Selecting a Classification Ensemble and Detecting Process Drift in an Evolving Data Stream

Alejandro Heredia-Langner[1], Luke R. Rodriguez[1], Andy Lin[1], Jennifer B. Webster[1]
[1]Pacific Northwest National Laboratory, 902 Battelle Boulevard, PO Box 999 Richland, Washington 99352

***Abstract*— We characterize the commercial behavior of a group of companies in a common line of business, or network, by applying small ensembles of classifiers to a stream of records containing commercial activity information. This approach is able to effectively find a subset of classifiers that can be used to predict company labels with reasonable accuracy. Performance of the ensemble, its error rate under stable conditions, can be characterized using an exponentially weighted moving average (EWMA) statistic. The behavior of the EWMA statistic can be used to monitor a record stream from the commercial network and determine when significant changes have occurred. Results indicate that larger classification ensembles may not necessarily be optimal, pointing to the need to search the combinatorial space spanned by the classifiers in a systematic way. Results also show that current and past performance of an ensemble can be used to detect when statistically significant changes in the activity of the commercial network have occurred. The dataset used in this work contains tens of thousands of high level commercial activity records with continuous and categorical variables and hundreds of labels, making classification challenging.**
***Key Words:* Ensemble classifiers, EWMA, optimization**.

1. Introduction

Approaches to mine and analyze streaming data that use a single classifier or a fixed ensemble assume that the classifiers at hand are, as a set, optimal for the problem under consideration. Under evolving conditions and at the rate at which data streams are generated in today's commercial, scientific, and security environments, it is unlikely that a single classifier, or even a fixed ensemble, can reliably provide an acceptable level of performance for a prolonged period of time. In addition to making reliable predictions in high volume data streams, researchers and analysts may also be interested in using classifiers to detect when the behavior of the data has changed significantly`. Significant changes in the predictive performance of a classification system signal the need for an analyst to get involved, determine the cause, and decide if the classifiers need to be updated. This situation is difficult when multivariate and complex data streams are involved, such as those considered here, containing large scale business activity of companies in the private sector.

Working with classification ensembles can be challenging because of the size of the combinatorial space that needs to be explored when searching for an optimal set for the current operating conditions. Exhaustive search is only possible if the number of distinct classifiers available is relatively small, while larger spaces can only be partially explored. Researchers in [1] investigate the performance of 15 classifiers on a variety of datasets using several search methods and optimality criteria and find that, in general, the best results are produced when using a direct search approach for the selection of an optimal ensemble. They also find that using a criterion that correlates strongly with the overall classification error to determine performance produces better results than using other measures of classifier diversity.

Even when an optimal ensemble can be found, it may remain so only for a narrow period, since high volume data rates usually mean that only a relatively small window of records is available to characterize the data stream and train the classifiers. The selection of an adequate window of training data can in itself be a difficult problem. Results in [2] indicate that using a fixed number of records can be problematic, since a wide window may make classifiers insensitive to trends and a narrow one may result in classifiers that simply chase the noise in the data. For this reason, it is important to detect when significant changes in the underlying distribution of the data stream have occurred.

An additional difficulty arising when applying a classification ensemble to a data stream is determining how and whether the ensemble should be modified. In [3] examples are presented of a dynamic weighted majority ensemble method where individual classifiers (also called base learners or experts) are selected based on the performance of the ensemble. In that approach, new individual classifiers are added to the ensemble when a threshold of poor performance by the current ensemble has been crossed, while the influence of some base learners currently present may be down-weighted if their individual performance is poor. Results in [3] are encouraging, but the size of the ensembles considered can become large, unless this number is explicitly restricted.

To address the issue of the changing nature of streaming data, also known as concept drift, and the detection of a point where new records should be obtained for training and selecting a new ensemble, we propose the use of an exponentially weighted moving average statistic to detect significant concept drift and the use of a small population of classifiers to build a classification system. In our approach, different combinations of individual classifiers are used to find an ensemble that is optimal for the current conditions and that can help estimate the effect that each individual classifier has on the overall classification rate. The ensemble selected can then be

applied to a stream of new records for as long as a stable and acceptable level of performance is maintained. In this way, a new window for re-training and finding a potentially new classification ensemble can occur only when necessary.

