ISSN: 0976-3104
REGULAR ISSUE
Lavanya

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

STUDY AND APPLICATION OF DYNAMIC INCREMENTAL REGRESSION IN SEASONAL CROP MARKETING SUPPORTED BY SONN

K. Lavanya

School of Computing Science and Engineering,VIT University,TN, INDIA

ABSTRACT

Seasonal crop Marketing includes market infrastructure modeling, providence of crop market information and multi level conservatory, advisory, training services through synchronized networks. In modern world it became challenging with the most recent technologies and association of commission agents who keep their margins and move the produce further. In the through course of action of marketing the farmer gets the lowest price and the ultimate consumer pays the premier as the association of more agents in the entire supply chain process. To benefit the farming community the internal crop marketing network in the nation needs to be incorporated and strengthened. Internet can be effectively used to strengthen the supply, marketing chain for agro based companies leading to better price realization by farmers. It enables all outputs of farming to customers at sensible price without compromising on the quality of the produce. Here a novel ensemble model of incremental approach of regression associated with Self Organized Neural Network (SONN) clustering is proposed that helps farmers in decision making of agricultural goods based on real time and equipped marketing information in domain of agricultural networked channel pattern. The model was tested against vellore district crop data and proved better performance against conventional prediction methods.

Received on: 07th-Jan-2015
Revised on: 16h-Feb-2016
Accepted on: 19th-Feb-2016
Published on: 20th-Apr-2016

KEY WORDS

Agricultural marketing, Incremental Regression, Neural Networks, Prediction

***Corresponding author: Email:** lavanya.k@vit.ac.in; **Tel:** +91-9786361286

INTRODUCTION

The mounting trend of agricultural production has invoked new challenges in terms of finding market for the marketed surplus. The changing global agricultural scenario insists the need to review the policies related to pricing, marketing and trading of agricultural commodities. Agricultural marketing reforms and creation of marketing infrastructure has been initiated to respond to the market needs and consumer preferences. The exploitation of intermediaries and traders prevented farmers to ensure better prices and timely payment for their produce. Exporters, processors and retail chain suppliers cannot obtain straight from the farmers as the produce is required to be channelized through regulated markets and licensed traders.

Agricultural Market network is set of cooperatives at the local, regional, state and national levels .If the seasonal crop production is enomorous it ends up in low price for product. But if the same product is transported to another market where there is market need and consumer preference considerable profit could be achieved. At present scenario it is impossible to know the market information of products from a single place. In agricultural sector the amplified marketing information flow has a positive effect in decision making but, collecting and disseminating information is often complicated and expensive. Data mining techniques has a great deal of interest in recent years that increases the amount of marketing data provided to all participants, decreases the cost of disseminating the information and facilitates collected information for better price realization.

Forecasting the price movements is a major issue in confronting producers, brokers and sellers. Automatic learning systems create extensible data-driven applications by inferring the appearance and behavior of data entities at run time from the database. These rule based technologies concentrate in static data which are less successful. The proposed dynamic marketing extraction using SONN increases the accuracy of precipitation in forecasting of predicting prices in agricultural products wholesale markets.

LITERATURE SURVEY

Agriland – Nellikuppam, Tamil Nadu project includes upgrade market and commercial information among farmers, train farmers on the latest developments in agricultural technology, provide access to farmers to the precise markets through reasonable credit and transportation solutions and to support farmers in raising their income in triple means

