The Realization of a Type of Supermarket Sales Forecast Model & System

Fen Cai
College of Mathematics and Computer Science
Quanzhou Normal University
Quanzhou, China
e-mail: caifen@qztc.edu.cn

YanMin Luo
College of Computer Science and Technology
Huaqiao University
Xiamen, China
e-mail: lym@hqu.edu.cn

***Abstract*—This essay solves the problem of supermarket sales forecast. In allusion to that the forecast targets are numerous, as well as the features of high volatility and having evident seasonality of the forecast series, the system adopts combined forecast method, and designs a kind of forecast method selecting algorithm integrating bagging approach and significance statistics checking approach. The system is realized and through testing, it is able to make effective forecast on most of the forecast targets.**

Keywords-sales forecast; decision-making support; data mining; ensemble method

I. INTRODUCTION

The research group has developed a supermarket sales forecast system based on data mining technology and statistical analysis method to help enterprise managers rationally adjust the structure of commodities and make decisions on purchase, sales, inventory and promotion. Both the demands and experimental data of this research subject come from a big chain supermarket located in Quanzhou City, Fujian Province.

The forecast of sales is a matter of time series forecast. There are many methods can be used in time series forecast. The main forecast methods for stationary series are Moving Average Approach and Simple Exponential Smoothing Approach, while to forecast the linear trend, the classical A Linear Regression or Holt's model can be used. With regards to the forecasting of non-linear trend regression, Polynomial Regression and Exponential curve model can be used. If the series include seasonal elements, then Winter's model or Multiple Regression Predicting Method with seasonal dummy viable can be tried. As to series including multiple factors, it would be more appropriate to choose Decomposing Prediction and ARIMA model. Besides, there are Grey Prediction model, Neural Network and Hybrid Neural Network etc. [1-3] In case of a specific forecasting object, deep discussion can be made to find out the most appropriate forecast method or combination of methods, so that the forecasting value and the observed value would have the best goodness-of-fit. In this system, the forecasting objects are being chosen and typed in by the user. Different commodity, different category and even different commodity of same category in the supermarket have different data pattern. And the data pattern can be different as the length of selling period differs. The test results show that, in the face of all possible sales series model of numerous commodity objects, a specific forecast method can not produce good results always. The solution is to use combination of forecast methods, i.e. to set up a library of forecast methods, having several forecast methods included in the library, and to select and empower the methods. Presently, research of combined forecast focuses on the matter of empowerment [4-5]. [6] proposes a Forecast Method Selection Model with Personnel Selection Approach as its basic theory. While [7] proposes a forecast model based on Linear Regression and Exponential Smoothing. None of these models are suitable to be applied to the sales data of supermarket. This essay puts suitable methods in the library of forecast methods and designs a selection algorithm of forecast methods integrating Bagging Theory and Statistical Testing Method according to the characteristics of commodity data in supermarket. In allusion to the forecast series, the forecast model or combination of models with best performances will be found through this algorithm. And then the forecast and error estimation will be made according to the model.

II. SUPERMARKET SALES FORECAST MODEL

A. The Forecast Methods in the Library

With regards to the forecast of commodities sales volume, which methods are more suitable? What kind of method should be used in the forecast of time series depends on the pattern of data. Whether the data is stationary or non-stationary, and whether the data includes trend, seasonal or circulating factors? The scale of historical data, and the forecast period required by the Demands all have impacts on the selection of forecast method. Through study, the time series of commodities sales have the following characteristics:

1) Most of the sales series have obvious seasonal or cyclical variation. Therefore, it would be better for the selected method to have the ability of separating seasonal factors.

2) Generally the volatility of series is relatively high. The commodities sales fluctuate significantly; generally it would not be a smooth curve, but an irregular zigzag form. Therefore, Easy Exponential Curve and A linear regression are not suitable for the forecast of commodities sales volume.

3) The system retains all data from the year 2008 up to now. The scale of data fulfils the requirements for at least 4 periods as requested by Winter's model, Multiple Regression with Seasonal Dummy Variable, Decomposing Prediction and ARIMA Model etc.

4) The Demand requires for the forecast of at least 3 post-stage data. So, neither Moving Average nor Easy Exponential Smoothing is appropriate, because their forecast period are relatively short.

