Implementing Enhanced AdaBoost Algorithm for Sales Classification and Prediction

Von Kirby P. German, Bobby D. Gerardo, and Ruji P. Medina

***Abstract*—In today's data driven economy, retail businesses rely on information systems that monitor and process their daily transactions. These huge amount of data being processed on a day-to-day basis can be utilized to forecast sales for inventory management, and decision-making. In this paper the AdaBoost algorithm is used in classification and prediction of data. While it is known to be capable of processing both variable and numerical values, it is quite certain that processing data, represented as facts, is faster in digital form. This allows the algorithm to process the conditions digitally. The original raw facts presented in this study are in variable forms. To better improve performance, the first part of the algorithm converts the facts and represent them in numerical, computable values. This allows the rest of the algorithm to process the entire set of data numerically, and evidently faster, resulting to a better performance of the algorithm. The use of this innovative technique improves the performance of the extraction methods used in data mining which is very important for business support and decision making. The future of this research can explore the development of a Decision Support System (DSS) for the purpose of predictive analysis and support of business decisions.**

***Index Terms*—AdaBoost, classification, prediction, data mining.**

I. Introduction

Most of the retail businesses plan to attract the customers to the store and make profit to the maximum extent by them. Once the customers enter the stores they are attracted then definitely they shop more by the special offers and obtain the desired items which are available in the favorable cost and satisfy them. If the products as per the needs of the customers then it can make maximum profit the retailers can also make the changes in the operations, objectives of the store that cause loss and efficient methods can be applied to gain more profit by observing the history of data the existing stores a clear idea of sales can be known like seasonality trend and randomness. The advantage of forecasting is to know the number of employees should be appointed to meet the production level. Sales drop is bad thing forecasting sales helps to analyze it and it can overcome through the sales drop to remain in the competition forecast plays a vital role [1].

The advances for producing and collecting data have been advancing rapidly. At the current stage, lack of data is no longer a problem; rather the inability to generate useful information from data is. The explosive growth in data and database brings about the need to develop new technologies and tools to process data into useful information and knowledge intelligently and automatically. Data mining, therefore, has become a research area with increasing importance [2]. It is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems [3].

Data mining has great importance in today's highly competitive business environment. It is largely used in several applications such as understanding consumer research marketing, product analysis, and demand and supply analysis, electronic commerce and so on. In this research the extraction of data from a database is simulated using adaptive boosting method for sales classification and prediction.

II. Related Works

Researchers in the field of data mining always try to find innovative techniques so as to improve the performance of the extraction methods used in data mining as they usually use history of the different transactions done in finding the data as it will be useful for future use. This data collection can be used by them to predict the customer behavior and their interests [4].

The future and emerging trends in managing information systems is very important in today's digital economy. Data mining techniques are best suited for the classification, useful patterns extraction and predications which are very important for business support and decision making. Historical inventory data can indicate market trends and can be used in forecasting which has great potential in decision-making and strategic planning [5].

AdaBoost is a proven powerful algorithm for classification that is widely used in various fields such as biology, computer vision, and speech processing. Unlike other powerful classifiers, such as SVM, AdaBoost can achieve similar classification results with much less modification of parameters or settings [6].

In other studies, the AdaBoost algorithm is integrated in various kinds of weak classifiers to enable learning to get a strong classifier. It takes advantage of different kinds of classifications and makes sure the result to a maximum level of accuracy. AdaBoost integrated several classification

Manuscript received September 20, 2017; revised November 3, 2017. This research is a dissertation requirement for the completion of the doctoral degree in Information Technology of the Technological Institute of the Philippines, Quezon City, Philippines.

V. K. P. German is with the Far Eastern University, Manila, Philippines. (e-mail: vkgerman@feu.edu.ph).

B. G. Gerardo is with the West Visayas State University, Iloilo City, Philippines. (e-mail: bgerardo@wvsu.edu.ph).

R. P. Medina is with the Technological Institute of the Philippines, Quezon City, Philippines. (e-mail: ruji_p_medina@yahoo.com).

doi: 10.18178/ijtef.2017.8.6.577

methods is used to classify the samples and get a stable classification rule. An empirical study of classifying retail outlets of a tobacco market in a city in China is done in order to prove the method is feasible [7].

