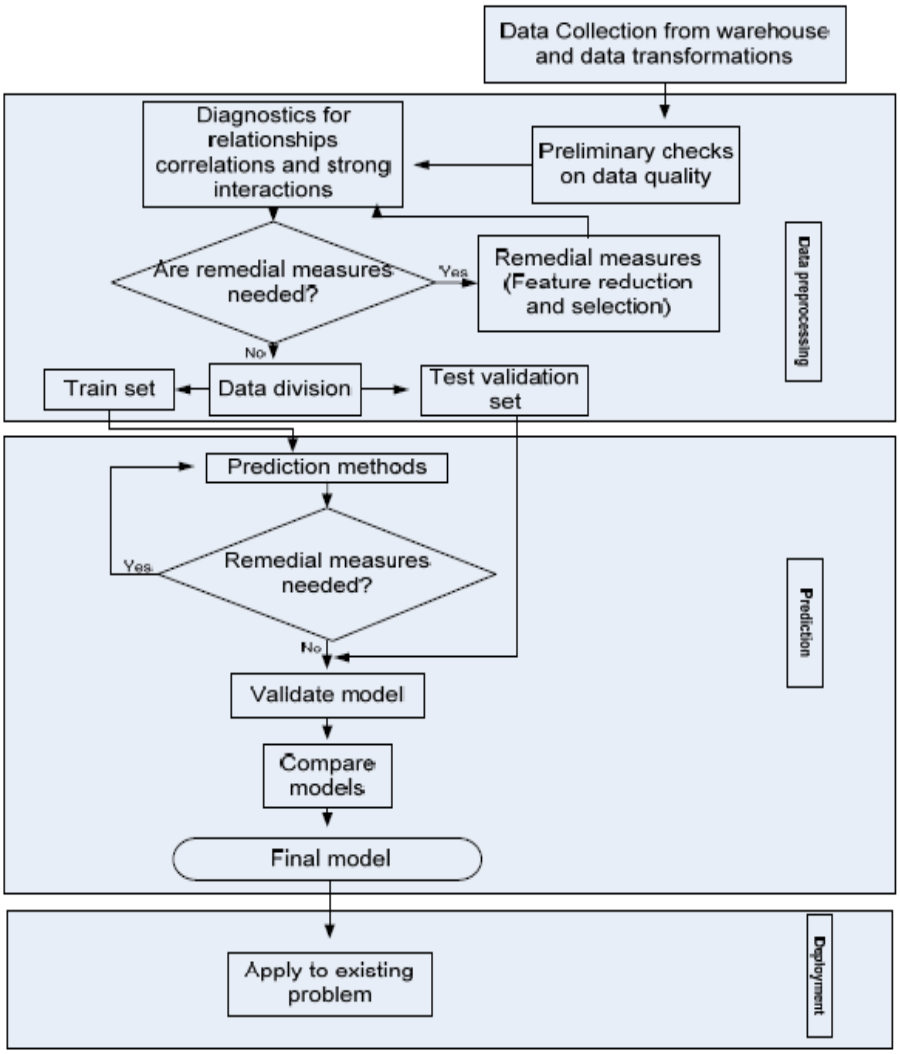

Fig. 1. The stages of predictive data mining.

a diminished dimensional space. Essentially, the extraction of foremost parts adds up to a fluctuation expansion of the first factor space. PCA permits to utilize a diminished number of factors in resulting investigations furthermore, can be utilized to wipe out the quantity of factors, however with some loss of data. In any case, the disposal of a portion of the first factors ought not to be an essential target when utilizing PCA. PCA is proper just in those situations where the greater part of the factors emerge "on a break even with balance." This implies the factors must be measured in similar units or at minimum in practically identical units, and they ought to have changes that are generally comparable in size. For all investigation circumstances, PCA can be suggested as an initial step. It can be performed on an arrangement of information before playing out some other sorts of multivariate investigations. During the time spent doing this, new factors (elements) called main segments (PCs) can be framed in diminishing request of significance, so that (1) they are uncorrelated and orthogonal, (2) the primary key part represents as a great part of the changeability in the information as could reasonably be expected, and (3) each succeeding segment represents however much of the rest of the inconstancy as could be expected. The implementation of PCR technique is as shown in Fig. 2. The PCA is processed utilizing solitary esteem decay (SVD), which is a strategy that disintegrates the X grid into a unitary network U, and a corner to corner lattice S that have an indistinguishable size from X, what's more, another square network V which has the extent of the quantity of sections of X.

