Association Pattern Discovery of Import Export Items in Ethiopia

Mezgeb Manaye
Department of Information Technology, Nefas Silk Polytechnic College, Addis Ababa, Ethiopia
mezgebman@gmail.com

Berhanu Borena
PhD Candidate, Addis Ababa University, Ethiopia
berhanuborena@gmail.com

Abstract

This paper examines the application of data mining to detect association pattern of customs administration data with market price and currency exchange rate in Ethiopia. The association rule method of data mining is used in this paper to generate the interesting pattern from the data. The research focused on the objective of identifying relationships between attributes of custom data and market price to clearly understand the nature of import-export items in Ethiopia.

The results of the experiments carried out using association rules has discovered that the technique of data mining is applicable to generate knowledge from import and export items in custom administration. Algorithms such as Apriori, Tertius, PredictiveApriori and FliteredApriori were used to generate the associations. One of the resulted associations indicates there is strong link between market price and textiles imported.

The implication of this research finding is to clearly identify the association of import-export items with the market price and the effects of those items on the market price and currency rate in Ethiopia.

Keywords: Association Rule; Data Mining; Customs Administration, Market Price

1. Introduction

The incapability of human beings to interpret and digest accumulated huge data and make use of them for decision-making has created the need for development of new tools and techniques for automated and intelligent analysis. As a result, the discipline of knowledge discovery or data mining, which allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified, has evolved into an important and active area of research [18].

Data mining is an automated process employed to analyze patterns in data and extract information [27]. Central to data mining is the process of modeling data set. There are widely used generic modeling techniques. Some of these are Neural Networks, Agent Networks, Genetic Algorithm, Decision Trees, and hybrid models [2].

Data mining's extraction of meaning from huge databases is exactly what companies are looking for to increase profits through describing past trends and predicting future trends [6]. It is about extracting interesting patterns from raw data. There is some agreement in the literature on what qualifies as a "pattern" (association rules and correlations [20, 21, 22, 23] as well as clustering of the points [12], are some common classes of patterns sought), but only disjointed discussion of what "interesting" means. Most works on data mining study patterns to be extracted automatically, presumably for subsequent human evaluation to the extent in which they are interesting. Patterns are often deemed "interesting" on the basis of their confidence and support [20, 21, 22, 23], information content [13, 17], and unexpectedness [1].

Modern data mining is used intensively and extensively by financial institutions (for credit scoring and fraud detection) [16], retailers (for market segmentation and store layout), and manufacturers (for quality control and maintenance scheduling) [11].

Similarly various studies have addressed the different aspects of customs administration and market price by using data mining techniques and other statistical approaches. Numerous data mining-related studies have been undertaken to analyze customs data, with results frequently varying depending on the socio-economic conditions and infrastructure of a given location.

2. Related Work

Import-export activity of Ethiopia is much related with the country's economy and policy. As this research focuses more on aspects that influence price from custom perspective, we reviewed prior works that dealt on factors that affect price of commodities.

The effect of depreciation of the Birr on major export products of Ethiopia: the case of hides and skins was presented by Ali [4]. The paper analysed the effects of trade and exchange rate policies on one of Ethiopia's agricultural export items, hides and skins. The study answers what happened in the export of hides and skins when the exchange rate depreciates continuously for the last 17 years. The empirical findings of the study reveal that real exchange rate is one factor, among many others, that affects the volume of export of hides and skins. Hence, "it is recommended that policy towards liberalized exchange rate determination should be complemented by other policy measures which are in harmony with the economic agenda of export enhancement". The paper show the impact of Birr exchange fluctuation on export, but does not assess the vice versa and the association between Birr and import product.

The Impact of Trade Liberalization on the Ethiopia's Trade Balance was presented by Hailegiorgis [7].The model used for the analysis of impact of trade liberalization on trade balance was based on export equation of Santos-Paulino and Thilwall [24]. The study examines the impact of trade liberalization on the Ethiopia's trade balance using the data over the period 1974 to 2009 from NBE (National Bank of Ethiopia). The country has undertaken serious trade reforms, either as part of major macroeconomic reforms and commitments with international regulations, or by decisions driven by a process of internal adjustment for the last two decades. One of the anticipated gains from the trade liberalization policies adopted by Ethiopia is improved export performance. The researchers analyzed the impact of trade liberalization on Ethiopian trade balance. However, when it was examined with the application of export equation, it showed that trade liberalization led to the deterioration of the trade balance or too fast of an increase in imports. Thus, "it was deduced the evidence that the trade liberalization worsens trade balance due to more imports than exports". The paper showed that the trade liberalization worsens trade balance due to more imports than exports, but did not show the association between import/export items with the market price and currency exchange and which item could highly affect the market price and currency rate exchange.

