The Use of an Association Rules Matrix for Economic Modelling

Raouf Veliev*, Alex Rubinov* and Andrew Stranieri**

*School of Information Technology and Mathematical Sciences, University of Ballarat, Victoria, Australia

**Donald Berman Laboratory for Information Technology and Law, Department of Computer Science and Computer Engineering, La Trobe University, Bundoora, Victoria, Australia

E-mail: {rveliev@krause.ballarat.edu.au, amr@ballarat.edu.au, stranier@latcs1.cs.latrobe.edu.au}

Abstract

Knowledge discovery techniques have not been widely used with macroeconomic data despite the social and political importance inherent in accurate economic forecasting. Rather, economic forecasting is currently performed with the use of models that rely heavily on theoretical assumptions. Economic assumptions are invariably contentious and model predictions are often rejected or accepted based on subjective perceptions about assumptions. We present an application of KDD that generates a forecasting model that avoids economic assumptions by focussing entirely on existing data. Although, association rules are typically used for finding interesting patterns in data, this is not a strategy we employed. Our approach differs in that all possible association rules between variables representing the current state of an economy in a quarter and the state in the next quarter are generated to form a matrix. A metric based on the support, confidence and expected probability of each rule is then derived. The system has been used to perform economic analyses of existing and future government policies. The system has been developed using data from the US economy over the last thirty years.

1. Introduction

A national economy is such a complex object of study that direct analysis is not possible. Therefore, economists have limited research to the study of models of economies. However the development of a macroeconomic model is heavily dependent on a researcher's own subjective view of the world and theoretical background and beliefs. For example, hypotheses generated by researchers who accept Keynesian assumptions are quite different from hypotheses from Classical theorists.

Hypotheses are not only dependent upon the subjective beliefs of their creators but can easily become obsolete. Completely different economic systems can emerge in different times in different countries and be described by different models.

Thus, if making assumptions and deriving hypotheses about an economy leads to subjective models, and successful theories do not last long, then the following questions arise: Is it possible to eliminate model dependence on the subjective researcher's assumptions about features and properties of the object of study?; Can there exist an approach that automatically generates a hypothetical basis for constructing a model?; Can this approach be applied in different times to different types of economic systems?

In this paper we suggest an approach that uses domain data only. Our approach is based on the technique of deriving association rules. Association rules (AR) were introduced by [2] and still attract the attention of many computer science researchers [7], [10] but to date no attempt has been made to apply AR to macroeconomic data.

Association rules are typically used to find interesting rules between attributes in a large database. Here, we use AR to generate a data structure we call an *Association Matrix (AM)*. The Association matrix is a simple model that is automatically generated from macroeconomic data without the need for any background assumptions.

We have trialed the Association Matrix using macroeconomic data from the US economy. Results indicate the approach is suitable for economic policy analysis.

0-7803-5871-6/99/$10.00 © 1999 IEEE

The Association Matrix is described in the next section. Following that we describe the application of our approach with US data and then offer concluding remarks and future research.

2. Association Matrix

Our approach is based on only one general assumption that the current state of an economy determines its state in the next period of time. This assumption seems reasonable if we take into account two arguments:

- A macroeconomic system typically changes slowly. If the intervals of time when observations are made is small enough (quarter or month), then it is unlikely that huge changes in the state of an economy will be noticed. Therefore, we expect that macroeconomic variables will predict values in the next period of time.

- Each state of the economy represents the result of numerous activities of many different economic agents participating in economic life. Most agents form expectations about the future that influence their behaviour. This assumption is broadly accepted in economic modelling. These types of expectations are called "adaptive expectations" [3], [4] [8].

The problem of forecasting the value of variables in the next time period converges to determining a mapping:

$$F: Y(t) \text{ -> } Y(t+1) \qquad (1)$$

where *"Y(t)"* is a set of macroeconomic variables in period of time *"t"*. It is necessary to find the contribution of each variable in this mapping. The question arises: Why not use statistical techniques to estimate this mapping such as those employed in [5]? We might have built a system of equations:

$$Y(t) = f(Y(t+1)) \qquad (2)$$

Then, we could have found coefficients of these equations, which would have represented weights of a system's impact on each macroeconomic variable. However, to estimate the system of equations we must assume the functional form of this dependence because the exact form is unknown. The adoption of one functional form over another is a very strong assumption about the nature of the dependence and is best avoided [6]. Furthermore, low quality data and the presence of noise can also significantly distort the picture of dependence. This especially applies to macroeconomic data, which reflects a high level of data aggregation.

To avoid with these difficulties we formalise the mapping (1) into a set of Association rules:

IF *variable(i)* equals *value(k)*,

THEN *variable(j)*" equals *value(l)* (3)

The degree of certainty of this rule will be defined by confirmed cases in the past using the metrics support, confidence and expected predictability.

The interestingness of mined rules is typically defined by a minimum support and confidence of mined rules. However, we do not set a minimum value for support and confidence in generating association rules, but instead generate all rules possible. These rules are used to construct an association matrix.