We demonstrate this approach though an example using a set of commercial records from nearly 400 companies in the automotive field. Characterizing this set is challenging because it contains continuous and categorical features and because the records can be relatively vague and prone to contain erroneous entries caused by humans entering the data. This means that perfect classification may not be achievable.

2. Materials and Methods

The PIERS records (Port Import/Export Reporting Service, [4]) database contains international trade information from vessels arriving to or departing ports in the U.S. The database contains millions of records with information such as port of entry/departure, estimated value of the shipment, tonnage, and brief descriptions of the contents of the shipment, among many other fields. The information in the PIERS records can be used to research whether and what kinds of relationships exist between certain commercial entities, and, as shown in [5] and [6], is a useful source of data in business analytics.

The data available in the PIERS database can be challenging to analyze because the number of features, or fields, available for each shipment is large and the fields include numerical, categorical and text data. The information available in PIERS can also be fairly ambiguous, such as when a single tariff code is used to describe the contents of a shipment, and the code may cover a wide variety of items. In spite of this, PIERS records contain valuable information about the activity of commercial actors, and this information can be aggregated and analyzed in ways that are meaningful for describing the behavior of companies operating in a business group or network.

For the present work, we employed 52353 records for the year 2013 for 396 companies in the automotive industry. These records are of interest because they contain transactional information between commercial actors in what can be considered a fairly well defined line of business. The first objective of this work was to determine if the records can be used by machine learning algorithms to adequately classify the companies they belong to. The features selected as inputs for the classifiers are shown in Table 1.

Table 1. Name, description and type of the features used to build the classifiers

Feature Name	Description	Type
YRMTH	Combined year and month date of record	Continuous
CTRYCODE	Code for country of origin/destination	Categorical
FCODE	Foreign port code	Categorical
USCODE	US port code	Categorical
HSCODE	Harmonized System Code, for tariff purposes	Categorical
QTY	Quantity shipped, integer	Continuous
MTONS	Metric Tons shipped	Continuous
TEUS	Twenty-Foot Equivalent Units, integer	Continuous
VALUE	Estimated value, USD	Continuous
CONVOL	Container Volume, m^3	Continuous

This set of records is challenging for classifiers because it is highly unbalanced (some companies have thousands of records to train on while others have only a few), there are more than 80 different countries involved in the trades and hundreds of tariff codes used for the items shipped by the companies in this network.

The classifiers employed include a Naïve Bayes (NB) classifier, a k-nearest neighbor (k-NN) classifier and two classification trees. The classification trees employ different split criteria: Gini's diversity index (GDI) and maximum deviance reduction. All results in this document were obtained using MatLab [7]. Because only four individual classifiers are involved, it was possible to explore the performance of different ensembles using a factorial approach, where all possible combinations of the four classifiers are applied to the same training/testing partitions of the data.

3. Results

There are 16 classifier combinations, including one where none of the four individual classifiers is used. The case with no classifiers represents a truly naïve predictor, used to establish a lower bound on performance. The truly naïve predictor produces labels for new records randomly, but in the same proportions as those found in the training set. For example, if 50% of the records in the training set belong to a single company, the truly naïve classifier will predict, with probability of 0.5, that any given record in the test set will belong to that particular company.

The dataset was divided repeatedly and independently into training and test sets using a fixed number of records for training and testing, and the names of the 396 companies as labels. This resulted in a training/testing partition of roughly 74%/26%. Predicted company labels for the test records were obtained directly when a single classifier was used, or by averaging scores when more than one classifier was involved, breaking ties randomly. The process of training and testing was carried out repeatedly and independently to assess the performance of the classifiers. Boxplots of the fraction of mislabeled test records in 10 independent trials are shown in Figure 1. The labels on the x-axis of Figure 1 indicate which classifier combination was used on the test sets, and the individual classifiers are identified in the plot.