THE IIOAB JOURNAL
ISSN: 0976-3104

in next five years. Agmarket scheme provides funds to state and national level institutions managing the markets and executing market-led expansion activities and thus has no distinct gender specific provisions under the scheme. It covers market, price, infrastructure and advertising related data for proficient marketing. The designed system uses a quadratic forecasting model of linear time series to forecast the cost, and compares the prediction results by using different time series and different training data to identify the best prediction model to foretell the price in sites [1]. A study of the ASD-DM Methodology [2] helps arrive at an understanding of the need and relevance of dynamic and agile methods of software engineering in data mining procedures. [3] Lists out applications that evaluate different data mining techniques for their accuracy in predicting the default credit payments thereby giving more insight into the performance abilities of various predictive data mining techniques. [4] wherein describes a direct and an iterative forecast method of prediction using neural networks that is highly relevant to this work. In [5], various regression techniques such as the Ordinary Least Squares (OLS), Principal Component Regression (PCR) and Latent Root Regression (LRR) methods are defined and their relevance due to shelf life prediction is discussed. [8] Applies data mining techniques to categorize the momentous variables that measure network intrusion from the wealth of raw network data and perform competent susceptible assessment based on those variables. [10] Deals with appropriate regression techniques and compares four diverse techniques on selected agriculture data. [11] selects significant patterns for the effective prediction of heart attack using Multi-layer Perceptron Neural Network with Back-propagation as the training algorithm.[12] states that statistical and forecast analyses on data sets refers the growth of olive fly by present characteristics (big variability and non-linearity) which makes complex to be treated mathematically. According to [14] ANNs are useful when data are unsupervised because they can discover from the data. ANNs, however, do not provide a clean model to a problem. It is hard to know how they come to their conclusions, because they are like a "black box", only providing a final outcome and not what causes the result. Textual articles appearing in the leading and influential financial newspapers are taken as input by [15].Then the daily closing values of major stock market indices in developed countries are predicted. Textual sentences describe not only the cause but also the reason behind it. Exploring textual information in addition to numeric time series data improves the eminence of the input. From the background study it is clear that the statistical and data mining techniques is applicable on datasets of reasonable size. But in practical, the increasing size of datum deviates the prediction accuracy. Hence our paper aims in extracting the user defined crop price datum by SONN clustering followed by incremental linear regression that showcases improved accuracy.

MATERIALS AND METHODS

The proposed system extracts market data table from online pages. It detects the RSS Feed or "<td>" " <tr>" tags on web pages and transforms it into DOM tree, Further obtains the node values of the "<td>" tags by defining the extraction rules and stored in the database which is given as input to the knowledge extraction engine. The transformed data are preprocessed into the format required for the storage of the data that includes procedures to deal with null values and outliers. Marketing analysts may wish to categorize a group of agricultural commodities to buy, sell, or hold. Neural networks technology is widely used to cluster such quite complex and numerous irrelated variables. SONN are interconnected networks of sovereign processors that, by changing their associations (known as training), discover data in clusters. The incremental linear regression of cost runs to update the regression coefficients when the target data of regression is available. This process operates parallel with the stored data to economize on the number of database accesses**.** The Price forecasting process is a complex system that contains many uncertain factors, it is hardly exactly speaking that it is merely a linear or nonlinear system. Therefore, the modeling of precipitation forecasting should contain some linear and nonlinear characteristics.

Extraction of online dynamic crop prices using DOM

Online websites provides publicly accessible contents as RSS Feed (Really Simple Syndication) for data analysts. RSS immediately publishes regularly updated contents (*i.e.*, price information) in generalized (XML) format [13]. The following algorithm filters the featured price from web pages/RSS using DOM (Domain object model).

1. Achieve the web page of an appointed URL
2. Convert pages into XML and convert into DOM tree
3. Define extraction rules and identify data block using XPATH
4. Extract table using DOM or SAX or Regular Expression
5. Store Extracted data into database

