GDP Forecasting through Data Mining of Seaport Export-Import Records

H Raymond Joseph†

***Abstract*—With the ever increasing ubiquitousness of globalization through international trade, principally on sea, there seems to be a direct correlation to a nation's Gross Domestic Product(GDP). Traditionally, in literature, structural models have predicted GDP correlation with the export-import tonnage on a cross-section of commodities. In this paper, machine learning and data mining techniques on publicly available, export and import tonnage of commodities at sea ports of the nation in question are analysed. Algorithms are then considered that output real GDP forecasts for the fiscal. The dataset for the exercise consists of daily export and import tonnage at a given port. Several ports in the country of interest are then considered. With data for several years and the accompanying GDP forecast on a daily basis, the question provides a challenging supervised learning problem to be analysed, with an appropriately sized data set, that is expected to generalize.**

***Index Terms*—GDP Forecasting, Seaport Data Analysis, Export-Import Analysis, Machine Learning and Macroeconomics.**

I. Introduction

For the purpose of definition, GDP is the total market value of all final goods and services produced in a country in a given year, equal to total consumption, investment and government spending, plus the value of exports, minus the value of imports.[1]

Correlations drawn between Export-Import volumes and real GDP have been widely researched in Economics and Econometrics literature. There also exist quantitative models that seek to model the correlational behavior between these two Macroeconomic variables. Meanwhile, the large data-set available on tonnage and volume of Exports and Imports at a nation's Seaports and the corresponding GDP forecasts make the problem appropriate to be considered within the purview of Machine Learning and Data Mining. For instance, Owokuse investigates "Causality Between Exports, Imports and Economic Growth".[2] Ben-David and Loewy[1998] argue that an increase in exports means:Increase in employment in export sector industries which, in turn, increase income and GDP, reallocating resources from less productive sectors to exports industry and enhancing capacity utilization exports growth promotes GDP growth.[3]

Traditionally, GDP forecasts have been produced and utilised by several agencies ranging from Investment Banking Corporations, Ratings Agencies and Governments.[4] Many such agencies make forecasts about the expected GDP and make appropriate changes pertaining to spending, capital utilisation and leverage. Spending on GDP analytics forms an important part of research spending in these organisations. Such forecasts have been made from time to time to reflect dynamic changes in the economy.

As mentioned, there is an observed correlation between GDP and Export-Import. The nature of this dependency is not very clear, and very few mathematical models exist, that explicitly relate these quantities. Hence, the problem is bought within the purview of Machine Learning and Data Mining.

II. Application Context for GDP Forecasting Using Machine Learning on Export and Import Data

A. Importance of GDP forecasting

- Economic forecasts of GDP are very important for determining monetary and fiscal policy.[5]
- If the GDP is really expected to increase, then inflation may pick up and the Banks may need to raise interest rates. If the GDP is likely to continue to shrink, the Banks may need to pursue further quantitative easing.
- Another issue for any national monetary authority is that interest rate changes can take up to 18 months to have an effect. Therefore when interest rates are changed, they are trying to set the optimal rates for the future economic situation.

B. Machine Learning on Export and Import Data

The machine learning algorithm employed is expected to give various factors weights. For instance the weights for categories in codes $39-40$, Plastics and Rubbers, (see table) may be very different from those for categories $72-83$, Metals. Hence the algorithm is expected to assign appropriate weights.

Another important aspect to be noted is that, for most countries, Exports and Imports in all the categories may not necessarily be non-zero. There may exist several goods and commodites that are not traded at all by the country. Our model is able to allow for this.

C. Correlation between Export-Import and GDP

Disagreements persist in the empirical literature regarding the causal direction of the effects of trade openness on economic growth and hence the GDP. Michaely (1977), Feder (1982), Marin (1992), Thornton (1996) found that countries exporting a large share of their output seem to grow faster than others.[6] The growth of exports has a stimulating influence across the economy as a whole in the form of technological spillovers and other externalities. Models by

†The Author wishes to thank the Shastri-Indo Canadian Institute and the DFAIT, Canada, for generous funding.

Grossman and Helpman (1991), Rivera-Batiz and Romer (1991), Romer (1990) posit that expanded international trade increases the number of specialized inputs, increasing growth rates as economies become open to international trade.[7] Buffie (1992) considers how export shocks can produce export-led growth.[8] Export growth is often considered to be a main determinant of the production and employment growth of an economy and its GDP. Similarly, Import growth is expected to have adverse effects on GDP. Export expansion and openness to foreign markets is viewed as a key determinant of economic GDP growth because of the positive externalities it provides. For example, firms in a thriving export sector can enjoy the following benefits: efficient resource allocation, greater capacity utilization, exploitation of economies of scale, and increased technological innovation stimulated by foreign market competition. [9]

