A Machine Learning Approach for Business Intelligence Analysis using Commercial Shipping Transaction Data

Lisa Bramer, Samrat Chatterjee, Aimee Holmes, Sean Robinson, Steven Bradley, and Bobbie-Jo Webb-Robertson

Abstract— **Business intelligence problems are particularly challenging due to the use of large volume and high velocity data in attempts to model and explain complex underlying phenomena. Incremental machine learning based approaches for summarizing trends and identifying anomalous behavior are often desirable in such conditions to assist domain experts in characterizing their data. The overall goal of this research is to develop a machine learning algorithm that enables predictive analysis on streaming data, detects changes and anomalies in the data, and can evolve based on the dynamic behavior of the data. Commercial shipping transaction data for the U.S. is used to develop and test a Naïve Bayes model that classifies several companies into lines of businesses and demonstrates an ability to predict when the behavior of these companies changes by venturing into other lines of businesses.**

Keywords- **Incremental machine learning; Naïve Bayes model; Business intelligence; Commercial shipping data**

I. Introduction

MANY "intelligence" problems are particularly challenging because of the complexity of the underlying phenomenon and the lack of consensus on "ground truth" that drives the need to have a team of expert analysts apply their collective knowledge. In some cases, the volume and velocity of data to be analyzed makes the application of machine-based reasoning desirable to assist these domain experts in their analysis, but many new analytic advances are needed to realize such an operational capability.

This study utilizes the Port Import/Export Reporting Service (PIERS) data [1]—a comprehensive database of U.S. international trade—to drive the research for developing advanced intelligence capabilities.

The PIERS data consists of commercially available U.S. import and export shipping transactions, which are typically used for competitive business intelligence. In this paper, this data is utilized specifically to: 1) characterize the lines of business (LOB) to which a particular company belongs based on their procurement activity, and 2) detect possible dynamic changes in LOB as a company's procurement behavior varies. From a business intelligence perspective, it is important to understand when competitors make significant changes to their business operations, especially expansions into new lines of business. While the use of PIERS data is focused on a business intelligence problem, it serves as a proxy to address analytic challenges that may be applicable to other domains.

We begin by discussing key analytic challenges and past work. This is followed by a description of the PIERS dataset and our machine learning based methodology for LOB classification. The results of our algorithms are presented next. We conclude with a discussion of possible extensions of this work.

II. Analytic Challenges and Past Work

Our research approach is driven by commercial shipping transactions for a set of companies over a ten-year period, and produces hypotheses about whether these companies are changing their LOB. While, at first glance, this may seem straight forward, there are analytic challenges that are discussed below; along with a summary of past work using the PIERS data.

A. Dynamic Models

Standard supervised machine learning techniques may be applied to build models that classify a company to a LOB based on features extracted from the PIERS shipping records. However, a company changing its procurement behaviors does not necessarily indicate that it is expanding into a new line of business. If a majority of companies within an LOB happen to adopt similar new procurement behaviors, then one could just as accurately infer that these companies are not expanding into new LOBs, but are simply reacting to a dynamic business environment that is having an impact on the LOB as a whole. Rather than inundating analysts with inaccurate hypotheses, we would want the models to detect this LOB-wide behavior change and evolve accordingly.

The research described in this paper is part of the Analysis in Motion Initiative at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

L. Bramer is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (phone: 509-375-4553; fax: 509-375-2522; e-mail: lisa.bramer @pnnl.gov).

S. Chatterjee is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: samrat.chatterjee@pnnl.gov).

A. Holmes is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: aimee.holmes@pnnl.gov).

S. Robinson is with the Human Centered Analytics Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: aimee.holmes@pnnl.gov).

S. Bradley is with the Cyber Innovation & Operations Center, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: steven.bradley@pnnl.gov).

B-J. Webb-Robertson is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: bobbie-jo.webb-robertson@pnnl.gov)

B. *Hypothesis Rationale*

Models that generate inductive as well as deductive hypotheses could be useful for domain experts. For example, it may be helpful for a user to be alerted that a company is suddenly behaving in ways that are no longer consistent with its previously classified LOB, it appears to be also useful for expert analysts to know *why* the models have reached that conclusion. For example, a classification model may compute over the last 90 days that the likelihood has dropped from 98% to 90% that Ford Motor Company is an automobile manufacturer; however, this doesn't provide the analyst with the insight required to assess whether the models took into account observations that she missed or whether she believes that the models are flawed, which is critical for model steering.

