Use AI Technology to Analyse Corporate Goods Price Index

Qi Fan
School of Computer Science and Technology, Huaibei Normal University
Huaibei City 235000, Anhui Province, P.R. China
Email: fanqimai@hotmail.com

Chang-jie Zhu, Quan-gui Chen, Jian-yu Xiao, Bao-hua Wang, Yan-yan Guo
School of Computer Science and Technology, Huaibei Normal University
Huaibei City 235000, Anhui Province, P.R. China

***Abstract*—In this work we took advantage of AI technique (data mining) to analyze the dataset of Corporate Goods Price Index (simple for CGPI). The public-use dataset from Chinese Bank is used in this research. We carry out a thorough research, and compare twenty-four kinds of different models and present the optimal results.**

***Index Terms*—Corporate Goods Price Index, AI, Data Mining, Weka**

I. INTRODUCTION

Corporate Goods Price Index (CGPI) is a investigation and statistics system in China, which is approved by Chinese statistical bureau, and implemented by Chinese Bank. The reprehensive products in CGPI includes 791 kinds of products, which stand for overall material products which cover agricultural products, minerals, oil and coal products, and processing products. It also contains primary product, intermediate product, and final products, even consumer goods and investment goods.

Data mining is commonly considered as the process of extracting the useful knowledge and rules from large collections of data [1]. Data mining techniques have been proven to be successful application in many areas, incluing Marketing, Economy, Medicine, Science and so on [2]. In this work we took advantage of AI technique (data mining) to analyze the dataset of CGPI.

II. RELATED WORK

After the literature survey, we found that there have been only a few studies on the CGPI analysis using statistical approaches [3]. Chen used multivariant statistical approaches to do the linear regression analysis to CGPI. The study of Chen [3] shows that multivariant statistical approaches is an possible method to study the relationship among the statistical parameters of CGPI, and estimate the current price to monitor the macrography economics. However, unfortunately Chen only investigated one approach, and this research should be still preceded deeply. More data mining algorithm could be investigated and be compared in this research.

III. DATASET AND METHOD

In this study, we adopted the dataset of CGPI from 2000 to 2009 in P.R. China. The data source is from the website of Chinese Bank, which includes 5 attributes and 117 records. The summary of the dataset is shown as Appendix A.

We have used Weka to experiment with these algorithms. Weka is a open source data mining software, which contains the tools for data pre-processing, classification, regression, clustering, association rules, and visualization [4].

Five performance measurements are used in this research: correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error.

IV. EXPERIMENT RESULTS

The overall results are shown in Appendix B, and we picked up the optimal results shown in Appendix C. After comparing, we could find that M5Rules algorithm has the best performance.

A. The data model of function

Functions.LeastMedSq

$$\begin{aligned} &OverallInd\_ex = \\ &0.2226 * Aripro\_ + 0.017 * Minepro\_ + \\ &0.078 * Oil\_\Pr o + 0.6882 * \Pr oce\_\Pr o + \\ &(-0.5797) \end{aligned} \quad (1)$$

Functions.LinearRegression

$$\begin{aligned} &OverallInd\_ex = \\ &0.226 * Ari\_\Pr o + 0.017 * Mine\_\Pr o + \\ &0.0782 * Oil\_\Pr o + 0.6816 * \Pr oce\_\Pr o + \\ &(-0.3746) \end{aligned} \quad (2)$$

Functions.PaceRegression

$$\begin{aligned} &OverallInd\_ex = \\ &(-0.3746) + 0.226 * Aro\_\Pr o + \\ &0.0179 * Mine\_\Pr o + 0.0782 * Oilpro\_ \\ &+ 0.6816 * \Pr oce\_\Pr o \end{aligned} \quad (3)$$

Functions.SMOreg

doi:10.4304/jsw.6.11.2143-2145

$$\begin{aligned} &OverallInd\quad ex = \\ &+0.3173 * (normalized\quad) Ari\ \Pr o + \\ &0.0394 * (normalized\quad) Mine\ \Pr o + \\ &0.1823 * (normalized\quad) Oil\ \Pr o + \\ &0.6061 * (normalized\quad) \Pr oce\ \Pr o \\ &-0.0476 \end{aligned} \tag{4}$$

B. The data model of rules

Rules.M5Rules

Rule: 1

$$\begin{aligned} &If \\ &\Pr oce \Pr o > 101.285 \\ &Then \\ &OverallIndex = 0.225 * Ari \Pr o \\ &+0.0168 * Mine \Pr o + 0.0789 * Oil \Pr o \\ &+0.6861 * \Pr oce \Pr o - 0.7029[47/0.703\%] \end{aligned} \tag{5}$$

Rule: 2

$$\begin{aligned} &If \\ &\Pr oce \Pr o > 98.175 \\ &Then \\ &OverallIndex = 0.235 * Ari \Pr o \\ &+0.0178 * Mine \Pr o + 0.0757 * Oil \Pr o \\ &+0.6805 * \Pr oce \Pr o - 0.8761[43/0.87\%] \end{aligned} \tag{6}$$

Rule: 3

$$\begin{aligned} &OverallIndex = 0.2218 * Ari \Pr o \\ &+0.0205 * Mine \Pr o + 0.0809 * Oil \Pr o \\ &+0.6764 * \Pr oce \Pr o + 0.0201[27/1.234\%] \end{aligned} \tag{7}$$