III. The Problem Formulation - Learning Data and Prediction Output Structure

The problem data-set is an N dimensional data vector, where$N$ is the number of classes of commodities considered. Each dimension of the data vector class is a $2-$tuple numerical - volume exported and volume imported. Hence, we have several data vectors for several days under consideration. The problem data-set is expected to consider 25 or more years for the purpose of sampling. Therefore, the data-set is large enough for Machine Learning purposes (365×25 data vectors). The classification system for commodities is the widely used International Harmonic System. A brief table of the classification is outlined as shown below in Table 1.[10]

Table 1: International Harmonic System of Classification.[10]

Code	Commodity
01-05	Animal & Animal Products
06-15	Vegetable Products
16-24	Foodstuffs
25-27	Mineral Products
28-38	Chemicals & Allied Industries
39-40	Plastics / Rubbers
41-43	Raw Hides, Skins, Leather, & Furs
44-49	Wood & Wood Products
50-63	Textiles
64-67	Footwear / Headgear
68-71	Stone / Glass
72-83	Metals
84-85	Machinery / Electrical
86-89	Transportation
90-97	Miscellaneous
98-99	Service

The Table 1 is presented only for the purpose of completeness. Note that $N = 16$ for this case. For the purpose of GDP forcasting, the factors may need to be weighted, within a data vector. For example, commodities entailed within section 72-83 (Metals) may be given more weights, as assigned by the learning algorithm. Corresponding GDP forecasts made are available for use by the learning algorithm. Hence, the problem reduces to a supervised learning problem.

At this point it must be borne in mind that for predicting GDP on day $t+t'$the Export-Import trade has to be forecasted on day $t + t'$. This setback is underscored by the fact that International Export-Import trades are predictable - contract agreements are entered into well before the goods are actually delivered. However, this can also be viewed as a sub-problem of forecasting trade volumes at time $t+t'$, given trade volumes upto time t. This complication will not be considered since, by the earlier assumption, trade volumes of commodities are considered to be predictable. Also, the apparent diificulty in predicating Import-Export trades is underscored by a causality relationship between these factors.

The problem formulation in qualitative term, alongwith dependencies is represented below pictorially:

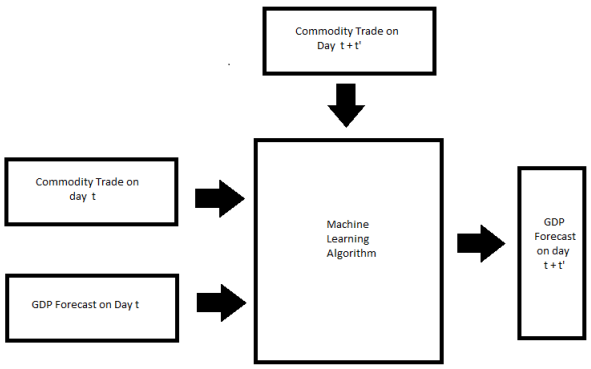

Figure 1: Pictorial Representation of the Problem Formulation.

IV. A closer Look at The Data - Structure and Frequency

Presented in this section is a data-set example. The country considered is India. In particular, this is the data for the Chennai Port, on 16^{th}March, 2013.

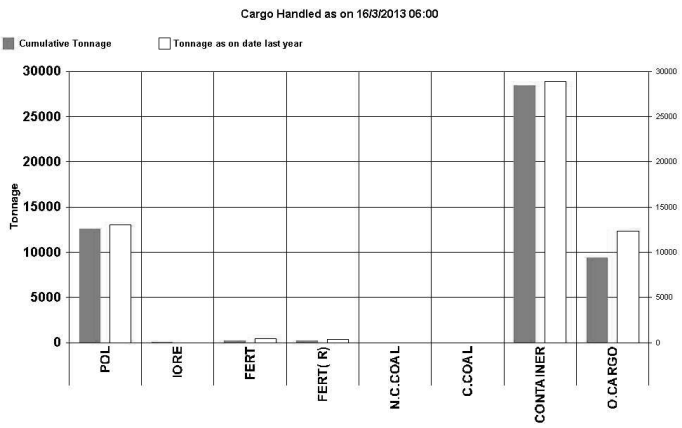

Figure 2: Data-Set Example.

V. Possible Algorithms and the Learning Process

Several algorithms can be considered for the purpose of this learning process. As such in this section their applicability is discussed.

A. *Support Vector Machines - Multi-Class SVM approach*

Support Vector Machine (SVM) ia very specific type of learning algorithms characterized by the capacity control of the decision function, the use of kernel functions and the sparsity of the solution. Established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error, SVM is shown to be very resistant to problems of over-fitting thus, achieving an high generalization principle. Also, SVM is equivalent to solving a linearly constrained quadratic programming problem, so that the solution of SVM is always unique and globally optimal, unlike neural networks training which requires nonlinear optimization with the danger of bumping into a local minima.