Other applications of AdaBoost is in vehicle detection. The results of the experiment demonstrate that the rapid incremental learning algorithm of AdaBoost is designed to significantly improve the performance of the algorithm, and it also yield better or competitive vehicle classification truths compared with several state-of-the-art methods, presenting their possible actual applications [8].

With the huge amount of data that are available in businesses, this can be statistically analyzed to conduct predictive analytics. While artificial intelligence can help computer systems learn on their own through machine learning, and identify potential customers, track leads and thereby come up with the most precise prediction about their future course of action [9]. Predictive Analytics can provide the framework within which intelligent decision making could be made possible with higher business efficiencies.

III. Classification and Prediction

The predictive techniques used in data mining can be divided into two major operations: classification (or discrimination) and prediction (or regression). The aim of these two operations is to estimate the value of a variable (called the 'dependent', 'target', 'response', 'explained' or 'endogenous' variable) relating to an individual or an object as a function of the value of a certain number of other variables relating to the same individual, identified as the 'independent' variables (also called the 'explanatory', 'control' or 'exogenous' variables) [10].

Classification is a technique that uses data to generate a model which assigns data items to one of several distinct categories. This machine leaning method is capable of processing large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The model can be used to make predictions based on the data items which can be used in businesses like hospitals, whether a patient has dengue fever (yes, no) based on various medical test data; and banks, whether a customer is qualified for a loan determined by his risk category (low, medium, high) based on the applicant's financial history [11].

There are many different classification algorithms and techniques, including SVM, Naive Bayes, neural network and logistic regression. Adaptive Boosting Classification is a technique used to predict results by combining simple rules derived from a set of training data. Each simple rule is given a weight and each new input is predicted by taking the weight of each rule to arrive at a result. The capability of each rule to predict an outcome is improved by combining them with others, hence it is called boosting. Throughout this paper, Adaptive Boosting is implemented which is a far more effective technique than using the traditional ambiguous single rule technique [12].

The Classification process involves following steps [13]:

a) Create training data set.
b) Identify class attribute and classes.
c) Identify useful attributes for classification (Relevance analysis).
d) Learn a model using training examples in Training set.
e) Use the model to classify the unknown data samples.

IV. Adaptive Boosting

AdaBoost is a popular ensemble classifier that combines the output of several weak learners to obtain a strong classifier. The weak learners could be any other classifier, such as a decision stump, decision tree, logistic regression, SVM, etc. There are two requirements placed upon the weak learners:

a) The accuracy of these weak learners should at least be better than random guessing for arbitrary, unknown distributions of the training data. For instance, in a binary classification problem, the training data accuracy of the weak learner (which is the percentage of correctly classified examples) should be strictly greater than 0.5.
b) The weak learner should be able to handle weighted training examples. Given these constraints on the weak learners, AdaBoost provides a framework to combine these weak learners to obtain a final classifier whose accuracy is significantly higher than the accuracy of any single model, the weak learners.

In each iteration, AdaBoost attempts to improve upon its errors for particular examples in the training set by minimizing the errors for those in the previous model. In each iteration, the weak learners place higher weights on training examples that have been particularly difficult, allowing it to focus on all of the data, rather than ignoring a subset.

The pseudo-code of the AdaBoost learning algorithm:

Let us assume we have N training examples {(x1, y1), (x2, y2) … (xi, yi)… (xN, yN)} where xi represents the feature vector for the ith training example and yi represents the corresponding label (0 or 1). Let us also initialize a set of weights, wi, over the set of training examples to be 1/N, equal for each training example initially.