Another paper examined Foreign Exchange Rationing, Wheat Markets and Food Security in Ethiopia [14]. The paper showed the relation between Ethiopia's domestic cereal markets and the international market. According to the paper, there was a remarkable growth in Ethiopia's agricultural production and overall real incomes (GDP/capita) from 2004/05 to 2008/09. Due to this growth, the prices of major cereals (teff, maize, wheat, and sorghum) have fluctuated sharply in both nominal and real terms. International prices of cereals also fluctuated widely, particularly between 2006 and 2008. However, the links between Ethiopia's domestic cereal markets and the international market are by no means straightforward. Among the major staples, only wheat is imported or exported on a significant scale. Frequent changes in trade and macro-economic policies, movements in international prices and fluctuations in domestic production have at times eliminated incentives for private sector imports of wheat. "Domestic wheat prices have been above wheat import parity prices since May 2008, indicating that it would be profitable for private traders to import wheat if they had access to foreign exchange at the official exchange rate".

A recent paper that examined container clearance/release times in gateway of African ports, is an example of analyzing data extracted from Customs IT systems [19]. This study provided information on why clearance/release times "are widely recognized as a critical hindrance to economic development". It also demonstrated the interrelationships between logistics performance of consignees, operational performance of port operators and efficiency of customs clearance operations. Moreover, the paper showed customs

data mining can enable to identify and describe clearance and forwarding agents, shippers and shipping line strategies and their impact on time release.

Clifton and Gengo [3] used data mining to develop custom intrusion detection filters. The researchers' approach was developing custom filters that reduce the false alarm stream based on known "normal behavior" in a particular environment. They used commercial intrusion detection systems, but filter out produced alarms that fit a pattern caused by normal operation at that site. The difficulty with this approach is building these filters, and determining what is normal operation at a site. While much less costly than building a complete intrusion detection system, it still requires considerable human effort. To reduce this effort, they used data mining technology to discover alarms caused by normal operation. They developed custom filters data mining model based on sequences of alarms using sequential association mining. The idea is that a sequence of operations that are normal in a particular environment may contain operations that look like a potential intrusion. However, the complete sequence is unlikely to be duplicated in an intrusion, so alarms that are part of the complete sequence can be ignored. The problem is in identifying such normal sequences. They use frequent episodes to identify frequently occurring sequences of alarms. An episode is a sequence of alarms that occurs within a specified time window.

Yan-hai and Lin-yan [10] used cluster data mining rule to study and apply data mining to structure risk analysis of customs declaration. The cluster method of data mining is used in the paper to divide the cargo into seven types so that customs can put the mainly inspection force to the high risk level cargo. The results showed that this kind of method can be used to reform the operation mode of customs inspection. Through the basic cluster analysis of cargo, they found that multi-variance statistical analysis is a very important method, and also more effective than the other tools, which can offer the basement of quantity analysis, and can support decision-making.

Another paper examined risk management systems using data mining in developing countries' customs administrations [9]. The researcher used data mining (descriptive statistics) to be successful in accurately targeting declarations that present a risk of infraction, to carry out prior work on data analysis, on descriptive statistics. This work requires customs to identify the characteristics of declarations that, in a preceding period, have resulted in an infraction, and then deducing the 'statistical regularities' in those infractions. The researcher used a database that covers twelve months. Data comes from detailed declarations and monthly statements of customs infractions for the two main offices in Dakar. Finally, the researcher combined these risk profiles to facilitate the right decision with regard to referring the declaration to a particular customs clearance channel. Generally the paper showed the risk management on the custom administration to improve the problem to facilitate the decision, but did not show the association of import/export items with the market price and currency exchange.