THE IIOAB JOURNAL
ISSN: 0976-3104

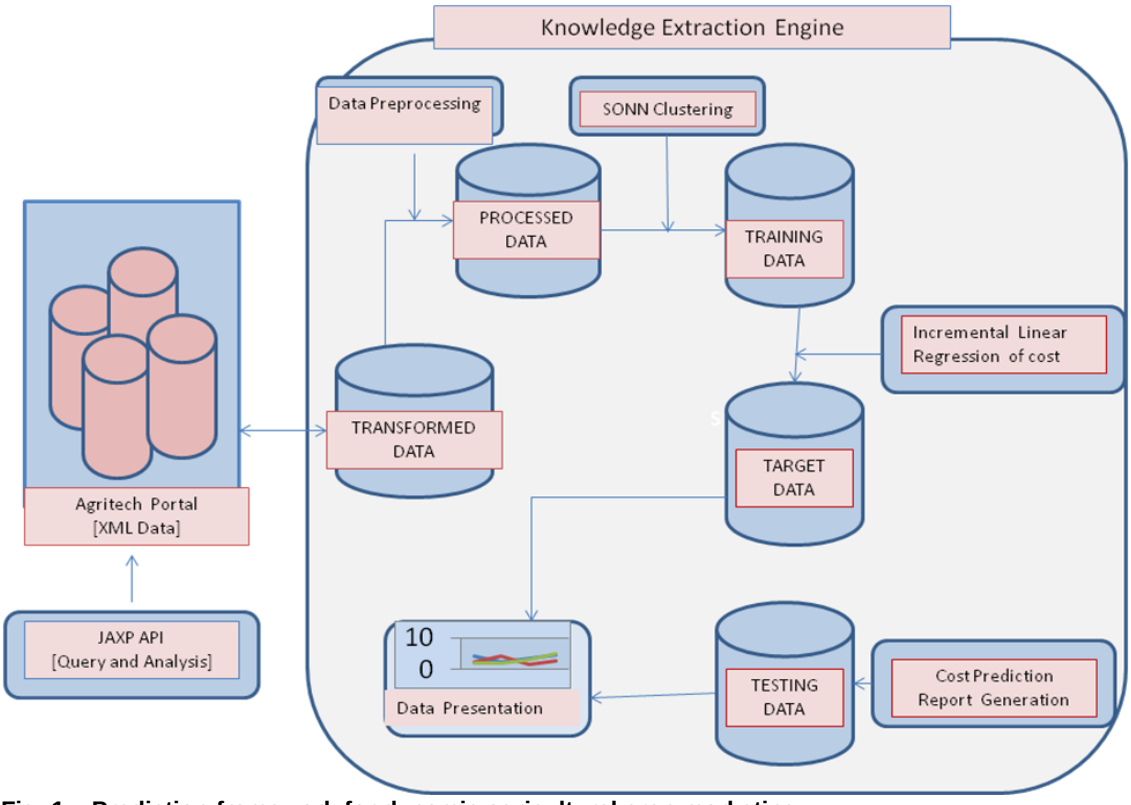

Fig: 1. Prediction framework for dynamic agricultural crop marketing

SONN clustering for specific crop and market categorization

Clustering is a process of unsupervised learning which aims at discovering new set of physical or abstract objects into reference set of classes. The cluster has the characteristics of more similarity within the group and is described as unsupervised learning of a concealed data concept. Let the set of *n* points $\{x_1, x_2, \ldots, x_n\}$ be represented by the set *S* and the K clusters be represented by $C_1, C_2, \ldots, C_K$. Then

$$C_i \neq \phi \qquad for i = 1,\ldots..,k,$$

$$C_i \cap C_j = \phi \qquad for \qquad i = 1,\ldots..k, j = 1,\ldots..,k \qquad and \qquad i \neq j$$

$$and \bigcup_{i=1}^{k} c_i = s$$

In SONN an input pattern has n features and is represented by a vector *x* in an *n*-dimensional pattern space. The network links the input training set to an output pattern. The output space is supposed to be a 1-dimensional or 2-dimensional array of output nodes. The neurons are connected like a lattice, usually a one or two-dimensional array, which is placed in the input set and is linked over the inputs distribution. To each processing unit in the SONN lattice is associated as similar dimension input vectors. Using the weights of each processing unit as a set of coordinates the lattice can be positioned in the input space. During the learning stage the weights of the units change their position and "move" towards the input points. When the map is visualized the inputs can be associated to each cell on the map. The cells closely contains similar values can be considered as a cluster on the map. These clusters are formed during the training phase using the available data.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table :1. Sample Clustered crop price from online RSS Feed [Data courtesy http://www.agmarket.nic.in]

Market	Date	Variety	Modal price	Max price	Min price
Ammoor	1/2/2013	ADT 37	1553	1747	1650
Ammoor	2/4/2015	ADT 36	1639	1671	1655
Ammoor	2/5/2010	ADT 36	1639	1671	1655
Ammoor	4/6/2011	ADT 36	1675	1672	1672
Ammoor	4/8/2015	BPT	1015	1105.54	1978
Ammoor	4/7/2012	Other	915.03	1882.46	1810
Ammoor	3/9/2014	A. Ponni	1625	1684	1480

Incremental approach based marketing price prediction

The mathematical formulation for the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets of the points from the curve is known as least squares fitting technique. Assume that $\{x_i, y_i , i = 1, 2, \cdots, n\}$ are independent bivariate observations from the pair of response–explanatory variables {X, Y }, To describe the relationship between Y and X, a typical regression model is described by

$$E(Y/X) = \mu_0 + \mu_1 x \quad (1)$$

where the intercept μ_0 and μ_1 the slope are unknown regression coefficients. We assume that each observation, Y, can be described by the model

$$Y = \mu_0 + \mu_1 x + \varepsilon \quad (2)$$

Where ε is a random error with mean zero and (unknown) variance σ2. The random errors corresponding to different observations are assumed to be uncorrelated random variables.