For each iteration t until T:

Step 1: learn a weak classifier (ht) with current set of weights wi

Step 2: compute the training error (εt) of the classifier ht

Step 3: define $\alpha t = 0.5*(\ln(1- \varepsilon t)/ \varepsilon t)$

Step 4: increase the weights on the misclassified examples by a factor of $e \alpha t$ and renormalize the weights wi

During each iteration, the set of weights (wi) are adjusted in such a way that in the next iteration there is more emphasis on mis-classified examples in the previous round. This ensures that complementary features (rules) are picked during the different rounds of AdaBoost. As a result, AdaBoost's key benefit is that it can create a non-linear decision boundary for the classification problem at hand by combining the decision boundaries of these weak learners from different iterations.

The final AdaBoost classifier is then given by $H(xi) = \Sigma t\ \alpha t$ ht(xi) where $ht(xi)$ is the decision of the tth weak classifier and αt is the corresponding weight given to that decision in the final classifier. If the weak learners (ht s) are decision trees, then you can intuitively imagine the final classifier as a way to combine multiple rules with a certain weight (αt) for each of the rules [14].

V. Methodology

The purpose of the enhanced algorithm is to predict whether if a product will be sold or not considering the season and discount variables. As shown in Fig. 1, the sales and inventory data or the dataset is converted from string data into zero-based integer form for more efficient processing and then stored into machine memory as a matrix. The last part or output of the program shows the classification model used to make a sales prediction for a particular product (item) based on whether it is summer or rainy season, and when it is discounted, in promo or in regular price. The outcome to be predicted has only two values, yes or no.

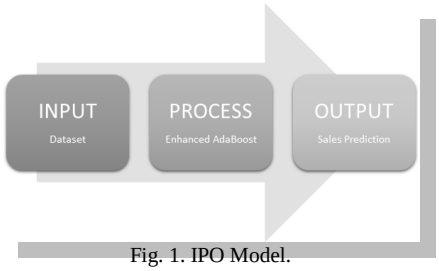

Fig. 1. IPO Model.

VI. Program Structure

Fig. 2 shows the Main method, which begins by setting up hardcoded strings for the fields or features. Then it reads off the values for these features from a locally attached database file. This composes the training data that will be used in the entire program.

RAW TRAINING DATA

Product	Season	Discount	Sold
Item4	Summer	Promotion	Yes
Item4	Rainy	Discounted	Yes
Item2	Summer	Regular	Yes
Item2	Summer	Discounted	Yes
Item1	Rainy	Promotion	Yes
Item3	Summer	Discounted	Yes
Item3	Rainy	Regular	No
Item3	Summer	Regular	No
Item1	Rainy	Discounted	No
Item4	Rainy	Promotion	No

Fig. 2. Values that were read from the local database.

```
TRAINING DATA IN INT FORM
3 0 2 -> +1
3 1 1 -> +1
1 0 0 -> +1
1 0 1 -> +1
0 1 2 -> +1
2 0 1 -> +1
2 1 0 -> -1
2 0 0 -> -1
0 1 1 -> -1
3 1 2 -> -1
```

Fig. 3. Training data converted to Integer.

The method RawTrainToInt as illustrated in Fig. 3, converts the training data in string form to zero-based integers and stores those integers into an int[][] matrix named train. RawTrainToInt calls a helper method named ValueToInt.

The train matrix has the dependent variable values (Result) stored in the last column. For the purpose of this sample program, the entire training data set into is stored and accessed from machine memory for faster processing.

The program then determines the weak learners presented in Fig. 4 using method MakeLearners and a program-defined class Learner. This class contains six fields namely feature, value, predicted, error, epsilon and alpha. The feature field holds an integer that indicates which independent variable is the key to the learner. The value field holds an integer that indicates the value of the feature. The predicted field is -1 or +1, depending on whether the actual category for the feature value is Yes or No. The error field is type double and is the raw error rate associated with the weak learner on the training data. The epsilon field is a weighted error term. The epsilon for a weak learner is an error term that takes into account the internal D weights assigned to each training item. The epsilon values are used by the adaptive boosting algorithm to compute the alpha weights. To summarize, there are two sets of weights used in adaptive boosting classification. The alpha weights assign an importance to each weak learner and are used to determine an overall prediction. An epsilon error is an internal error associated with a weak learner that's used to compute the alpha weights. Each training tuple has an internal weight (given the name D in adaptive boosting literature) that's used to compute the epsilon errors.

```
WEAK LEARNERS
[0] IF Product IS Item2 THEN Sold IS Yes   (raw error = 0.00)
[1] IF Product IS Item3 THEN Sold IS No   (raw error = 0.33)
[2] IF Product IS Item4 THEN Sold IS Yes   (raw error = 0.33)
[3] IF Season IS Summer THEN Sold IS Yes   (raw error = 0.20)
[4] IF Season IS Rainy THEN Sold IS No   (raw error = 0.40)
[5] IF Discount IS Regular THEN Sold IS No   (raw error = 0.33)
[6] IF Discount IS Discounted THEN Sold IS Yes   (raw error = 0.25)
[7] IF Discount IS Promotion THEN Sold IS Yes   (raw error = 0.33)
```