Finding association rules that trade support optimally against confidence was presented by Scheffer [25]. The methodology used was knowledge Discovery in Databases (KDD). KDD is the process of discovering useful knowledge from a collection of data. When evaluating association rules, rules that differ in both support and confidence have to be compared; a larger support has to be traded against a higher confidence. The solution proposed for this problem is to maximize the expected accuracy that the association rule will have for future data. The contributions of confidence and support to the expected accuracy on future data was determined. Then a fast algorithm that finds the n best rules which maximize the resulting criterion is presented. "The PredictiveApriori algorithm returns the n rules which maximize the expected predictive accuracy; the user only has to specify how many rules he or she wants to be presented".

Scalable parallel data mining for association rule was presented by Karypis and Kumar [5]. The algorithm of the Apriori class is based on the simple observation: if a given itemset is not frequent, then none of its supersets can be frequent. They have a level-wise behavior: they start with k=1 by evaluating singleton itemsets, and base the

computations performed at step k on the results of the previous iteration k-1. This level-wise behavior has been often criticized because of the consequent multiple scans of the dataset, one for each level. A lot of research has been thus devoted to minimize the number of the dataset scans.

Confirmation-Guided Discovery of First-Order Rules with Tertius was presented by Flach and Lachiche [15]. This paper deals with unsupervised discovery of rules in first-order logic. The researchers defined statistically well-founded confirmation measure to induce rules that are in some sense unusual or interesting. Then they described a complete best-first search algorithm that uses an optimistic estimate of the best confirmation of possible refinements of a rule to prune the search. The algorithm works on a function-free first order Prolog representation, which enables application to a wide range of structured domains, and to include background knowledge as part of the heuristic evaluation.

From the above reviewed papers, we can understand that there is research gap on the area of association mining of import and export items and the relationship between those items and the economy regarding market price and currency rate of different countries. Thus in this paper, association rule data mining technique is used to study the relationship between major export items (coffee, livestock, and oil seeds0 and import items (food, fuel, and textiles) and market price and currency rate.

Consequently we addressed the following research question in this paper:

- What sort of association exists between major import-export items, and market price or currency exchange rate in Ethiopia?

3. Method

This part of the paper discusses the data set, tools, and approaches used for attribute selection, dimensionality reduction, and model building.

3.1 General Approach

We followed CRISP-DM (Cross-Industry Standard Process for Data Mining) data mining methodology with appropriate modification to fit into the problem domain at hand. CRISP data mining process model describes commonly used approaches that expert data miners use to tackle problems. Polls [8] showed that it is the leading methodology used by data miners. Accordingly, based on situational analysis on this study, business and data understanding were the first tasks. Then follows data preprocessing tasks relevant to the data mining goal identified. Model building and evaluation along with a possible recommendation to integrate the resulted pattern or knowledge with the existing one is the last stage.

3.1 Data Collection

This study used data obtained from Ethiopian Revenue and Customs Authority (ERCA), Central Statistics Agency (CSA), and the National Bank of Ethiopia. The total dataset for the study contains import export records from 1997-2012, market price from 2000-2012, and currency rate from 1997-2012. Based on the availability of data, a total of more than 750,000 import export information was described with 10 attributes.

3.3 Tools

Weka data mining tool was used as a data analysis tool. Weka is a machine-learning system written in Java. It was adopted for undertaking the experiment in data mining, which includes several algorithms that can be used for feature/attribute selection and model building.

4. Experiment and Results

The first stage of the data mining process is to select the related data from available databases to correctly describe a given business task [9]. There are at least three issues to be considered in the data selection. The first issue is to set up a concise and clear description of the problem. The second issue would be to identify the relevant data for the problem description. The third issue is that selected variables for the relevant data should be independent of each other. Variable independence means that the variables do not contain overlapping information.

Using the right data for data mining task is one of the primary keys for successful data mining [26]. For this research, the 12 years initial data were collected from three different government organizations (i.e.,

ERCA, CSA and NBE). The initial data that was collected from ERCA contain 704,573 raw data.

Data preparation or preprocessing is always important in a machine learning and pattern recognition process. The purpose of data preparation is to clean the data as much as possible and to put it into a form that is suitable for the selected data mining software. Starting from the data extracted from the source database maintained by ERCA, CSA, and NBE, a number of transformations were performed before a working dataset was built. The activities during this phase included data cleaning, data selection, attribute or feature selection, transformation and aggregation, integration and formatting, discretization and binarization, dimensionality reduction, minimizing noises, and handling missing values.