$X = U.S.V^T$

U = Orthonormal (M × M) matrix

S = Diagonal (M × N) matrix

Where diagonals are known as the singular values and which decrease monotonically. When these

singular values are squared, they represent the Eigen values.

C. Partial Least Squares

Another prescient information mining system is the Partial slightest squares (PLS) procedure. PLS is a technique for demonstrating input factors (information) to anticipate a reaction variable. It includes changing the information (x) to another variable or score (t) and the yield information (y) to another score (u) making those uncorrelated elements and expelling co-linearity between the information and yield factors. A direct mapping (b) is performed between the score vectors t and u (Fig. 3). The score vectors are the estimations of the information on the stacking vectors and q. Moreover, a standard part like investigation is done on the new scores to make stacking vectors (p and q). Rather than chief part investigation (PCA), PLS concentrates on clarifying the relationship framework between the information sources and yields however PCA harps on clarifying the differences of the two factors. PCA is an unsupervised procedure and PLS is managed. This is on the grounds that the PLS is worried with the relationship between the info (x) and the yield (y) while PCA is just worried with the connection between the information factors x.

Fig. 4 speaks about the mapping area between the t and u scores. The benefit of PLS is that it draws out the most extreme measure of covariance clarified with the base number of parts. The quantity of inert components to show the relapse model is picked utilizing the decreased Eigen variables. The Eigen components are proportionate to the particular qualities or the clarified variety in the

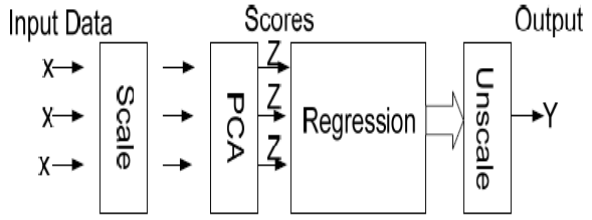

Fig. 2. Schematic diagram of PCR.

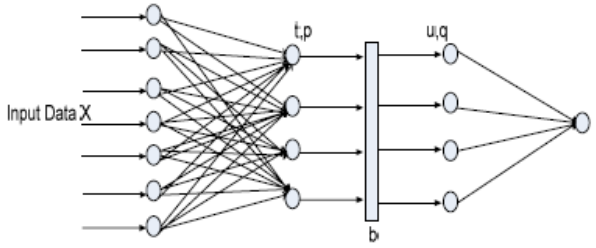

Fig. 3. Schematic diagram of the PLS inferential design.

PC determination and are regularly called the Malinowski's decreased Eigen esteem (Rosipal and Krämer, 2006). At the point when the decreased Eigen qualities are fundamentally equivalent, they represent commotion.

PROCEDURE AND METHODOLOGY

The methodology implemented in this research study is described in this section. This is the strategy defined to evaluate the efficacy of different predictive data mining techniques using a unique dataset. Fig. 4 shows the flow chart of the methodology used in this research work. The basic stage is theverify for any hidden relationships between the process variables in the data set. The data set is first pre-processed, as well as the preliminary diagnosis done on the data set to gain an insight into the characteristics of data set. The relationship between the parameters is assessed by plotting the contributions over the output of raw data set. The information is preprocessed by scaling or institutionalizing them to decrease the level of scattering between the factors in the informational index. The connection coefficients of each of the different informational collections are figured to check more on the connection between the info factors and the yield factors. This is trailed by finding the singular value decomposition of the informational collections changing them into principal components.

At this stage, the informational indexes are partitioned into two a balance of, setting the odd number information focuses as the "preparation set" and the significantly number information focuses as the "test approval informational collection." Now the preparation information for every informational index is utilized for the model building. For each preparation informational index, a prescient information mining system is utilized to manufacture a model, and the different strategies for that method are utilized. This model is approved by utilizing the test approval informational collection. Nine model ampleness criteria are utilized at this phase to quantify the decency of fit and sufficiency of the expectation. The outcomes are introduced in Table 1. The model is relied upon to perform well when distinctive informational collections are connected to it. In this review, the inaccessibility of various however comparable genuine informational collections has constrained this review to utilizing just the test informational collection for the model approval. This is not a major issue since this work is constrained to model examination and is not principally worried with the outcomes after organization of the model.