C. *Machine Learning with Streaming Data*

Desirable features of machine learning models from streaming data involve: 1) accounting for *recent history* when making predictions, and 2) allowing the models to *evolve* or *update* with the data streams. Conditioning predictions based on history, with moving training windows, is an approach that addresses the first case above. For the second case, a machine-learning algorithm that incrementally learns over the data and updates the model with new training instances appears to be appropriate. Giraud-Carrier [2] describes incremental learning as applied to tasks and algorithms. An incremental learning task involves the availability of training examples over time; and an incremental learning algorithm, also referred to as a *memoryless online* algorithm, produces hypotheses that depend on past hypothesis and the current training example.

D. *Past Work using PIERS Data*

Limited applications were found in the open source literature that involved the use of PIERS records for data mining. Jeske et al. [3] describe a platform for generating synthetic data for testing data mining tools. They implemented a resampling data generation algorithm using the PIERS data.

Das and Schneider [4] describe an anomaly detection problem and discuss the use of unsupervised methods applied to categorical datasets, including: association rule; likelihood; and bayesian network based approaches. The authors implemented a likelihood-based approach using the PIERS data to detect unusual shipments among all imports into the country. The focus was on detecting unusual combinations of attribute values in the data.

III. Datasets

Our study analyzed PIERS import data records [1], from January 2005 to December 2014. The PIERS database contains records for every company importing or exporting goods in the U.S. For this study, we selected a subset of these companies, in particular 17 companies that could be categorized within one of three lines of businesses. These companies were selected because they had a large number of records available and had a well-defined LOB. Future analysis will incorporate other lines of business and companies. PIERS data is rich with shipment related information and at times is noisy with possibly inconsistent data entries. Access to the PIERS data records was made possible due to the establishment of a strategic goods testbed (or data library) at PNNL [5]. The PNNL testbed team has created a centralized data location and with a single agreement allows access to the PIERS data for research purposes. The lines of businesses include: 1) Automotive, 2) Clothing, and 3) Appliance. The Automotive companies chosen were BMW, Ford, Honda, Hyundai, Nissan, Toyota, and Volkswagen. Clothing companies were Guess, Gymboree, Hennes & Mauritz, J Crew, Levi, and Ralph Lauren. Finally, the appliance companies considered were Bosch, Electrolux, General Electric, and LG Electronics. The 10-year shipping record counts associated with these companies ranged from 108,828 for LG Electronics to 7,572 for Gymboree.

In addition to the companies mentioned above, we also merged records for several pairs of companies belonging to different lines of businesses (where over time, the record counts from the starting LOB company incrementally decreases and the other LOB company increases). The motivation behind this merge was to test whether our classification algorithms can detect changing LOB over time. Several hybrid companies were formed with several different rates of change. For illustration purposes, we examine one such hybrid, which started with records from Ford and slowly injected records from Old Navy into the data over time.

Categorical	Quantitative
Date Shipper Shipper Address Consignee Consignee Address Carrier Country of Origin Port of Arrival Port of Departure U.S. Destination HS Code Short Commodity Description	Weight (lb, kg, etc.) Measure (cubic ft, etc.) Quantity (bags, pkgs, etc.) Estimated Value

Table 1. Examples of PIERS record attributes by variable class.

Every record in the dataset has as many as 54 different attributes. These attributes contain information about the shipper, shipment, and arrival/departure locations. Table 1 presents a selection of these attributes separated by variable class: quantitative or categorical. A challenge working with this dataset was the identification of attributes that characterize and classify companies into a LOB and can help detect deviations with dynamic changes in procurement behavior. One such challenge is that limited quantitative variables are available, and the variables that are available are recorded with many different units of measurement. Additionally, in some records, no units of measurement are recorded. Many categorical variables are available including but not limited to: the final destination of the shipment, the departure port, the harmonized system (HS) code for tariff purposes, and a commodity short description.