The estimates of μ_0 and μ_1 should result in a line that is a best fit to the data. One way to find the values of μ_0 and μ_1 is to minimize the sum of squares of the vertical deviations from the estimated regression line. This criterion is known as the method of least squares. Suppose that we have *n* pairs of observations: (x_1, y_1), $(x_2, y_2) \ldots\ldots (x_n, y_n)$.Using equation(1), we may express the *n* observations in the sample as:

$Y_i = \mu_0 + \mu_1 x_i + \varepsilon_i$ i=1,2,…….n

and the sum of the squares of the deviations of the observations from the true regression line is:

$$L = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \mu_0 - \mu_1 x_i)^2 \quad (3)$$

The least square estimators of μ_0 and μ_1, say $\hat{\mu}_0$ and $\hat{\mu}_1$ must satisfy

$$\left.\frac{\partial L}{\partial \mu_0}\right|_{\hat{\mu}_0, \hat{\mu}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\mu}_0 - \hat{\mu}_1 x_i) = 0 \quad (4)$$

$$\left.\frac{\partial L}{\partial \mu 0}\right|_{\hat{\mu}_0, \hat{\mu}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\mu}_0 - \hat{\mu}_1 x_i) = 0 \quad (5)$$

Simplifying equations (4) and (5), we get,

THE IIOAB JOURNAL
ISSN: 0976-3104

$$n\hat{\mu}_0 + \hat{\mu}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (6)$$

$$\hat{\mu}_0 \sum_{i=1}^{n} x_i + \hat{\mu}_1 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} y_i x_i \quad (7)$$

Equations (6) and (7) are called the least squares normal equations. The solution to the normal equations result in the least square estimators $\hat{\mu}_0$ and $\hat{\mu}_1$. These values are the same as the values in equations (17) and (18) with slight difference in the form of the equation.

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \quad (8)$$

$$\hat{\mu}_0 = \bar{y} - \hat{\mu 1}\bar{x} \quad (9)$$

where $\bar{x}$ is the average of $x_1, x_2, \ldots, x_n$, and $\bar{y}$ is the average of $y_1, y_2, \ldots, y_n$.
In estimating the regression coefficients one can make use of the equations (8) and (9). Along with the values of regression coefficients being stored in the database, one can also store the values of $\sum_{i=1}^{n} y_i x_i$ $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$ and n. With this while new data is being inserted into the database, the above mentioned values can be updated with minimum number of mathematical operations. For example, if a new entry into the records is made with the value (x_{n+1}, y_{n+1}), only the following few operations will be needed thereby reducing the complexity by a large amount:
n=n+1

$$\sum_{i=1}^{n+1} y_i x_i = \sum_{i=1}^{n} y_i x_i + y_{n+1} x_{n+1} \quad (10)$$

$$\sum_{i=1}^{n+1} y_i = \sum_{i=1}^{n} y_i + y_{n+1} \quad (11)$$

$$\sum_{i=1}^{n+1} x_i = \sum_{i=1}^{n} x_i + x_{n+1} \quad (12)$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 + x_{n+1}^2 \quad (13)$$

This process would just need 2 multiplication and 5 addition operations thus reducing the computation time by a great amount. Also, the regression coefficient values can be calculated in parallel with data updation. The only overhead in this method is the memory space needed to store the extra information

RESULTS

THE IIOAB JOURNAL
ISSN: 0976-3104

The experimental analysis in finding significant price for specific crop prediction results are presented in this section. The Agricultural marketing data set is preprocessed successfully by removing duplicate records and supplying unknown values. The obtained data set, resulting from preprocessing, is then clustered using SONN clustering via CMSR - Cramer Modeling Segmentation & Rules. The crop prices dataset we have used for our experiments is obtained from Agritech portal of Tamilnadu agricultural University. With the help of the dataset, the appropriate crop price predictions are extracted using the proposed approach. The training data include the location based separated price data from 2010-2014, is tested against the year 2015 and comparative results are introduced in **Table I**. In **Fiure 2**, the upshot of the model and its results are graphically represented.