According to the above-mentioned characteristics of series, the system attempts to bring Decomposing Prediction, Winter's model and ARIMA(p,d,q)(P,D,Q)^s model (autoregressive integrated moving averages) into the library of forecast methods. All the three methods have the ability of separating seasonal factors, and have at least mid-term forecast period. Besides, the data scale meets the three methods' requirements for data scales. The library of forecast methods also includes linear regression, Holt Exponential Smoothing and Multi-stage Curve, because when the Decomposing Prediction decomposes the series into error component, seasonal component and trend and cycle component, these three methods would be applied in the modeling of abstracted trend and cycle component. At the meantime, attempts are being made in the modeling of trend and cycle component using ARIMA model without seasonal factors.

B. The Selection Algorithm Design

Firstly, the system adopts the approach of truncation to get 95% of the data series as the training set, leaving the remaining 5% to be used for the assessment of model. It tries every method in the library of forecast methods, using threshold value R^2 to conduct the first round of selection. The model fulfilling the set threshold value will be kept back. The model with slightest root-mean-square error will be picked out from the remaining method ensemble, assuming to be M_{min}, the other model will then going through statistical significance check with M_{min} one by one, this is the second round of selection. Eventually, the non-statistical significance model against M_{min}, together with M_{min} is the model ensemble with best performance. The retaining model and M_{min} together will conduct forecast on the original data series, and conduct error calculation for the truncated data. The forecast average and error average will be exported as the final conclusion. The model ensemble, together with model parameter and error value will all be saved into the model database, and being used in the forecast of same forecast object within short time.

The algorithm summarizing of forecast method selection is described as follow:

Algorithm :Selection Algorithm
Input:
- X:given data series
- R^2: threshold value
- Forecast methods in the library
- Test of Significance Approach

Output:
- X': forecast value
- M: The selected model or combination
- $Err(M)$:forecast error

Procedure:
BEGIN
(1) Adopt the approach of truncation, dividing X into training set D and testing set T;
//The first round of selection
(2) m:=0 ;
(3) For i=1 to l do //for all methods in the library
(4) Using training set D to train the model M_i ;
(5) If $R^2(M_i)>R^2$
(6) Then
(7) M_i is kept;
(8)m:=m+1;
(9) Endif
(10)Endfor
//The second round of selection
(11) $M_{min}:=M_1$;
(12) For j=1 to do
(13) Compute RMSE(Mj);
(14) If RMSE(Mj)<RMSE(Mmin)
(15) Then $M_{min}:=M_j$;
(16) Endif
(17) Endfor
(18) n:=m ;
(19) For j=1 to do
(20) If M_j is significant difference against M_{min}
(21)Then M_j is sifted out;
(22) n:=n-1;
(23) Endif
(24) Endfor
(25) Using every model in M to carry out forecast on X, get the forecast value X_j' ;
(26) Using every model in M to carry out error computation on T, get the RMSE $err(M_j)$;
(27) Output $X'=avg(X_j')$;
//the average value of each forecast value will be given as the final output
(28) Output forecast error $err(M)=avg(err(M_j))$;
//and the error of final output is the average value of each error
(29) Output the adopted model or combination of models $\boldsymbol{M}$
END

What worth further explanation in the algorithm are:

1) The training set obtained using the approach of truncation.

The general practice for model error estimation is dividing the dataset into training set and testing set by sampling, using training set to train the model, and using testing set for error estimation. However, sampling is not suitable for the testing of time series model because only continuous series set without time breakpoint can be used as training set for the study of model. The system adopts the approach of truncation, dividing the whole sample zone into front part and later part. The front part is used as training set

for the study of model, while the later part is used as testing set for the assessing the model and providing error value of the assessment. To ensure enough sample for the study of model, the system cut out 5% of the data as testing set.

2) Increasing the accuracy of forecast by using the approach of bagging [8].

Through the first round of selection, theoretically, the model with smallest error value will be chosen for the forecast of given sample data, and the final forecast value will be obtained. Yet the testing proves that, if two or more than two models both perform good, then it would be better for every model to participate in the forecast of given sample data, and take the average value of these forecast values as the final result of the final forecast return value. This approach is bagging of ensemble method, usually being used to increase the accuracy of predictor and classifier.