Accordingly, the data which was in a relational database format that were stored using different application programs is first exported into a single table format of an Excel sheet. This is mainly because the Weka tool supports a single table data format for processing. Moreover, removal of some attributes was done since they were irrelevant to the objective of this research, redundant, and have no values at all. Through attribute creation and aggregation of attribute values for the import export data set, a total of 10 features were used for many sided analysis.

4.1 Model Building

The first task of the experiment was to understand and heuristically identify attributes or features related to the goal of the machine learning task which will obviously be evaluated by the machine learning process through attribute selection. Feature or attribute selection is deciding on the data to be used for analysis. Criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types. It particularly covers selection of attributes (columns) in a table. The reason of selecting features are the time it takes to build a model increases with the number of variables and blindly including extraneous columns can lead to incorrect models [26]. Thus given the data mining task mentioned above, 10 attributes were identified to be relevant and have been selected. Descriptions of the attributes are shown in Table 1.

Table 1: List of selected attributes

Attribute	*Data Type*	*Description*	*Remark*
Month	Number	Export or Import month for each item.	
Year	Number	Export or Import Year for each item.	
Price	Currency	The price of each item in the market	Initially it was named UNIT PRICE
Item_Code	Number	Code of the import or export items	Initially it was named HS_CODE
Item_Description	Categorical or nominal	Description of the import or export items	Initially it was named HS_DESCRIPTION
Currency rate	Number	Monthly average currency exchange rate.	Initially it was named Average Weighted Rate
CIF_Value (ETB)	Number	The total cost of imported or exported items in Ethiopian Birr	
CIF_Value (USD)	Number	The total cost of imported or exported items in US dollar	
ITEM_TYPE	Categorical or nominal	Show either item is imported or exported item.	Derived
Net_Mass (Kg)	Number	The net weight of the items	

4.2 Result

The results of the experiments confirmed that the techniques of data mining are applicable to generate knowledge from import and export items data in custom administration. The result indicates more than 78% is imported textile. This imported textile is highly associated with price and currency rate. This indicates that imported textile is an item that is highly associated with the market price and by extension currency exchange rate in Ethiopia. Next to textile is food, an item that is associated with market price and currency rate. Also from the four association rule algorithms, Apriori is the fastest algorithm and Tertius is the slowest when compared with all other algorithms that we used in this research.

5. Conclusion and Recommendation

This research attempted to study the possible application of data mining techniques, especially association rule, to identify import and export items' association with market price in Ethiopia. The study was conducted in five major phases, namely problem domain understanding, data understanding, data preparation, model building, and evaluation. However, since data mining task is an iterative process, these steps were not followed strictly in a linear order, but through spiral approach where significant improvement is achieved.

The result of the experiment shows imported textile is an item that highly is associated with market price and currency exchange rate. The domain experts in custom administration must properly handle this item. This means that they must properly follow up this item in order to minimalize its immediate implication on market price. The policy makers also should find solutions (like manufacturing those textile items in Ethiopia as import substitution or tax avoidance) to control the market price and currency exchange rate inflation.

The researchers used different association rule algorithms namely Apriori, Tertius, PredictiveApriori, and FliteredApriori to construct a model. From these algorithms, Apriori and FliteredApriori generate similar model with almost equal time. The time it takes to generate 100 rules by Apriori was 2.075 seconds and FliteredApriori was 2.90 seconds. But both Tertius and PredictiveApriori generate different models than Apriori and FliteredApriori. Tertius was the slowest algorithm when compared with all other algorithms. It took more than an hour to generate 10 rules and most of the generated rules were two itemset rules. PredictiveApriori generates the rule that only associates numerical data and was also slower than both Apriori and FliteredApriori. PredictiveApriori took 5 minutes to generate 100 rules. However, given the fact that there has been treatment in the data in the process of normalization (such as converting random values of a year into monthly based), the result might be affected by the changes.

In general, the results from this study were encouraging. It was possible to implement data mining techniques on the import and export items of custom administration data and market price. It is the authors' belief that a more thorough study using data mining techniques can help to understand the association of import and export items with market price in Ethiopia. The results of this experiment could be employed as an input for the decision making process for decision makers of custom administration and consumer market association.