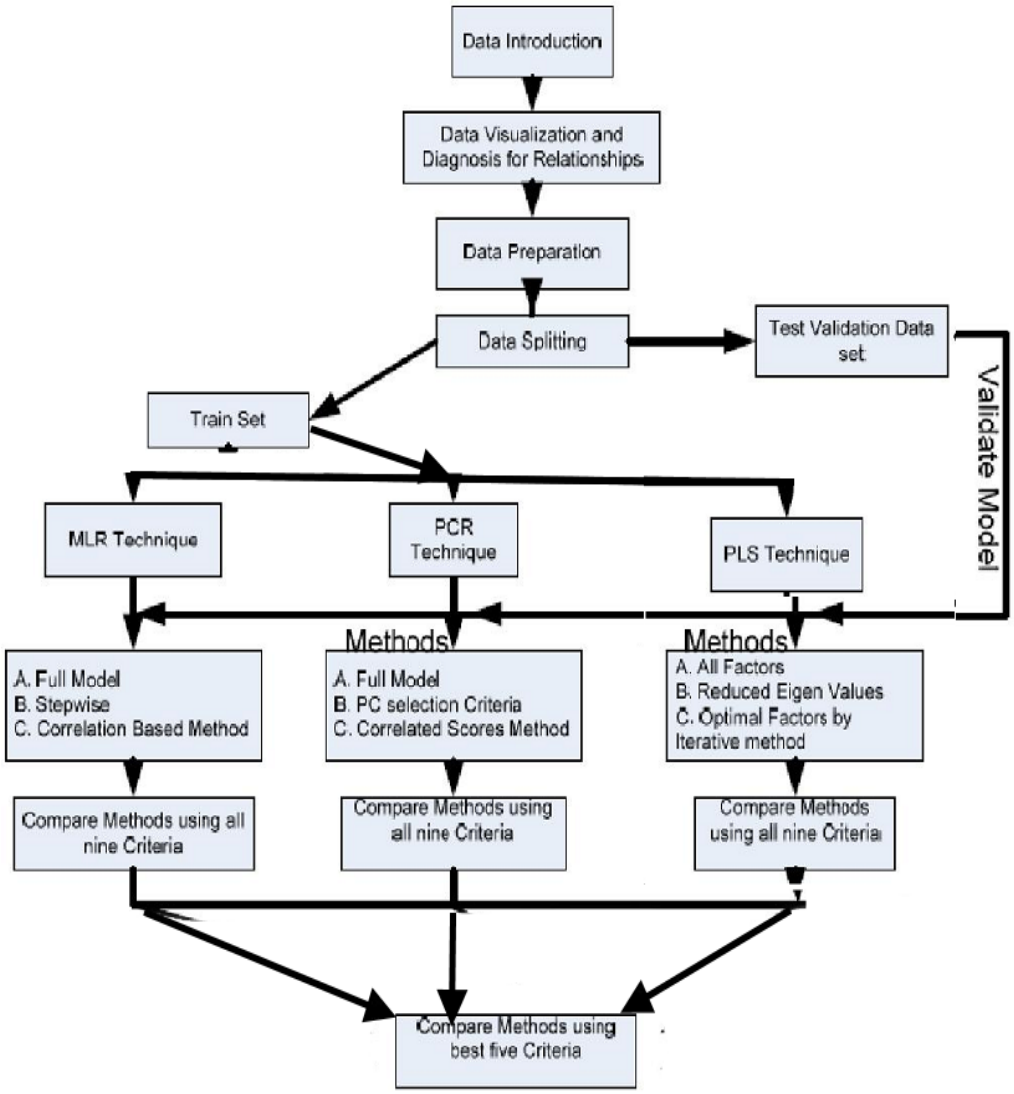

Fig. 4. Flow chart of the methodology.

Table 1. Statistical metrics of data set

RMSE	R^2	F	p
1.28046	0.60357	105.662	1.56854e-67

At last, all the three techniques are looked at (in view of their outcomes on every informational collection) utilizing extremely solid model sufficiency criteria. The best outcome gives the best expectation method or calculation for that specific kind of informational index.