The rows of AM represent the "IF" parts of association rules. The columns reflect the "THEN" parts. Each variable of the system and each value from a set of values for the variable define a cell. Hence, the intersection of row (*variable(i), value(k)*) and column (*variable(j), value(l)*) reflects the rule (3). This cell stores a coefficient representing the degree of certainty of the rule (3). The coefficients for each rule, consequently for each cell in the matrix are obtained in the same way.

The degree of certainty of a rule can be calculated in various ways. For our purposes a coefficient that represents a degree of certainty should satisfy the following conditions:

1) It must be proportional to the support of a rule, because the support value determines the

probability of appearance of this dependence in the data;

2) It must be proportional to the confidence of a rule, because the confidence value determines the trueness of a rule;

3) It must be inversely proportional to the expected predictability of a rule. The expected predictability is the frequency of occurrence of the "THEN" items of the rules. The higher the expected predictability, the less significance the rule has, the more likely the "THEN" item happens anyway and less likely it depends on the "IF" item.

The association matrix represents the hypotheses for our future model and can be automatically generated for any macroeconomic system in any country. It does not depend on the modeller's personal assumptions about the object's nature or its behaviour and it can be fairly easily calculated. In the next section a description of a simple model, which we can build on the basis of this matrix, is given.

3. Example of a model

In this section we develop a simple macroeconomic model, which is based on the association matrix described above. The suggested model can be written in the following form:

$$X(t+1)=AX(t) \qquad (4)$$

where "A" is the association matrix. "$X(t)$" represents a vector of values, which are degrees of certainty for each variable of the system to hold each value from the set of possible values for this variable. These values represents the trueness of the "IF" part of the association rules as illustrated in Figure 1. For example, we see from this Figure that the degree of certainty that variable 1 will have value 1 at time T+1 if it had value 1 at time t is 35.5%

Thus, knowing the degree of certainty for each value in the current period of time for each variable in the system we multiply them by coefficients of association matrix. Then, we summarise them by columns to obtain the accumulated degree of certainty of each value in the next period of time for each variable representing the "THEN" parts of AR in columns. These values of degree of certainty form a vector "$X(t+1)$" and we start the process again. Provided the period of time for measuring the variables is relatively small, say, a month or a quarter for macroeconomic indicators, we can assume the linearity of the model reflected in the matrix transition.

	T+1						
T	V1	V1				...	Vn
			Value1	...	*ValueM*		
		Value1	35.5%		11.4%		
		...					
		ValueM	67.7%		27.3%		
	...						
	Vn						

Figure 1. Association matrix

The described model is able to generate the next state of the economy on the basis of its current state. Thus, having determined the initial point we can recursively evaluate the next point and construct a chain-trajectory of economic system development.

The forecasting model consists of the following blocks: 1) *The Input*, which represents the vector of the degree of certainty that variables will hold particular values in the current period of time; 2) *The association matrix*, which represents the parameters of the model based on the derived association rules between current and the next states; 3) *The Output*, which represents a vector of the degree of certainty that variables will hold particular values in the next period of time.

Variables that reflect a government's economic policies cannot be predicted. It is not appropriate to try to predict them, because they can be changed at any time by government decree. The set of these variables is represented in the *Control Block* of the system and can be manipulated by the user for scenario testing. This block allows the user to test different economic policy decisions and perform "what if" scenarios. According to different values of the control variables of the *Control Block* different economic development trajectories can be generated and the consequences of these decisions can be compared and analysed. The next section describes an implementation of this model to the economic system of the USA.

4. Implementation of the model

In this section we described the data used for the implementation of the model, how the association rules have been obtained and the structure of the forecasting model applied to the economic system of the USA.

4.1. Data

The data chosen for our experiment was obtained from the Federal Reserve Economic Data, which is supported by The Federal Reserve Bank of Saint Louise. The total number of variables is 412. The records are made on an annually, quarterly or monthly basis depending on a particular variable. The records report economic indicators between 1900 and 1998. However, the values for some economic indicators were not collected for the entire period. The goal of data pre-processing was to prepare a data sample with variables recorded for the same period of time and with the same frequency. The largest possible sample we could obtain contained 100 variables recorded from 1960 until 1997, quarterly.

The most common form of trajectory for the development of most macroeconomic variables in the short-run is similar to a monotonically increasing or decreasing function and can be fairly well approximated by a linear function. In order to capture more complex underlying dependencies in an economy, we substituted values of the variables for the change from one quarter to the next. For each variable we found the minimum and maximum value of change in percentage and divided the intervals into four sub-intervals.

The final stage of data transformation was to classify the value of change of each variable into the appropriate sub-interval category. This allowed us to reduce the set of possible values of the variables to the set of {one, two, three, four} which was done in order to map a 100-dimensional Quantitative Rules problem to a Boolean Rules problem. Any algorithm for finding Boolean Association Rules could then be applied [10]. In our application 80 variables were chosen to represent the Input block, while the remaining 20 were allocated for the Control block of the system.