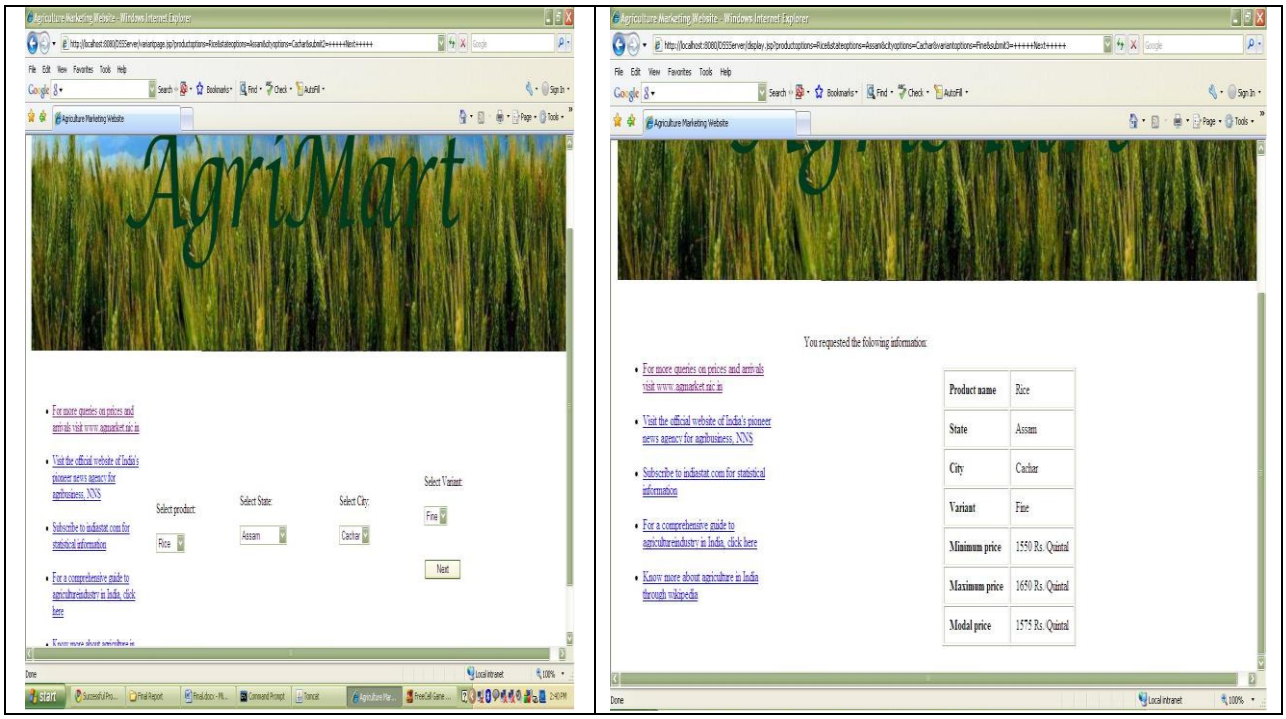

Fig: 2. Predicted modal price for user specified crop from online RSS Feed

The proposed algorithm works in two basic phases. The first phase involves the extraction of desired market data to determine user specified crop price with respect to the number of data set supplied. The second phase is the prediction of crop agricultural prices for marketing. For experimental purpose we randomly divided the target data set into two groups called training dataset (90%) and testing(unknown)dataset (10%) .Before executing, data cleaning and pre processing are performed. The training dataset is used to build the knowledge mining model and the testing dataset is used to detect over fitting of the rules based on the threshold decided by the domain intelligent system. In order to validate our approach we have compared it with some conventional decision making algorithms and the results are shown in **Table– 2 and 3**.

The cause and effect relationship of marketing demand for user specified crop variety with respect to the total demand in the Indian marketing committees was analyzed using incremental linear regression statistical technique and the results are tabulated **Figure– 3**.