Then, how to identify the model with good performance? Among the remaining models after the first round of selection, the algorithm first choose the model with smallest root-mean-square error, and all other models will conduct statistical significance check with it one by one.

In the fitting models, suppose there are *k* series spots, then the error of every series spot can be taken as different independent sample in probability distribution, and generally it follows the pattern of distribution *t* with *k*-1 degree freedom, in which, *k* equals to the number of series spots. Conducting presumption testing *t-test*, suppose the testing is passed, then these two models is "identical", or the error rate of the average value of the two is "zero". If this presumption is declined, it shows that the difference of these two models is statistical significant, i.e. they have differences, and then the model with higher error rate will be sifted out. And every non-statistical significant model will be retained. Then the corresponding forecast values can all be regarded as inputs of bagging algorithm for increasing the accuracy of the forecast.

To forecast the same data series using M_1 and M_2 respectively, the statistics of the significance testing *t* of the error shall be calculated according to the following formula:

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{\mathrm{var}(M_1 - M_2)/k}} \quad (1)$$

In this formula, $\overline{err}(M_1)$ is the average value of M_1 model's error, $\overline{err}(M_2)$ is the average value of M_2 model's error, and $var(M_1\text{-}M_2)$ is the variance of the two models' difference:

$$\mathrm{var}(M_1 - M_2) = \frac{1}{k}\sum_{i=1}^{k}[err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2))]^2 \quad (2)$$

Significance Level *Sig* is chosen to conduct presumption testing for the retained 5%.

III. The Analysis And Realization Of The Sales Forecast System

A. System Architecture

The system architecture drawing of the sales forecast system is shown as Fig. 1:

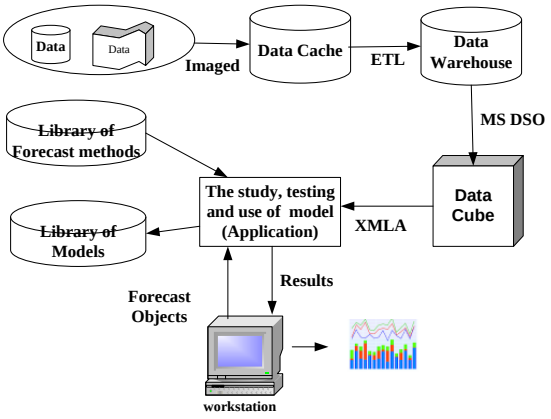

Figure 1. System architecture of the Sales forecast system.

The procedures are described as follow:

1) All heterogeneous data will be completely imaged in the data caching zone.

In order not to cause overdue burden to the supermarket's operation database, the setting up of data caching zone is quite necessary. The caching zone stores blank database. All sheet structure and data are established and imported by the data abstracting program, totally identical with all kinds of heterogeneous data. The extracting of data adopts the approach of full flow extracting, and it is realized through auto extracting program, while the extracting frequency can be set manually, either daily, weekly or other value is acceptable.

2) The establishment of data warehouse.

Through the combining and summarizing, computing viewing, integrity check as well as cleaning and loading of the caching data, the data warehouse SALES is established. The data warehouse includes data sheet with preset structure.

3) Setting up data cube.

The data in the data warehouse will be gathered as data cube saved in the server of Analysis Services. The star data cube being set up based on the sales forecast model is shown below as Fig. 2:

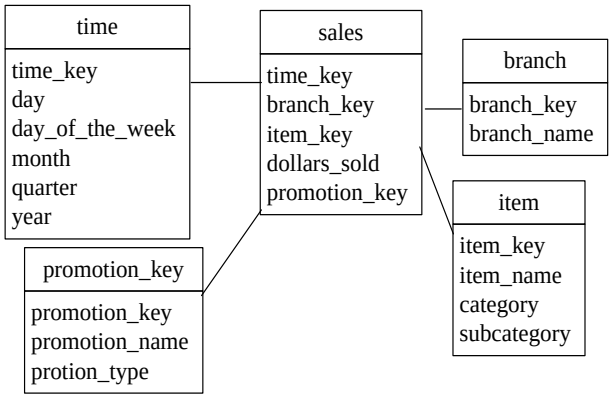

Figure 2. Sales Star

4) The study, testing and use of forecast model.