IV. Methodological Approach

Our modeling methodology is comprised of five steps: 1) identification of key data attributes, 2) creation of a data-driven library of attribute values, 3) selection of a machine learning model, 4) training and testing strategy, and 5) model evolution plan. A description of each methodological step follows.

A. *Selected Data Attributes*

The choice of data attributes was driven by their potential to characterize a company within a LOB. We explored the evolution of several attributes over time for various companies, and our attribute set for further analysis included: 1) *Commodity Description*, 2) *U.S. Destination*, and 3) *Port of Departure*. All three selected attributes contain text information. Commodity description contains blocks of text associated with the shipment and/or company. Since we only consider import data, U.S. Destination is listed a city within the U.S. where the shipment is headed, and Port of Departure is a foreign port where the shipment began its journey. Figure 1 presents an example frequency plot of words that are contained in the commodity description field of Hennes & Mauritz, over a subset of time. In this example, *Ladies* is the most frequently occurring word.

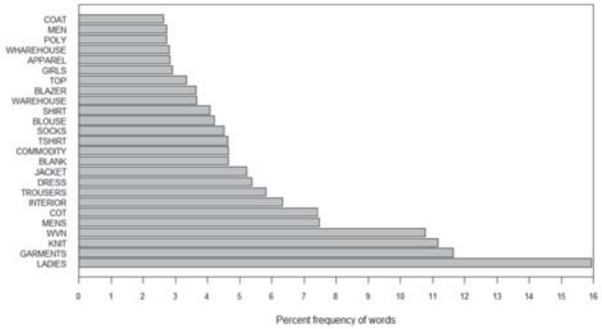

Fig. 1. Example frequency plot of words contained within commodity description attribute for Hennes & Mauritz.

Similarly, example frequency plots of shipment counts by U.S. Destination over time were prepared to assess variability of shipment location characteristics (see Figure 2 for data associated with Hennes & Mauritz). Each plot in Figure 2 corresponds to a different U.S. Destination city and each bar in a given plot corresponds to shipment counts for a chosen time block. In this example, the location with the most frequent spikes/bars (i.e. count of arriving shipments) is *New York City*.

B. *Library of Attribute Values*

Attribute values (or text strings) were first split to create a list of unique keywords, final U.S. destination cities, and departure ports for each company within the three lines of businesses. The percent occurrence frequencies of these unique attribute values were then computed for different blocks of time; and an overall mean percent frequency was evaluated. Table 2 presents an example of percent frequencies of keywords for Hennes & Mauritz. Similar frequency tables were created for the cities and departure ports. A minimum mean occurrence threshold level of 5% was chosen to select unique keywords, and a threshold of 2% was chosen for selecting cities and departure ports; leading to the creation of the attribute value library.

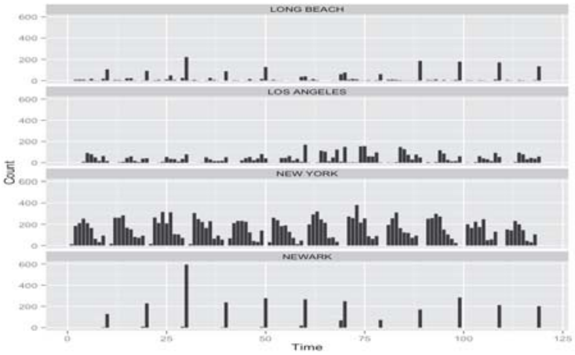

Fig. 2. Example frequency plot of shipment counts by U.S. destination for Hennes & Mauritz. Each plot corresponds to a different U.S. Destination city and each bar in a given plot corresponds to shipment counts for a chosen time block.

Three attribute value libraries were developed thereafter; one each for the keywords, cities, and departure ports covering information from all companies across all lines of businesses. Each library contains a list of primary attribute values along with their spelling and parts of speech variations found within the data records. For example, the keyword library list item *Auto* along with *Automobile*, *Automotive*, and *Autos*. These data libraries were key inputs for training the machine learning algorithms.