4.2. Deriving association rules

To derive association rules we used the "Mineset" software package, on a UNIX platform. The standard Apriori algorithm [1] has been used to generate rules. We prepared two tables: the second table was the same as the first one, except that the first record was missing. It allowed us to generate rules, which predict variables from the next ($t+1$) state of the system on the basis of the current (t) state. In order to reflect all possible dependencies in the data, the threshold for support and confidence has been set to the minimum value - 1.00 %. It allowed us to obtain almost all possible rules. The total number of rules exceeded 57,000. The significance of each rule has been presented by three indicators: confidence, support and expected predictability.

4.3. Forecasting Model

Once the association rules were obtained, the support, confidence and expected predictability of rules need to be combined in order to provide a value for each association matrix cell. The three metrics are combined in this study by simply taking their average. Thus, the association matrix cell value represents the association between a value on a a variable and a value on another variabel and is calculated from the association rule, A as

(support(A) + confidence(A) - expected predicability(A)) / 3.

Only those rules, which have confidence or support value less than 1%, have "null" value in the corresponding cell. For example, the variable v51 has the value of "one" only once in the data set, therefore at least the support value of all rules, which have v51 in their "IF" or "THEN" parts will be less than 1%. Thus, they do not satisfy the threshold value and "Mineset" will not show those rules.

The output values of the system are evaluated by multiplying the input vector by the association matrix. Hence, for each value of each variable in the next quarter we obtain the value, which represents the accumulated influence of the entire system on this variable.

Although this number is not the exact probability of this value, it represents a degree of certainty that this variable will have this value in the next period of time.

The degree of certainty varies from zero to 400*200/3 (when the support and confidence equals 100% and expected predictability equals 0% for each rule for this variable). Zero indicates that it is impossible for the variable to hold the value. The value "400*200/3" indicate that this variable will definitely have this value. We normalise this numbers to the interval {0,1} to keep a balance between input and output. This vector is transferred to the spreadsheet, which represents the next quarter forecasting system.

Before processing the next quarter, we pick up the value of each variable with the greatest value of the degree of certainty. These values generate the current point in the trajectory of development of the economic system. Then, the process starts from the beginning.

5. Results

The association matrix derived model was tested with data from the first two quarters of 1998 published by The Federal Reserve Bank of Saint Louise A comparison of the system's output with the real data showed that the changes in 55 out of 80 indicators were correctly predicted. The remaining 25 indicators, though not correctly predicted were still close to the real data. Most of these variables were from the financial group and price indexes. Many changes have recently occurred within sector due to the influence of changing global conditions. Therefore possible inaccuracies in these predictions resulted because we didn't include the world's economic description in our data set. Future research aims to test our approach with an expanded data set that capture global indicators in addition to national indicators.

Simulations of the system under different control regimes have been performed. Results were largely consistent with some existing macroeconomic models. For example, the impact of changes in taxation levels on gross domestic product and employment is consistent with the main propositions of Keynesian General Theory. The impact of interest rates on consumption and gross domestic product was consistent with Classical Theory. On the other hand, some new dependencies in the US economic system have been discovered, which are not covered by existing theories. For example, an increase in real national defense investment leads to a decrease in employment in transportation and a decrease in non-financial corporate business profit after tax. An increase in employment in the government sector causes an increase in total automobile credit outstanding.

6. Conclusions

In this paper we proposed an approach to enhance the economic modelling process with the use of association rules. A simple macroeconomic model, on the basis of this approach, has been built and implemented for the economic system of the USA. The simulation of the model showed that it could be used in economic policy analysis to test different government economic decisions and analyse different scenarios of economic development. It also allows us to explore which indicators are stable with regards to the Control block and can hardly be influenced and monitoring by government as well as those, which are especially sensitive to such a control.

7. References

[1]. Agrawal R., Srikant R., *Fast Algorithm for Mining Association Rules in Large Databases,* VLDB 1994: 487-499.

[2]. Agrawal R., Imielinski T., Swami A., *Mining Association Rules between Sets of Items in Large Databases,* SIGMOD Conference 1993: 207-216

[3]. Brayton F., Mauskopf E., Reifschneider P., Williams J., *The Role of Expectations in the FRB/US Macroeconomic Model,* Federal Reserve Bulletin, April 1997: 227-245.

[4]. Brayton F., Mauskopf E., *The federal reserve Board MPS Quarterly Econometric Model of the US Economy,* Economic Modelling, vol. 3, July, 170-292, 1985.

[5]. Fair C., *Testing Macroeconometric Models,* Cambridge, Mass., Harvard University Press, 1994.

[6]. Glymour C., Madigan D., Pregibon D., Smyth P., *Statistical Themes and Lessons for Data Mining,* Data Mining and Knowledge Discovery, vol.1 No.1, 1997.

[7]. Meo R., Psaila G., Ceri S., *A New SQL-like Operator for Mining Association Rules.* VLDB 1996: 122-133.

[8]. Murphy C., *An Overview of the Murphy Model,* Australian Economic papers, Supplement, 1988: 175-199.

[9]. Rubinov A., Nagiyev A., *Elements of Economic Theory,* Baku, Bilik, 1992 (in Russian)

[10]. Srikant R., Agrawal R., *Mining Quantitative Association Rules in Large Relational Tables,* ACM SIGMOD Conference on Management of Data, 1996.