The forecast algorithm as introduced in Sec. 2 shall be used for the train and forecast of model with regard to the forecast object as chosen by the client, returning the forecast value and forecast curve to assist the manager in decision-making. Other model parameters, including the model and the error value will all be saved in the model database, and being applied in the forecast of same commodity in the future, or being used for reference in the forecast of other commodity of the same niche.

B. System Developing Platform

The system uses MS SQL Server as the server of date-caching zone and data warehouse, and uses Analysis Services as the data cube server. Both server and client program use .Net structure, and were written in C# language. DSO (Microsoft Decision Support Object) module was introduced in the C# language. Through the use of this module, data cube can be set up in Analysis Services. The system uses XMLA (Microsoft XML for Analysis) to realize the barrier free interaction of multi-dimension data set and mining program in Analysis Services.

IV. ANALYSIS WITH EXAMPLE AND TESTING OF ALGORITHM EFFECTIVENESS

A. Testing 1

The existing data of monthly sales of "Darlie Toothpaste 225G"after promotion reduction at some outlet is listed as below:

TABLE I. MONTHLY SALES OF "DARLIE TOOTHPASTE 225G"

Year	Sales Vol.
2008	569 585 552 362 439 965 1330 1387 830 369 976 1505
2009	2419 240 472 567 579 1183 1512 1669 809 583 359 322
2010	1177 2352 927 259 242 353 262 274 304 127 162 235
2011	382 599 356 332 270 264 861 1432 1361 658 874 278
2012	357 1155 1185 781 346 297 451 355 394 383 235 198
2013	587 605 412

Having cut off the last three data as training set and substituted the three prediction methods, the most optimized model as well as the model error of each method provided by the system is listed as below:

TABLE II. MODELS AND MODEL FITTING OF TESTING 1

Model No.	Model Description	R^2	RMSE
M_1	Winter's model	0.67	614.6
M_2	Seasonal Decomposition +ARIMA(2,2,1)	0.55	445.98
M_3	ARIMA$(1,2,1)(1,1,2)^{12}$	0.6	784.18

Generally, if the forecast model is desirable, then the forecast value must be referential, and the model's goodness-of-fit R^2 should be no less than 0.5. Therefore, suggest the threshold value R^2 is set to be 0.5, then after the first round of selection, M_1, M_2 and M_3 will all be retained. Among the three, M_2 has the smallest RMSE, thus M_1 and M_3 will carry out statistical significance testing against M_2 respectively. Through the testing between M_1 and M_2, we can find out the statistical volume $t=0.867<z=sig/2=0.025$'s tabular value 2.005, which shows there is no difference of statistical significance between M_1 and M_2, that the difference between them is stochastic. So M_1 is retained. Through the testing between M_3 and M_2, we can find out the statistical volume $t=0.25>z=sig/2=0.025$'s tabular value 2.005, which shows there is difference of statistical significance between M_3 and M_2, that the difference between them is stochastic. So M_3 is sifted out.

The calculation of error based on the testing set is given as below:

TABLE III. ERROR ESTIMATION OF TESTING 1

Observe Value	Forecast Value of M_1	Forecast Value of M_2
587	614	642
605	620	697
412	332	363

The Output Error is :

RMSE$=(RMSE(M_1)+RMSE(M_2))/2=(49.5+55.1)/2=52.3$

The forecast value is given as below:

TABLE IV. FORECAST VALUES OF TESTING 1

Time	Model M_1's output	Model M_2's output	Final output
2013-04	240	208	224
2013-05	98	158	128
2013-06	347	109	253

B. Testing 2

The existing data of monthly sales of "Tsingtao Beer with 11% alcoholic strength thick can 330ml"after promotion reduction at some outlet is listed as below:

TABLE V. MONTHLY SALES OF "TSINGTAO BEER WITH 11% ALCOHOLIC STRENGTH THICK CAN 330ML"

Year	Sales Vol.
2008	931 856 923 1349 2165 2542 2470 1607 1887 1333 1181 833
2009	1170 1052 1083 1426 1482 1127 4519 2576 4595 2828 1882 1788
2010	2647 2052 1600 2649 2705 1643 4080 2536 2518 731 666 669
2011	717 2234 531 671 829 810 1076 723 1093 1444 688 605
2012	1536 654 1048 1743 819 1379 768 969 546 493 280 212
2013	306 212 314