Fig. 4. Weak learners generated by makelearners method.

```
Adaptive Boosting Algorithm found 8 good learners and associated alpha values.

GOOD LEARNERS AND THEIR ALPHA VALUES
[0] 6.91
[1] 2.31
[2] 0.63
[3] 3.15
[4] 0.59
[5] 4.49
[6] 0.68
[7] 0.63
```

Fig. 5. Best Learners and their Alpha values.

After the weak learners have been created, the program calls method MakeModel. This method is the heart of this adaptive boosting algorithm. The net result is a sorted List as indicated in Fig. 5, named bestLearners, of the indexes of the learners that were assigned alpha values.

The Main method then predicts the outcome through the method called Classify which is shown on Figure 6 below. This takes input from the text fields I the form which is associated for each of the features created at the beginning of the program.

Finally, the main program finishes with collating the outputs, composed by the final result along with the list and matrices generated throughout the program. These outputs are then printed into the form as the program exits. The sample program result is illustrated in Fig. 7.

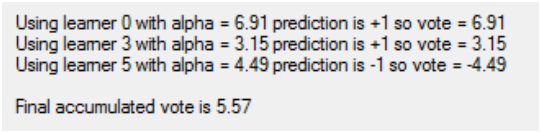

Fig. 6. The outcome is predicted by the classify method.

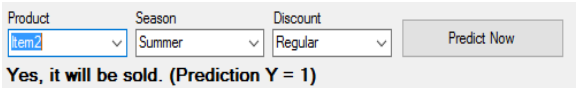

Fig. 7. Final outcome as displayed in the output.

VII. Results

An important enhancement on the AdaBoost algorithm presented in this paper is dealing with data that has numeric values. For example, suppose that the values for the discount feature, instead of being categorical "Regular," "Discounted" and "Promotion," were numeric, such as 1.5, 3.0 and 9.5. One of the major advantages of adaptive boosting classification compared with some other classification techniques is that adaptive boosting can easily handle both categorical and numeric data directly. You could create a dedicated learner class that has a friendly description similar to "if Discount is less than or equal to 5.5 then Result is Not Sold," or a more-complex learner along the lines of "if Discount is between 4.0 and 6.0 then Result is Sold." Adaptive boosting classification is best used when the dependent variable to be predicted has just two possible values. However, advanced forms of adaptive boosting can deal with multinomial classification.

VIII. Recommendation

Predictive analytics is very important in business operations and decision-making processes. The results of this research can be further used in developing Decision Support System (DSS) for Sales and Inventory Management. Tech-savvy retailers are looking towards predictive analytics to unleash the power data. Access to the right data mining and predictive analytics solutions can help a business take insightful decisions in today's volatile economic climate. Businesses use predictive analytics to set the bar in customer retention, inventory optimization and low-cost promotions which drive increases in profitability and market share.