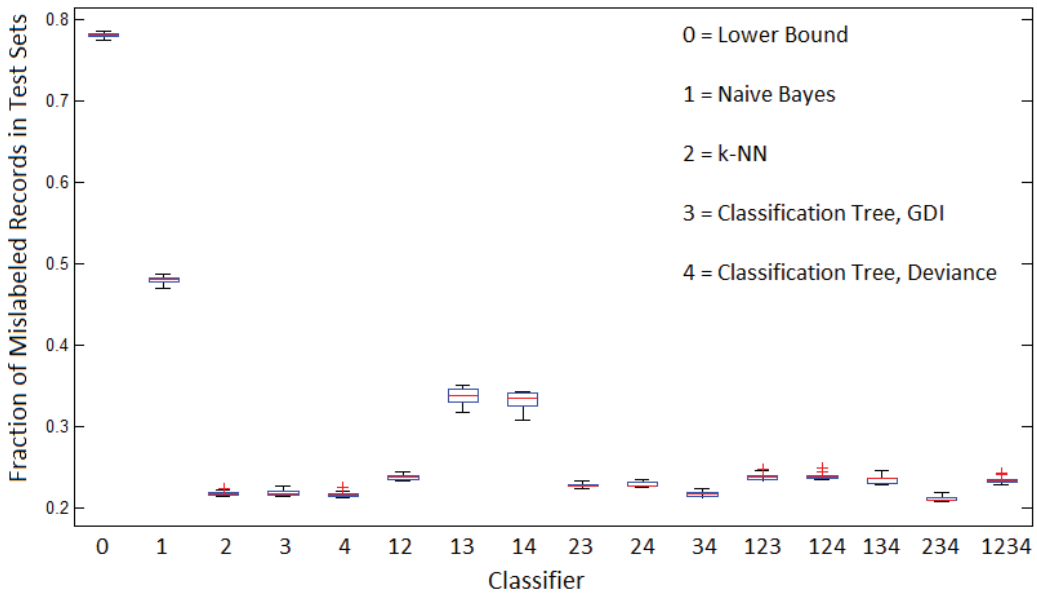

Figure 1. Boxplots of the fraction of mislabeled records in the test sets over ten independent train/test (76%/ 24%) partitions of 52353 records. The fraction of mislabeled records in 10 test sets is shown in the y-axis. The x-axis shows the coded values of the individual classifier or combination used. The boxes in the plot encompass the first, second and third quartiles, with whiskers denoting the most extreme points not considered outliers, and any outliers are marked with crosses.

Figure 1 shows that all the classifiers, alone or in combination, perform better than the truly naïve classifier used to establish a lower bound. It is also interesting to notice that the performance of the Naïve Bayes classifier can, in this case, be dramatically improved by combining its predictions with those from any other classifier available.

Figure 1 shows that there are several good options to choose from when it comes to selecting a classifier to predict the labels in this business network. The boxplots in Figure 1 can be used to choose the ensemble that minimizes the fraction of mislabeled records in a majority of test sets. However, the information gathered can also be used to estimate the impact that each classifier has on the observed error rate. This type of modeling may provide insights into how classifier diversity affects the results. The error rates obtained in the ten train/test trials that produced the results shown in Figure 1 were used as input for a generalized linear model, employing the percent error as the response. The model for the mean predicted percent error rate is:

$$\widehat{Error\ Rate} = \beta_0 + \beta_1 NB + \beta_2 kNN + \beta_3 GDI + \beta_4 DEV + \beta_{12} NB \cdot kNN + \beta_{13} NB \cdot GDI + \beta_{14} NB \cdot DEV + \beta_{23} kNN \cdot GDI + \beta_{24} kNN \cdot DEV + \beta_{34} GDI \cdot DEV + \beta_{123} NB \cdot kNN \cdot GDI + \beta_{124} NB \cdot kNN \cdot DEV + \beta_{134} NB \cdot GDI \cdot DEV + \beta_{234} kNN \cdot GDI \cdot DEV + \beta_{1234} NB \cdot kNN \cdot GDI \cdot DEV$$

where NB represents use of the Naïve Bayes classifier, kNN represents use of the k-NN classifier, GDI and DEV represent use of the respective classification tree, and the β_i are model parameters, which are estimated using iteratively reweighted least squares [8]. Model parameter estimates and related statistics are shown in Table 2.