Table: 2. Predicted prices of various products [Ammoor market of Vellore district]

Product	Variant	Projected Min (No. of Days)	Projected Max (No. of Days)
Rice	B P T	1917 (1), 1929 (2)	2120 (1), 2107 (2)
Rice	ADT 37	1899 (1), 1898 (2)	1899 (1), 1898 (2)

THE IIOAB JOURNAL
ISSN: 0976-3104

Rice	ADT 36	991 (1), 996 (2)	1017 (1), 1018 (2)
Rice	ADT 36	1017 (1), 1086 (2)	1113 (1), 1117(2)
Rice	Co 43	1056 (1), 1066 (2)	1056 (1), 1064 (2)
Rice	A. Ponni	1063 (1), 1070 (2)	1063 (1), 1070 (2)

Table: 3.Accuracy of predicted prices of various products

Actual Min (No. of Days)	Actual Max (No. of Days)	Accuracy % (Min)	Accuracy % (Max)	Number of Records
1950 (1)	2000 (1)	98.28%	94.34%	50
1900 (1)	1900 (1)	99.94%	99.94%	4
1081 (1), 1081 (2)	1092 (1), 1086 (2)	90.91%,91.46%	92.62%,93.32%	18
1200 (1), 1100 (2)	1300 (1), 1200 (2)	82.01%,98.71%	83.19%,92.57%	12
1081 (1)	1095 (1)	97.30%	96.31%	6
950 (1)	1300 (1)	89.37%	77.74%	5

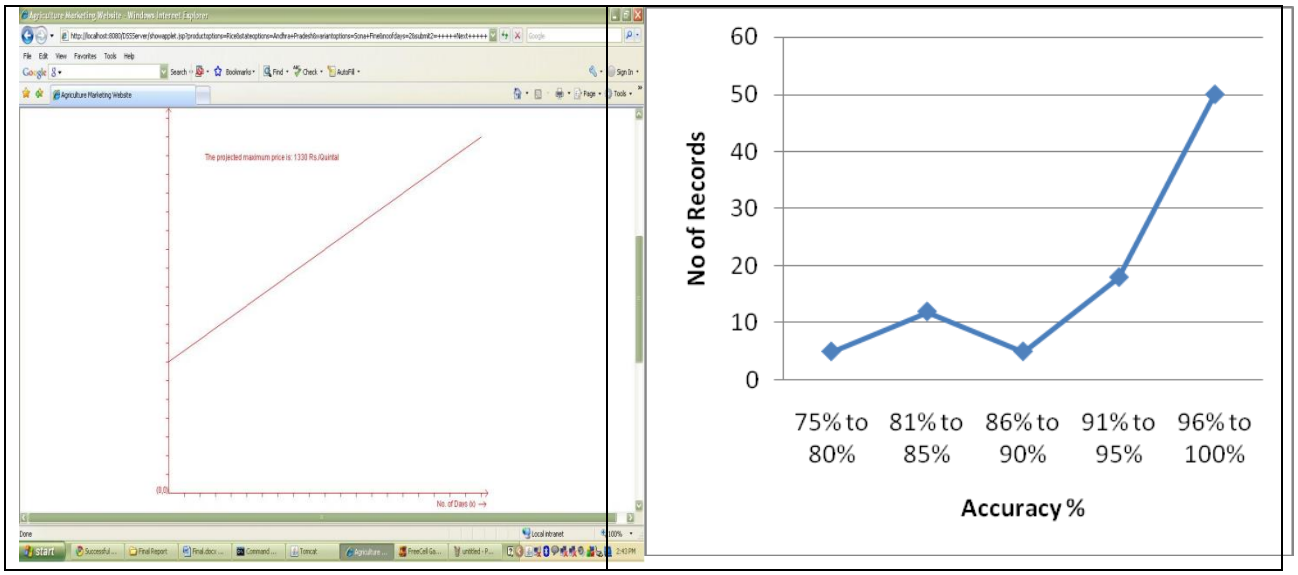

Fig: 3. Plot of accuracy against dataset size

CONCLUSION

The proposed decision support system fetches recent-most prices of regional agricultural products from various web based RSS feeds and analyses the price trend by incremental predictive data mining technique. An increase in the amount of testing data to compute regression coefficient, more is the efficiency of the predicted values. In the commercially available prediction data mining tools all collected data is used recursively to find out the regression coefficients. But this approach has increasing computation complexity and the number of mathematical operations needed to calculate the coefficients. But incremental approach for computing the regression coefficients increases the system's predicting ability and accuracy. This model can be enhanced by incorporating the use of stock availability and demand values for the agriculture products in the process of regression analysis.