RESULTS AND DISCUSSION

A. The Multiple linear regression (MLR) on housing data

The stepwise regression model is developed using MATLAB 2014®. The multiple linear regression model is developed from the variables that are significantly different from zero made the model. The dashed lines in Fig. 5, indicates the eliminated variables in the model. Those variables (dotted lines) whose confidentiality index is highand crossed the zero line were also not statistically different from zero and therefore deleted from the model development. The stepwise model was better in terms of R2, RMSE, F, cost objective, p (significant value) as shown in Table 1.

B. Principal component regression (PCR)

The train dataset that was selected for training was initially standardized from the standard deviation and mean of the data. This scaled training set were further used to scale the testing dataset. The loadings of principal components representing the dominant variables are shown in Fig. 6. The performance of the

predicted output against the testing data is presented in Fig. 7.

C. Partial least squares (PLS)

The covariance matrix of the PLS model were very much similar as those of PCR with corresponding number of principal components or factors because the data set was standardized and both used factors

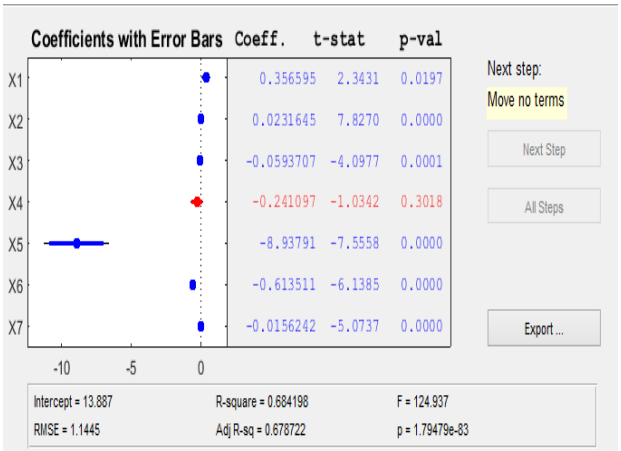

Fig. 5. Parameter estimations and confidence interval using stepwise regression.

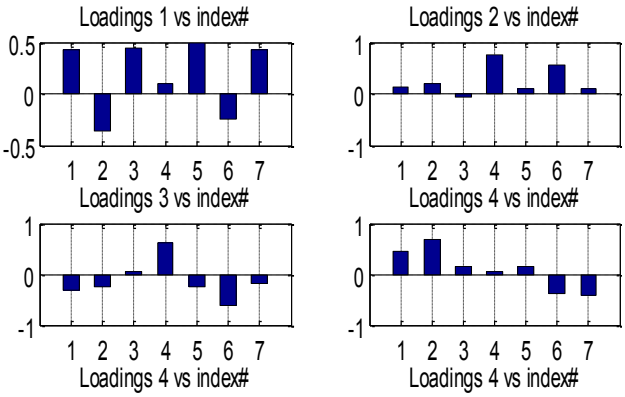

Fig. 6. PC Loadings showing the dominant variables in the PCs 1 to 7.

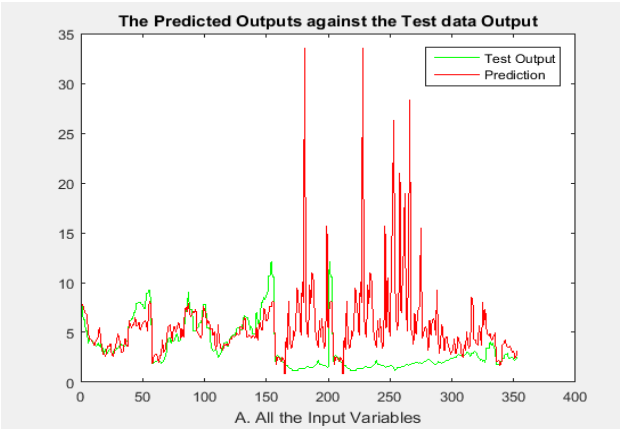

Fig. 7. The predicted output against the test data output.