Table 2. Percent error rate model parameter estimates and their corresponding standard errors, t- and p-values

Model Parameter	Estimate	Std. Error	t-value	p-value
β_0	78.10	0.1887	413.78	<0.0001
β_1	-30.20	0.2669	-113.14	<0.0001
β_2	-56.30	0.2669	-210.92	<0.0001
β_3	-56.10	0.2669	-210.17	<0.0001
β_4	-56.10	0.2669	-210.17	<0.0001
β_{12}	32.20	0.3775	85.30	<0.0001
β_{13}	41.90	0.3775	111.00	<0.0001
β_{14}	41.30	0.3775	109.41	<0.0001
β_{23}	57.20	0.3775	151.53	<0.0001
β_{24}	57.50	0.3775	152.32	<0.0001
β_{34}	55.80	0.3775	147.82	<0.0001
β_{123}	-42.70	0.5339	-79.98	<0.0001
β_{124}	-42.50	0.5339	-79.61	<0.0001
β_{134}	-51.00	0.5339	-95.53	<0.0001
β_{234}	-58.90	0.5339	-110.33	<0.0001
β_{1234}	53.10	0.7550	70.33	<0.0001

The model with the parameter estimates in Table 2 has an R^2 of 0.9984, an adjusted R^2 of 0.9983 and all of the parameters have highly significant p-values. Analysis of other performance statistics did not reveal major anomalies with the model. The model parameters in Table 2 show how the presence or absence of each classifier affects the predicted percentage error rate, and how individual classifiers interact with each other. The model can be useful because it can be employed to determine what level of improvement can be expected when adding or removing a particular subset of classifiers to a stream of data that retains the characteristics of the training sets.

Analysis of the model developed indicates that an ensemble that includes the k-NN classifier and the two classification trees produces the best predicted mean percentage error rate. However, the analysis also indicates that using any one of those three classifiers alone would produce results that are nearly as good. The model and parameter estimates also provide information of how the diversity in this set of classifiers affects the accuracy of the ensemble. As shown in Figure 1, the ensemble that contains all four classifiers does not result in the best rate of correct predictions for this particular dataset.

After selecting an ensemble with good performance, a key question that remains when applying the ensemble to streaming data is how to detect concept drift. A classifying ensemble can be expected to maintain a stable level of performance only as long as the characteristics of the new records remain more or less the same as those in the training data. For this reason, it is important to know when the behavior of the data stream has changed significantly, so that an analyst or monitoring system can be alerted and a new set of classifiers trained under the new conditions.

Selecting and training a new classification ensemble involves not only additional time and effort, but it also means that, during this time, classification of currently available data has to be put on hold. If the streaming data has not changed in meaningful ways and an ensemble with good performance is available, stopping to acquire new training data and selecting a potentially new ensemble is wasteful and could result in a process that may simply chase the natural noise in the data, increasing the variability of the predictions.

Monitoring the performance of a classification ensemble involves assessing when, and whether, a significant change in the data stream has occurred. Because the performance of a classification ensemble can be measured by its error distribution [9], it is important to find a way to detect when significant changes in the misclassification rate have occurred.

For this work, the performance of an ensemble is measured using an Exponentially Weighted Moving Average (EWMA) applied to a measure of classification error (see [10] for an excellent introduction to the EWMA). An EWMA is a weighted average of current and past observations, and it has been used extensively to monitor performance in industrial and scientific settings [11], [12].