Sample Attribute Value Percent Frequencies

Keyword	Time Block 1	Time Block 2	…	Overall Mean
Cot	0.185	0.185	…	0.207
Ladies	0.140	0.145	…	0.162
Knit	0.115	0.150	…	0.149
⋮	⋮	⋮	⋮	⋮

Table 2. Attribute frequency as a proportion of records.

C. *Machine Learning Algorithms*

A large number of features could possibly be extracted from the three selected attributes within the PIERS data.

Moreover, dependencies may also exist among these features. As a starting approach, a Naïve Bayes classification technique [6] was adopted for the LOB classification problem. The Naïve Bayes approach is based on Bayes theorem and assumes that conditional probabilities of independent variables are statistically independent. Three independent Naïve Bayes models, one for each LOB, were fit to training data from companies from all three LOB's. Thus, the conditional probabilities of a company being in each LOB do not necessarily have to sum to one.

The nodes or explanatory variables in the Naïve Bayes model were the proportion of records in a given timeframe that contained each of the items listed in the data libraries for keywords, cities, and departure ports. As a result, we had 163 total nodes (83 keyword types, 26 cities, and 54 departure ports). A probabilistic expression for the Naïve Bayes algorithm LOB classifier can be expressed as:

$$P(B_i|W_1, W_2, \dots, C_1, C_2, \dots\ D_1, D_2, \dots) \propto \prod_{j=1}^{83} P(W_j|B_i) \cdot \prod_{k=1}^{26} P(C_k|B_i) \cdot \prod_{q=1}^{54} P(D_q|B_i) \cdot P(B_i) \qquad (1)$$

where *B* refers to a LOB, *W*, *C*, and *D* refer to the proportion of records that contained a keyword K, a destination city *C*, and a departure port *D*, respectively. $P(B)$ is the prior probability of a LOB, and $P(x|y)$ is the conditional probability of event *x* given event *y* is observed.

D. Training and Testing

The first 5,000 records of each of the 17 companies were used as training data for each of the Naïve Bayes models. The explanatory variables were calculated for windows of training records of 150 records. Training cases were computed for moving windows of 150 records with a step size of 50 records. Each rolling window was evaluated on the attribute values of interest (for keywords, U.S. Destination, and Port of Departure) and a proportion of occurrence was calculated. The step between different windows was of size 50 records, and this rolling window process was repeated over all of the training records. Additionally, each training set summary record was assigned a response variable of one or zero (for each Naïve Bayes model: Appliance, Automotive, and Clothing) indicating the company's true LOB during the training period.

A separate Naïve Bayes model was fit from the training data for each LOB: clothing, automotive, and appliances, resulting in a total of three models. Fitting models for each LOB independently allows for the possibility of an individual company behaving in a manner similar to several LOBs and does not force predicted probabilities to sum to one. Predictions of a company's LOB can be generated for any reasonable moving block size at a true streaming level (i.e. a new predicted probability of each LOB can be generated with each new incoming record). However, for the purpose of demonstration here, the testing data that was then evaluated on these models was again created by a similar method to that described above for the training data (window width = 150, sliding windows), except that the step between different windows was of size 15 records. The testing data for each company was comprised of the remaining records for each company (after the first 5,000 records were removed for training purposes) over a ten year period as described previously.

V. Results and Discussion

We proceed by evaluating the predictive capability of the Naïve Bayes models with the 17 companies previously discussed. We then investigate the models' capability to detect changes in company behavior, by examining model performance for the aforementioned hybrid company.

The accuracy of each model for each company was assessed for the testing data. Figure 3 summarizes the accuracy of each model by company. Most clothing companies had a near perfect accuracy across all three models. The accuracy of the clothing LOB model is very accurate (greater than 95% accuracy) for all companies. However, the accuracy of the auto and appliance LOB models performed less accurately in the case of a few companies. For example, the auto LOB model incorrectly identified Bosch as an automobile company in more than half of the testing data cases. This behavior is not entirely unexpected as both the automobile and appliance industries involve electronics and other similar products. Upon further inspection, Bosch contained many records with keywords that were also seen in the automobile companies (e.g. parts, motor, etc.).