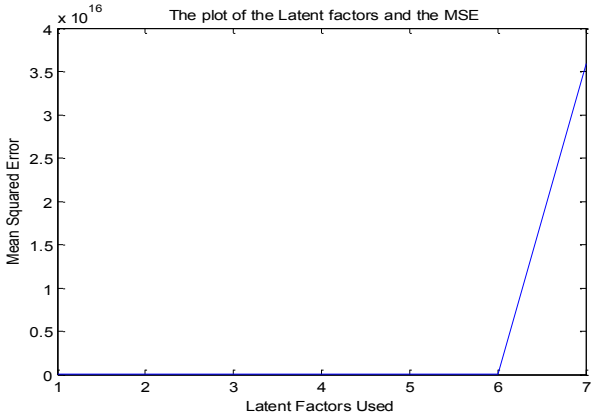

Fig. 8. MSE vs. latent factor.

or PCs. Iteratively using all the dominant factors and plotting the resulting mean square errors against the latent factors indicates the optimal factor which has the least MSE as shown in Fig. 8.

CONCLUSION

In this research study, the various data pre-processing techniques were used to organize the data set. Different data mining techniques such as multiple linear regression, principal component register and partial least square have been studied and implemented to obtain accurate output in terms of R^2, RMSE, F, cost objective, p (significant value). Results conclude that each technique has different performance. These diverse strategies were at first utilized on the one of a kind informational index and the best strategy in every system was noted and utilized for worldwide examination with different procedures for similar informational collection. From the investigation, it can be seen that the condition number of any information network has an immediate connection to the quantity of factually noteworthy factors in the model. The PLS, by and large, gave better models both on account of not well molded informational indexes and furthermore for informational indexes with excess information factors.

REFERENCES

Abusahmin, B.S., Karri, R.R. and Maini, B.B. (2017). Influence of fluid and operating parameters on the recovery factors and gas oil ratio in high viscous reservoirs under foamy solution gas drive. *Fuel*. 197 : 497-517.

Busahmin, B., Maini, B., Karri, R.R. and Sabet, M. (2016). Studies on the stability of the foamy oil in developing heavy oil reservoirs. Defect and Diffusion Forum, *Trans Tech Publications Ltd. Switzerland*. pp. 111-116.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and

Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. *AAAI Press Menlo Park*. 21.

Giudici, P. (2005). Applied data mining: Statistical methods for business and industry. *John Wiley & Sons, NY, USA.*

Han, J., Pei, J. and Kamber, M. (2011). Data mining: Concepts and techniques. *Elsevier, The Netherlands*.

Kantardzic, M. (2011). Data mining: Concepts, models, methods, and algorithms. *John Wiley & Sons, NY, USA.*

Karri, R.R. (2011). Evaluating and estimating the complex dynamic phenomena in nonlinear chemical systems. *International Journal of Chemical Reactor Engineering*. 9 : 38.

Karri, R.R. and Babovic, V. (2017). Enhanced predictions of tides and surges through data assimilation. *International Journal of Engineering, Transactions A: Basics*. 30(1) : 23-29.

Karri, R.R., Badwe, A., Wang, X., El Serafy, G., Sumihar, J., Babovic, V. and Gerritsen, H. (2013). Application of data assimilation for improving forecast of water levels and residual currents in Singapore regional waters. *Ocean Dynamics*. 63(1) : 43-61.

Karri, R.R., Wang, X. and Gerritsen, H. (2014). Ensemble based prediction of water levels and residual currents in Singapore regional waters for operational forecasting. *Environmental Modelling & Software*. 54 : 24-38.

Kidd, A. (2012). Knowledge acquisition for expert systems: A practical handbook. *Springer Science & Business Media*.

Liao, S.H., Chu, P.H. and Hsiao, P.Y. (2012). Data mining techniques and applications—A decade review from 2000 to 2011. *Expert Systems with Applications*. 39(12) : 11303-11311.

Linoff, G.S. and Berry, M.J. (2011). Data mining techniques for marketing, sales, and customer relationship management. *John Wiley & Sons, NY, USA.*

Rao, K.R., Rao, D.P. and Venkateswarlu, C. (2009). Soft sensor based nonlinear control of a chaotic reactor. *IFAC Proceedings Volumes*. 42(19) : 537-543.

Rao, K.R., Srinivasan, T. and Venkateswarlu, C. (2010). Mathematical and kinetic modeling of biofilm reactor based on ant colony optimization. *Process Biochemistry*. 45(6) : 961-972.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*. 34-51.

Torgo, L. and Torgo, L. (2011). Data mining with R: learning with case studies. *Chapman & Hall/CRC Boca Raton, Florida, USA*.