A pictorial representation of the problem formulation for using an SVM approach is given in Figure 3.

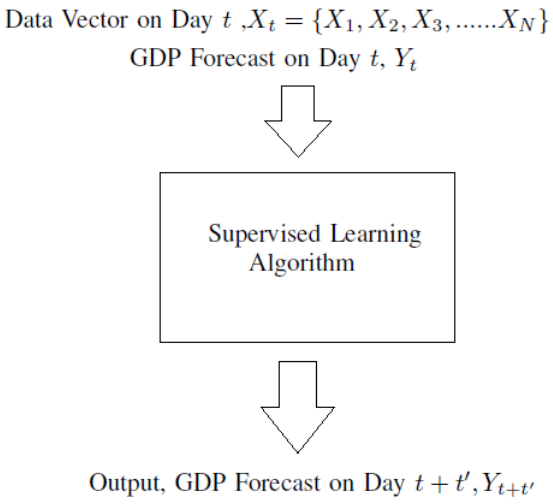

Figure 3: SVM Approach.

- The data vector is a vector of Nelements, where each element consists of Export and Import data on a given commodity.
- Hence, the problem becomes a N-dimensional supervised learning problem, with the supervising data aspect being the GDP on that day, corresponding to data within the data vector.
- The problem is reformulated as several binary classification problems.
- Each problem is such that it classifies all points over either a short-range of GDP or over the rest of the GDP range.
- The learning algorithm constructs a set of best-fit hyperplanes that separate these points.
- The GDP forecast varies over a short range, while the cargo tonnage has a larger variational range.
- This makes it simpler to have 'Pockets' wherein, Export-Import data vector falling within a certain accepted tolerance is mapped on to some GDP forecast.

Consider the problem of separating the set of training vector belonging with GDP forecasts,

$$G = \{(X_t; Y_t); t = 1, 2, ..., T\}$$

with the hyperplane

$$wF(X) + b = 0 \ ;$$

where, $X_t \in \mathbb{R}^{2N}$is the input vector on th t^{th}day, $y_i \in \{0, 8\}$is the known % GDP forecast.

Since the considered methodology is that of a Multi-Step SVM, notice that the classification SVM description is a binary classification problem. At the first classification, the classifier classifies the vectors of observed data into binary sets of (say) GDP growth forecast $>= 4\%$and GDP growth forecast $< 4\%$.

Then the classified data is once again input to the classifier but this time with constraints - $>= 2\%$,$<= 4\%$ and $< 2\%$ as one set and another set of $>= 4\%$,$<= 6\%$ and $>= 6\%, <= 8\%$ and so on. The classification continues until the forecast granularity reaches a desired stage such as 0.1%. [11]

B. *Fuzzy Set Based Genetic Learning Algorithms*

The genetic algorithms (GAs) are the procedure that searches the space of character strings of the specified length to find strings with relatively high fitness [12]. In preparing to apply the GAs to a particular problem, the first step involves determining the way to represent the problem in the chromosome-like language of GAs. An immediate question arises as to whether it is possible to represent many problems in a chromosome-like way. For this, we make use of fuzzy systems and the FAM matrix (Fuzzy Associative Memory).

Fuzzy systems are comprised of fuzzy sets, defined by their membership functions and fuzzy rules, which determine the action of the fuzzy system. The fuzzy rules can be concisely represented with one or more FAM matrices. The FAM matrix entries mainly depend on the subjective decision of an expert in each situation. GAs will then be used to adapt the FAM matrix entries so that the performance of a fuzzy system fits the desired behavior.

To apply genetic optimization to FAM matrix adaptation, we string the matrix entries together into a single long vector. The result is a very long binary vectors end to end. This is the chromosome upon which the GAs operate.

Here's a brief description of the algorithm:

- A single FAM matrix is used that deals with all the $16(N)$ classifications within the input data vector, as prescribed in the International Harmonic System.
- For the purpose of this method, each entry which comprised of Export and Import tonnage is made one entry which gives the signed difference between Import and Export.
- If we use r fuzzy sets for each input, i.e the possible GDP forecast interval $(0, 8)$ is divided into r partitions, then we will have r^{16}entries. Writing this as an appended vector from end to end, we will have $16 \times r^{16}$entries.
- The GA operates on the $16 \times r^{16}$entries.

From the preceding analysis, it is very clear that this number grows rapidly with r. However the analysis can be simplified by reducing N. The justification for reducing N lies in the arguments that:

- A certain country may not be either an Exporter or an Importer of all 16 categories of traded goods.

- Also, a number of commodities traded on the shores may be shown to economically have very little effect due to the volume in which it's traded in.

For the purpose of considering the workability, incorporating the above arguments, assume $N = 5$. Also, consider that$r = 10$. Then the number of entries for the GA to operate are, 5×10^5, which is still within reasonable limits for the Genetic algorithm to yield generalization. A pictorial representation of the Algorithm is as in Figure 4.