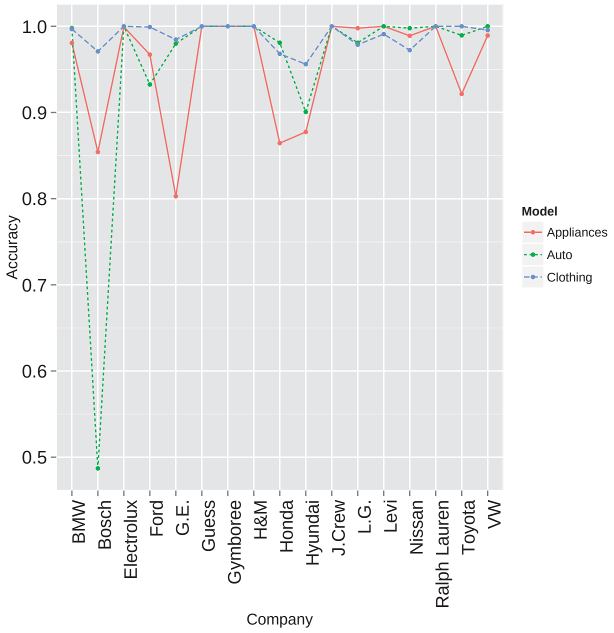

Fig. 3. Naïve Bayes model accuracy by company for three LOB models.

The overall model performance was assessed taking into account all companies. Because the number of records, and thus testing data points, varied from one company to another, we consider the first 250 summarized window testing data points. Table 3 summarizes the accuracy, false positive rate, and false negative rate for each of the LOB models. Overall, the three models are able to discriminate between different LOB's. Additionally, Figure 4 gives a receiver operating characteristic curve (ROC) for the clothing LOB model.

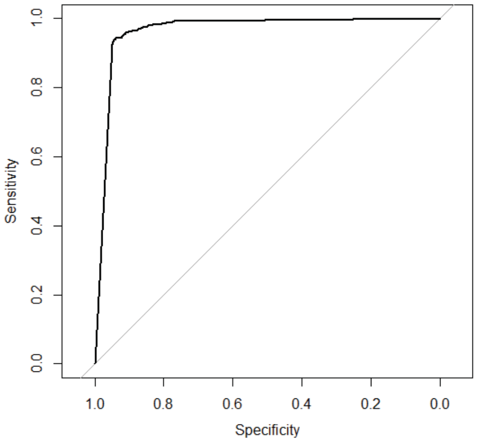

Fig. 4. ROC curve for clothing LOB model over all companies.

We further evaluate the models' ability to identify changes in company procurement/LOB behavior by generating predictions for a hybrid data set that transitions from a purely automotive company to adding a partial LOB in the clothing industry.

Model	Accuracy	FPR	FNR
Auto	0.9877	0.0204	0.0009
Clothing	0.9971	0.0045	0.0001
Appliances	0.9789	0.0100	0.0571

Table 3. Overall performance metrics for each LOB model.

Figure 5 shows the predicted probability of the hybrid company belonging to each LOB. In the beginning periods of the testing data when the testing data is comprised of just automobile records, the models classify the pure records correctly. Additionally it can be seen that the models pick up on the injection of clothing records into the testing data. However, the predicted probabilities tend to switch between the two models in a dichotomous manner. This behavior is due to some highly discriminate explanatory variables (e.g. keywords of auto or seatbelt). When these words appear in the dataset in any proportion, the records get classified as being from the automobile LOB. This dichotomous behavior continues, because the model is never updated to reflect changes in company procurement habits and entry into a new LOB.

VI. Conclusions and Future Work

We have demonstrated that Naïve Bayes classification models using keywords, destination cities, and ports of departures are able to effectively classify a businesses LOB, based on past procurement behavior. Additionally, these models are able to detect changes in a company's procurement behavior. However, the ability to model a company going into a second LOB and accurately model the company still participating in the original LOB was unsuccessful with only dichotomous training examples.