References

[1] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems." IEEE Trans. on Knowledge and Data Eng., 1996.

[2] Bigus, J.P., "Data Mining with Neural Networks in Solving Business Problems-From Application to Development to Decision Support", McGraw-Hill, 1996.

[3] Chris Clifton and Gary Gengo, "Developing Custom Intrusion Detection Filters Using Data Mining", 2000.

[4] Elias A. Ali, "The Effect of Depreciation of Birr on Major Export Products of Ethiopia: The Case of Hides and Skins", 2011.

[5] Han, G. Karypis and Kumar, "Scalable Parallel Data Mining for Association Rule", IEEE Transactions on Knowledge and Data Engineering, 2000.

[6] Han, Jiawei and Kamber, Micheline, "Data Mining: Concepts and Techniques and Applications", Morgan Kufman Publishers, 2001.

[7] Hailegiorgis Biramo Allaro", "The Impact of Trade Liberalization on the Ethiopia's Trade Balance", American Journal of Economics, 2(5): pp. 75-81, 2012.

[8] Dnuggets Polls, "Data Mining Methodology", 2007.

[9] Laporte, "Risk Management Systems: Using Data Mining in Developing Countries' Customs Administrations", World Customs Journal, 2011.

[10] Li Yan-hai and Sun Lin-yan, "Study and Applications of Data Mining to the Structure Risk Analysis of Customs Declaration", 2005.

[11] MJA Berry, "Data Mining Techniques: for Marketing, Sales, and Customer Support", John Willey & Sons, 1997.

[12] M. S. Chen, J. Han, and P. S. Yu., "Data Mining: An overview from a Database Perspective", IEEE Trans. on Knowledge and Data Eng., pp. 866–884, 1996.

[13] Padhraic Smyth, "Breaking Out of the Black-Box: Research Challenges in Data Mining", 2001.

[14] Paul Dorosh and Hashim Ahmed, "Foreign Exchange Rationing, Wheat Markets and Food Security in Ethiopia", Development Strategy and Governance Division, International Food Policy Research Institute – Ethiopia Strategy Support Program 2, October 2009.

[15] Peter A. Flach and Nicolas Lachiche, "Confirmation-Guided Discovery of First-Order Rules with Tertius", 2006.

[16] Philip K. Chan, "Distributed Data Mining in Credit Card Fraud Detection.

[17] P. Smyth and R. M. Goodman, "Rule Induction Using Information Theory", Proc. Intl. Conference on Knowledge Discovery and Data Mining, pp. 159-176, 1999.

[18] Raghavan, V., Deogun, J. S., and Sever, H., "Knowledge Discovery and Data Mining: Introduction", Journal of American Society for Information Science, Vol. 49, 1998.

[19] Refas, Salim and Thomas Cantens, "Why Does Cargo Spend Weeks in African Ports? The case of Doula, Cameroon", Policy Research Working Paper Series 5565, The World Bank, 2011.

[20] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. 20th Intl. Conference on Very Large Databases, pp. 487–499, 1994.

[21] R. Agrawal, T. Imielinski, and A. Swami., "Mining Association Rules Between Sets of Items in a Large Database." Proc. ACM SIGMOD Intl. Conference on Management of Data, pp. 207–216, 1993.

[22] R. Agrawal and J.C. Shafer, "Parallel Mining of Association Rules," IEEE Trans. on Knowledge and Data Eng., 8(6):962-969, December 1996.

[23] R. Agrawal and R. Srikant, "Mining Sequential Patterns", In Proc. 11th Intl. Conf. on Data Engineering, Taipi, Taiwan, March 1995.

[24] Santos-Paulino, A and A. P. Thirlwall., "The Impact of Trade Liberalization on Exports, Imports and the Balance of Payments of Developing Countries", Economic Journal, Royal Economic Society, 114(493), F50-F72, 02. 2004.

[25] Tobias Scheffer, "Finding Association Rules that Trade Support Optimally Against Confidence", 2004.

[26] Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, 3rd edition, 2005.

[28] Trybula, Walter J., "Data Mining and Knowledge Discovery", Annual review of Information Science and Technology (ARIST), pp. 197-229, 1997.