CONFLICT OF INTEREST

The author declares having no competing interests.

THE IIOAB JOURNAL
ISSN: 0976-3104

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

Institutional support was received.

REFERENCES

[1] Chun-Yan Y, Jun M, Yu-Yan Z .[2009]Online Price Extraction and Decision Support for Agricultural Products, *Proceedings of the 2009 International Conference on Information Management, Innovation Management and Industrial Engineering* , 04: 337-340.

[2] Alnoukari M, Alzoabi Z, Hanna S. [2008] Applying adaptive software development (ASD) agile modeling on predictive data mining applications: ASD-DM Methodology. In *Information Technology, 2008. ITSim 2008. International Symposium on* , 2: 1-6. IEEE.

[3] Yeh IC, and CH Lien. [2009].The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480.

[4] Hamzacebi C, DnAkay and F Kutay.[2009]Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. *Expert Systems with Applications*, *36*(2): 3839–3844.

[5] Doan XT, PT Kidd, RGoodacre, BD Grieve.[2008]Regression analysis for supply chain logged data: A simulated case study on shelf life prediction. *International Conference* in *Signal Processing, IEEE,Beijing*, pp. 2717–2720.

[6] Jovanovic N, V.Milutinovic and Z.Obradovic.[2002].Foundations of Predictive Data Mining. Neural Network Applications in Electrical Engineering,*Sixth seminar on Neural Network Applications in Electrical Engineering*, IEEE , 53 – 58**.**

[7] Alampalayam, S., and Kumar, A,(2004) An Adaptive and Predictive Security Model for Mobile Ad hoc Networks", *Wireless Personal Communication Journal* 29: 263-281.

[8] Wu,J., L.Huang and X.Pan. (2010).A novel bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting. *Third International Joint Conference* in *Computational Science and Optimization*,IEEE, Huangshan, Anhui ,pp. 466–470.

[9] Rub G. [2009] Data mining of agricultural yield data: A comparison of regression models. Advances in Data Mining. *Applications and Theoretical Aspects*, Springer, Berlin, pp. 24–37.

[10] Patil SB and YS Kumaraswamy. [2009] Intelligent and effective heart attack prediction system using data mining and artificial neural network. *European Journal of Scientific Research*, 31(.4): 642–656.

[11] Bellei E, D Guidotti, R Petacchi, LMnReyneri, and I Rizzi. [2001].Applications of neuro-fuzzy classification, evaluation and forecasting techniques in agriculture. *European Symposium on Artificial Neural Networks*, Belgium, pp. 403–408.

[12] Tan CN. [1997].An Artificial Neural Networks Primer with Financial Applications Examples in Financial Distress Predictions and Foreign Exchange Hybrid Trading System. Bond University.

[13] Lavanya K, Raguchander T, and Iyengar NCS. [2013] SVM Regression and SONN based approach for seasonal crop price prediction. *International Journal of u-and e-Service, Science and Technology*, *6*(4), 155–168.

[14] Wuthrich B, V Cho, S Leung, D Permunetilleke, K.Sankaran and J Zhang. [1998] Daily stock market forecast from textual web data. *International Conference in Systems, Man, and Cybernetics*, IEEE, San Diego, CA,pp. 2720–2725.

[15] Heydari M, and PHTalaee. [2011] Prediction of flow through rockfill dams using a neuro-fuzzy computing technique. *Journal of Mathematics and Computer Science*, *2*(3):515–528.

[16] Venugopal V, and Baets W. [1994]Neural networks and statistical techniques in marketing research: A conceptual comparison. *Marketing Intelligence & Planning*, *12(*7):30-38.

[17] Fan X, Li S, and Tian L. [2015] Chaotic characteristic identification for carbon price and an multi-layer perceptron network prediction model. *Expert Systems with Applications*, 42*(*8): 3945–3952.

[18] Nie G, Rowe W, Zhang L, Tian Y, and Shi Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12): 15273–15285.