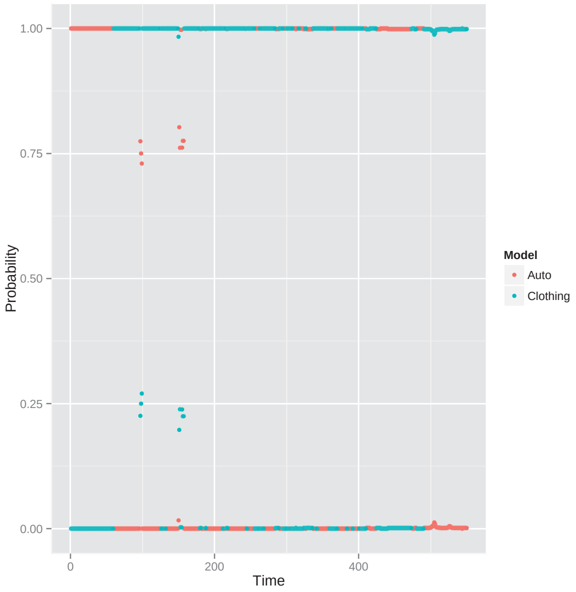

Fig. 5. Probabilities of LOB's for records in the hybrid auto and clothing company.

A class of algorithms that may naturally support predictive analysis on this streaming data may be found in the vicinity of incremental machine learning. Traditional machine learning approaches assume that a good training set is always available a priori and contains all the required knowledge to construct sufficient models that may applied to new examples or problems, which is not the case when changes in data dynamics are present. A wide variety of incremental learning algorithms have been developed in machine learning areas such as Bayesian networks [7-9], neural networks [6-7], support vector machines [10-12], and decision trees [13]. These methods should be adapted to automatically generate or retrain the incremental models to automatically evolve as drifts in company behavior and procurement features emerge in the data streams. Additionally, metrics for model evolution and the evolution of model features should be developed to help in eliciting domain expert feedback.

References

[1] JOC Group. (2015, March). PIERS Data. Available: https://www.piers.com/.

[2] C. Giraud-Carrier, "A note on the utility of incremental learning," *AI Communications*, vol. 13, pp. 215–223, 2000.

[3] D.R. Jeske, P.J. Lin, C. Rendon, R. Xiao, and B. Samadi, "Synthetic data generation capabilities for testing data mining tools," in *Proc. MILCOM'06 – 2006 IEEE Conference on Military Communications*, Washington, DC, October 23–25, 2006, pp. 3449–3454.

[4] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," in *Proc. KDD'07 – The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, August 12–15, 2007, pp. 220–229.

[5] J.B. Webster, L.E. Erikson, C. Toomey, and V.A. Lewis, “PNNL strategic goods testbed: a data library for illicit nuclear trafficking,” Pacific Northwest National Laboratory, Richland, WA Technical Rep. PNNL-SA-102611, 2014.

[6] StatSoft. (2015, March). Naïve Bayes Classifier. Available: http://www.statsoft.com/textbook/naive-bayes-classifier.

[7] Daly, R., Shen, Q., and Aitken, S. (2011). Learning Bayesian Networks: Approaches and Issues. *The Knowledge Engineering Review*, 26(2), pp. 99-157.

[8] Samet, S., Miri, A., and Granger, E. (2013). Incremental Learning of Privacy-Preserving Bayesian Networks. *Applied Soft Computing,* 13(2013), pp. 3657-3667.

[9] Cauwenberghs, G. and Poggio, T. (2000). Incremental and Decremental Support Vector Machine Learning. In *Proc. of NIPS*, pp. 409-415.

[10] Diehl, C.P. and Cauwenberghs, G. (2003). SVM Incremental Learning, Adaptation and Optimization. In *Proceedings of the 2003 International Joint Conference on Neural Networks*, pp. 2685-2690.

[11] Ralaivola, L. and d’Alche-Buc, F. (2001). Incremental Support Vector Machine Learning: A Local Approach. In *Proceedings of ICANN*, pp. 322-329.

[12] Chao, S. and Wong, F. (2009). An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining. In *Proc. of the 8th International Conference on Machine Learning and Cybernetics,* pp. 1694-1699.