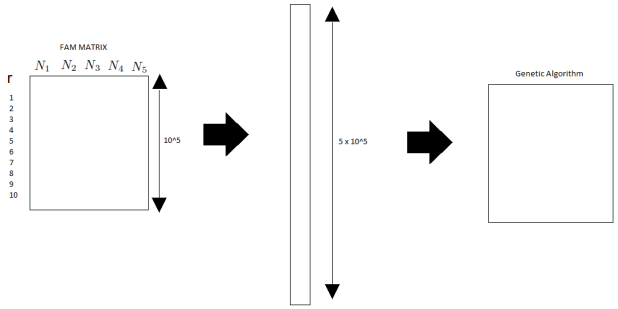

Figure 4: Using Genetic Algorithms.

C. Artificial Neural Networks

In recent times, much research has been carried out on the application of artificial intelligence techniques to the load forecasting problem.However, the models that have received the largest share of attention are undoubtedly the artificial neural networks (ANNs). The first reports on their application to the load forecasting problem were published in the late 1980's and early 1990's [13]. However, the models that have received the largest share of attention are undoubtedly the artificial neural networks (ANNs).

Artificial neural networks are mathematical tools originally inspired by the way the human brain processes information. Their basic unit is the artificial neuron. The neuron receives (numerical) information through a number of input nodes (four, in this example), processes it internally, and puts out a response. The processing is usually done in two stages: first, the input values are linearly combined, then the result is used as the argument of a nonlinear activation function. The combination uses the weights w_i attributed to each connection, and a constant bias term θ, with a fixed input equal to 1. The activation function must be a nondecreasing and differentiable function; the most common choices are either the identity function , or bounded sigmoid (s-shaped) functions, as the logistic one $y = \frac{1}{(1+e^{-x})}$. [14] Figure 5, shows the application of this learning model to the problem in hand.

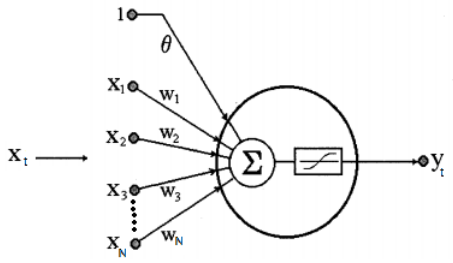

Figure 5: Artificial Neural Networks Approach.

VI. Conclusion

In this paper the problem of GDP forecasting was analysed with Seaport Export-Import data records as the learning resource. The usability of Learning Algorithms and problem formulation for these methods have also been discussed. The importance of GDP forecasting has been rightly emphasized in this paper, and hence it's utility. This paper can also be viewed as a quantitative indicator of the effect of Export-Import tonnage injected, on the economy of a nation.

The learning algorithms discussed are the Support Vector Machines, Fuzzy Set Based Genetic Learning Algorithms and also the Artificial Neural Network methods. Future research being carried out could focus on hybrid of these algorithms to yield more accurate forecasting.

References

[1] B Roffia, A Zaghini, Excess Money Growth and Inflation Dynamics, International Finance, 2007.

[2] T O Owokuse, Causality Between Exports, Imports and Economic Growth, Economics Letters, 2007.

[3] B David and Loewy, Free Trade, Growth and Convergence, Journal of Economic Growth, 1998.

[4] J Kitchen, R Monaco, Real-Time Forecasting in Practice, Business Economics, 2003.

[5] BS Bernanke, M Woodford, Inflation Forecasts and Monetary Policy, nber.org, 1997.

[6] J Thornton, Cointegration, Causality and Export-Led Growth in Mexico, Economics Letters, Elseiver, 1996.

[7] PM Romer, L A Rivera-Batiz, Economic Integration and Endogenous Growth, European Economic Review, 1991.

[8] E F Buffie, On the Condition for Export-Led Growth, Canadian Journal of Economics, 1992.

[9] E Helpman and P Krugman, Trade policy and market structure, 1985.

[10] International Harmonic Systems Classification.

[11] Wei Huang, et al, Forecasting Stock Market Movement Direction with Support Vector Machine, Computers and Operations Research, 2005.

[12] . Goldberg, D. E., Genetic Algorithm in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

[13] T. Czernichow, A. Piras, K. Imhof, P. Caire, Y. Jaccard, B. Dorizzi, and A. Germond, Short Term Electrical Load Forecasting with Artificial Neural Networks, Engineering Intelligent Syst., vol. 2, pp. 85–99, 1996.

[14] Henrique Steinherz Hippert, Carlos Eduardo Pedreira, and Reinaldo Castro Souza, Neural Networks for Short-Term Load Forecasting: A Review and Evaluation, IEEE Transactions on Power Systems, Vol. 16, February 2001.