Having cut off the last three data as training set and substituted the three prediction methods, the most optimized model as well as the model error of each method provided by the system is listed as below:

TABLE VI. MODELS AND MODEL FITTING OF TESTING 2

Model No.	Model Description	R^2	RMSE
M_1	Winter's model	0.544	751.88
M_2	Seasonal Decomposition +ARIMA(1,2,0)	0.962	132.6
M_3	ARIMA$(1,2,1)(1,1,2)^{12}$	0.412	943.48

Likewise, suggest the threshold value R^2 is set to be 0.5, then after the first round of selection, M_3 will be sifted out. Among the remaining models, M_1 and M_2, M_2 has the smaller RMSE, thus M_1 will carry out statistical significance testing against *M2*, and through the testing we can find out the statistical volume $t=6.91>z=sig/2=0.025$'s tabular value 2.005, which shows there is difference of statistical significance between M_1 and M_2, that the difference between them is not stochastic. So there is only M_2 being retained in testing 2.

The calculation of error based on the testing set is given as below:

TABLE VII. ERROR ESTIMATION OF TESTING 2

Observe Value	Forecast Value
306	324.56
212	233.97
314	182.32

The Output Error is :
RMSE=77.8
The forecast value is given as below:

TABLE VIII. FORECAST VALUES OF TESTING 2

Time	Model M_1's output
2013-04	233
2013-05	184
2013-06	138

C. Conclusion of Testing

The errors of both testing with examples are kept within a small scope. And from the result of analysis we can also see that, different testing object fits for different forecast method. And in fact, a lot of testing also suggest that, with regard to different forecast objects, it is not applicable to use a fixed forecast method. Therefore, it is more intelligent and having more expansibility for the system to set up a library of forecast methods, and to choose different forecast method according to different forecast situation.

V. CONCLUSION

Developing supermarket commodity sales volume forecast system with a library of forecast methods being set up, selecting the model or model group with the best goodness of fit, and the approach of bagging is adopted to further increase the accuracy of the forecast model. Through testing, the system can forecast the commodity sales volume in nearest 2-3 phases with relatively good result. The key works for next step is trying to add other methods to the database of testing methods, to measure the data mining period in terms of time, and to test the validity of mining algorithm as the category of testing objects further increases.

ACKNOWLEDGMENT

This work was supported by the department of education projects of Fujian(JA10240).

REFERENCES

[1] Jaya, M., and K. Sundar. "Forecasting of market capitalization through arima (with special reference to Indian information technology firms)," Asian Journal of Research in Business Economics and Management, Vol. 2, No.10, 2012, PP.1-21.

[2] Luo, B., W-W. Yan, and L. Wan. "Application of Time-series Decomposition with Dummy Variables to Cigarette Sales Forecast." Jisuanji Xitong Yingyong- Computer Systems and Applications,Vol. 21,No.12,2012.

[3] Pan, Youqin, Terrance Pohlen, and Saverio Manago. "Hybrid Neural Network Model in Forecasting Aggregate US Retail Sales." Advances in Business and Management Forecasting,No. 9,2013,pp. 153-170.

[4] Jiang, Guorui, and Ying Liu."Research on Collaborative Forecasting Model Based on CPFR." Software Engineering and Knowledge Engineering: Theory and Practice. Springer Berlin Heidelberg, 2012, pp. 523-529.

[5] Menezes L M d, Bunn D W. "Review of guidelines for the use of combined forecasts."European J of Operational Research,Vol. 120,No. 1 ,2000.

[6] GuangYu Zhu, HongSen Yan, "A combination forecasting method based on model evaluation and selection from forecasting-model-base."Control and Decision, Vol. 19, No. 7, 2004.

[7] Cheng Xu, Hongsen Yan, Qing Huang, "Design and Development of Sales Forecasting System Based on Module Technique on .NET

Paltform.”Chinese Journal of Computer Technology and Development, Vol 17, No. 2, Feb. 2007, pp. 27-30.

[8] Hillebrand, Eric, and Marcelo C. Medeiros. “The benefits of bagging for forecast models of realized volatility.” Econometric Reviews Vol. 29,No. 5,2010,PP. 571-593.