References

[1] A. Harsoor and A. Patil, "Forecast of sales of walmart store using big data applications," *IJRET: International Journal of Research in Engineering and Technology*, 2015.
[2] S. J. Lee and K. Siau, "A review of data mining techniques," *Industrial Management & Data Systems*, pp. 41-46, 2001.
[3] J. Ranjan, "Applications of data mining techniques in pharmaceutical industry," *Journal of Theoretical and Applied Information Technology*© 2005-2007.
[4] M. Al-Noukari and W. Al-Hussan, "Using data mining techniques for predicting future car market demand," *IEEE*, 2008.
[5] V. Ware and H. N. Bharathi, "Decision support system for inventory management using data mining techniques," *International Journal of Engineering and Advanced Technology*, vol. 3, issue 6, August 2014.
[6] R. Kaur and V. Chopra, "Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, issue 7, July 2015.
[7] K. Lu *et al*., "A retail outlet classification model based on adaboost," in *Proc. International Conference on Soft Computing Techniques and Engineering Application*, September 25-27, 2013.
[8] J. Bergstra *et al*., "A rapid learning algorithm for vehicle classification," *Information Sciences: An International Journal*, vol. 295, issue C, pp. 395-406, February 2015.
[9] Ketan. (November 2016). The Role of Predictive Analytics in Marketing & Sales. [Online]. Available: https://yourstory.com/2016/11/aa2147be3e-the-role-of-predictive-analytics-in-marketing-sales/
[10] Stephane Tuffery. 2 *Data Mining and Statistics for Decision Making*, UK: John Wiley & Sons, Ltd., 2011, ch. 11, pp. 521-528.
[11] S. S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," *Oriental Journal of Computer Science and Technology*, vol. 8, no. 1, pp. 13-19, April 2015.
[12] James McCaffrey. (April 2013). CLR - Classification and Prediction Using Adaptive Boosting. [Online]. Available: https://msdn.microsoft.com/en-us/magazine/dn166933.aspx
[13] T. C. Sharma and M. Jain, "WEKA Approach for Comparative Study of Classification Algorithm," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, issue 4, April 2013.
[14] R. Radhakrishnan. (January 29, 2015). *Implementing AdaBoost on MPP for Big Data Analytics*. [Online]. Available: https://content.pivotal.io/blog/implementing-adaboost-on-mpp-for-big-data-analyticsar

Von Kirby P. German was born in Quezon City, Philippines, he holds bachelor's degrees in business management majoring in entrepreneurship, and Information Technology from San Beda College and Saint Paul University respectively. He also has a diploma in computer science and master's degree in technology management from the University of the Philippines. At present, he is working on his dissertation in cloud computing and decision support systems for his doctoral degree in information technology at the Technological Institute of the Philippines, Quezon City.

He is a professor of digital marketing and e-commerce, and assistant program head for Business Administration at the Institute of Accounts, Business and Finance of the Far Eastern University, Manila. He specializes in the fields of business development, information systems project management, web design and development, digital and web marketing, and integrated marketing strategies gained through his work affiliation with leading local and international information technology companies. Currently, he is the director of Academic Partnerships of the Sales and Marketing Institute Philippines, and Certified Marketing Professional, Certified Sales Professional, and Certified Financial Markets Professional.

Assistant Prof. German is a member of the scientific program committee of the 14th National Conference on Information Technology Education (NCITE 2016). A strong advocate on digital literacy in the classroom, he is awarded as microsoft certified educator, microsoft innovative educator expert and microsoft ambassador in Education by Microsoft Philippines.

Dr. Bobby D. Gerardo is currently holding a position of Professor 6 at the College of ICT and the Vice President for Administration and Finance of West Visayas State University, Iloilo City, Philippines.

He has published more than 75 research papers in the national and international journals and conferences and received numerous research awards due to his quality published works. He is a referee to international conferences and journal publications such as in IEEE Transactions on Pattern Analysis and Machine Intelligence, Transactions on Knowledge and Data Engineering and Transaction on Computational Biology and Bioinformatics. His research interests lie in the area of IT protection and security, distributed systems, telematics, CORBA, data mining, web services and ubiquitous computing.

Dr. Ruji D. Medina is the Dean of the Graduate Programs and concurrent Chair of the Environmental and Sanitary Engineering Program of the Technological Institute of the Philippines, Quezon City. He holds PhD in Environmental Engineering from the University of the Philippines with sandwich program at the University of Houston, Texas.

He counts among his expertise environmental modeling and mathematical modeling using multivariate analysis.