C. The data model of trees

Trees.M5P

$$\begin{aligned} &\Pr oce \Pr o <= 101.285: \\ &\quad | \ ProcePro <= 98.175: LM1(27/0.591\%) \\ &\quad | \ ProcePro > 98.175: LM2(43/0.519\%) \\ &ProcePro > 101.285: LM3(47/0.703\%) \\ &LM\ num: 1 \\ &OverallIndex = 0.2252 * AriPro \\ &+0.0198 * MinePro + 0.079 * OilPro \\ &+0.68 * ProcePro - 0.4102 \\ &LM\ num: 2 \\ &OverallIndex = 0.2334 * AriPro \\ &+0.0178 * MinePro + 0.0761 * OilPro \\ &+0.6807 * ProcePro \\ &-0.7876 \\ &LM\ num: 3 \\ &OverallIndex = 0.225 * AriPro \\ &+0.0168 * MinePro + 0.0789 * OilPro \\ &+0.6861 * ProcePro - 0.7029 \\ &Number\ of\ Rules: 3 \end{aligned} \tag{8}$$

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Do not use abbreviations in the title unless they are unavoidable.

D. ANALYSIS

We use M5Rules model to analyze the parameters of CGPI.

$$\begin{aligned} &OverallIndex = 0.225 * Ari \Pr o + \\ &0.0168 * Mine \Pr o + 0.0789 * Oil \Pr o + \\ &0.6861 * \Pr oce \Pr o - 0.7029 \end{aligned} \tag{9}$$

It is clearly shown that the parameter of processing product have most influence on overall index and play determinative role. The parameter of processing product increases one point, then overall index will increase 0.6861 point. Secondly, the parameter of agricultural product play certain influence. The parameter of agricultural product increases one point, then overall index will increase 0. 225 point. However, the parameters of oil product and mine product have little influence on overall index, and could be ignored in some certain situation.

V. CONCLUSION

This study clearly shows that AI technique (data mining) is a good method to study parameters relation of CGPI. We could take good advantage of data mining algorithm result to track CGPI and monitor macroeconomy. In future, we could also investigate time series algorithm to predict CGPI.

ACKNOWLEDGMENT

This study was supported by grant No. 080240 from Huaibei City Science Foundation, and by grant No. KJ2009A090, No. KJ2010A298 and No. KJ2010B186 from Education Council of An'hui Province.

REFERENCES

[1] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med. 26, 1, 24 (2002).
[2] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med. 22, 215, 231 (2001).
[3] Chen Yun, Mei Suo. Use Multivariance Linear Regression to analyze CGPI. Economist.5,204,205(2008).
[4] http://www.cs.waikato.ac.nz/ml/weka/.

Appendix A Summary Of Dataset

Attribute	Type	Minimal	Maximal	Average	Standard deviation	Skewness	Valid records
OverallIndex	Continous	92.000	110.300	101.938	4.559	-0.017	117
AriPro	Continous	92.600	118.350	103.082	6.885	0.497	117
MinePro	Continous	84.200	123.500	106.312	8.834	-0.545	117
OilPro	Continous	83.900	126.800	107.598	9.982	-0.426	117
ProcePro	Continous	91.700	108.000	100.809	3.803	-0.091	117

Appendix B All Experient Results

		Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
functions	GaussianProcesses	0.9917	0.4769	0.6631	12.6123%	14.4715%
	IsotonicRegression	0.99	0.522	0.6399	13.8051%	13.9642%
	LeastMedSq	1	0.0329	0.0453	0.8695%	0.9890%
	LinearRegression	1	0.0336	0.0434	0.8880%	0.9472%
	MultilayerPerceptron	0.9996	0.1003	0.1312	2.6514%	2.8639%
	PaceRegression	1	0.0336	0.0434	0.8880%	0.9472%
	RBFNetwork	0.8429	1.9678	2.4429	52.0363%	53.3118%
	SimpleLinearRegresion	0.9883	0.5689	0.6922	15.0438%	15.1063%
	SMOreg	1	0.0337	0.0436	0.8919%	0.9506%
lazy	IBk	0.9935	0.4197	0.5372	11.0997%	11.7231%
	KStar	0.9941	0.3601	0.5077	9.5228%	11.0804%
	LWL	0.9008	1.5976	1.9828	42.2461%	43.2693%
meta	AdditiveRegression	0.9608	0.9612	1.2695	25.4180%	27.7030%
	Bagging	0.9917	0.4914	0.5874	12.9942%	12.8183%
	CVParameterSelection	-0.2803	3.7816	4.5824	100.0000%	100.0000%
	EnsembleSelection	0.9886	0.5603	0.6849	14.8166%	14.9462%
	MultiScheme	-0.2803	3.7816	4.5824	100.0000%	100.0000%
rules	ConjunctiveRule	-0.0459	4.7514	5.8976	125.6450%	128.7029%
	DecisionTable	0.9763	0.7024	0.9856	18.5746%	21.5087%
	M5Rules	1	0.0283	0.0362	0.7491%	0.7898%
	ZeroR	-0.2803	3.7816	4.5824	100.0000%	100.0000%
trees	DecisionStump	0.8173	2.1486	2.6173	56.8167%	57.1159%
	M5P	1	0.0286	0.0367	0.7550%	0.8011%
	REPTree	0.9878	0.58	0.7078	15.3370%	15.4463%

Appendix C Optimal Experiment Results

		Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
functions	LeastMedSq	1	0.0329	0.0453	0.8695%	0.9890%
	LinearRegression	1	0.0336	0.0434	0.8880%	0.9472%
	PaceRegression	1	0.0336	0.0434	0.8880%	0.9472%
	SMOreg	1	0.0337	0.0436	0.8919%	0.9506%
rules	M5Rules	1	0.0283	0.0362	0.7491%	0.7898%
trees	M5P	1	0.0286	0.0367	0.7550%	0.8011%