Performance of a classification ensemble applied to streaming data can be characterized and monitored by the number of misclassified observations in a fixed number of records. In this work, the number of misclassified observations in every ten records was used. If c_t is the number of misclassified observations at period *t*, that is, a period that involves ten consecutive records, then the EWMA statistic at period *t* is given by:

$$z_t = \lambda c_t + (1 - \lambda) z_{t-1}$$

where the value for z_0, needed for the first set of ten records (t=1), is computed as the average number of misclassified observations per ten records in the data used to train the ensemble. The EWMA statistic can be monitored using control limits given by:

$$UCL = \bar{c} + k\sqrt{\frac{\lambda\bar{c}}{2-\lambda}} \quad \text{and} \quad LCL = \bar{c} - k\sqrt{\frac{\lambda\bar{c}}{2-\lambda}}$$

where UCL is the Upper Control Limit, LCL is the Lower Control Limit, $\bar{c}$ is the average rate of misclassification in every ten records in the training data, and k and λ are constants chosen so that, if no concept drift is present, the EWMA statistic remains relatively stable and within the control limits. Values of $k = 3$ and $0.05 \leq \lambda \leq 0.25$ are common in practice, but other values can also be used [10]. The LCL and UCL need to be calculated using a training set that is representative and stable, that is, data where no concept drift has occurred.

To test the usefulness of the EWMA in detecting concept drift, the records used to generate the results shown in Figure 1 and Table 2 were used to compute LCL, UCL and EWMA values to monitor the error produced by the optimal classification ensemble. As stated previously, the number of misclassified observations in every ten records in the training data was used to compute the EWMA statistic and calculate the LCL and UCL values. Misclassification rates were obtained using predictions from an ensemble containing the k-NN classifier and the two classification trees. A plot of the EWMA statistic for 12,000 test set records, k =2.7 and λ = 0.08 is shown in Figure 2.

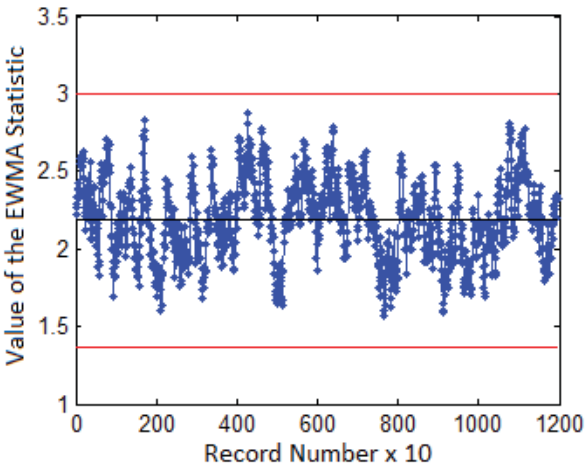

Figure 2. EWMA plot for 12000 test records. Training data for the number of misclassified observations in every ten records and a classification ensemble that includes the k-NN classifier and the classification trees with the GDI and deviance split criteria were used to compute the centerline, LCL (bottom horizontal line) and UCL (top horizontal line) shown in the plot.

Figure 2 shows the behavior of the EWMA for the optimal ensemble applied to test records, that is, records that were not used in training the classifiers. The statistic plotted in Figure 2 shows that the ensemble remains stable around the center line, with fluctuations showing the natural variability of the classification process. Figure 2 indicates that the error rate for the ensemble selected remains close to two observations mislabeled in every ten records, which is consistent with the behavior for this ensemble in Figure 1.

It is of interest to determine if the EWMA shown in Figure 2 can be used to detect changes in the behavior of companies in this particular business network. These changes may come about if, for example, one or more companies in the set start producing shipment records that are more commonly associated with other companies in the network. This type of change in behavior could be the result of individual companies making incursions into new markets or entering new lines of business, information that would be of interest to business analysts.

To investigate if this type of change would result in a significantly different behavior of the EWMA statistic, test sets were produced where the labels for records from a pair of companies were exchanged. This swap impacts around 6% of the total number of records in the test set, leaving the majority of the records unchanged. Using the same EWMA parameter values shown in Figure 2 on the test set with changed records for two companies produces the results in Figure 3.

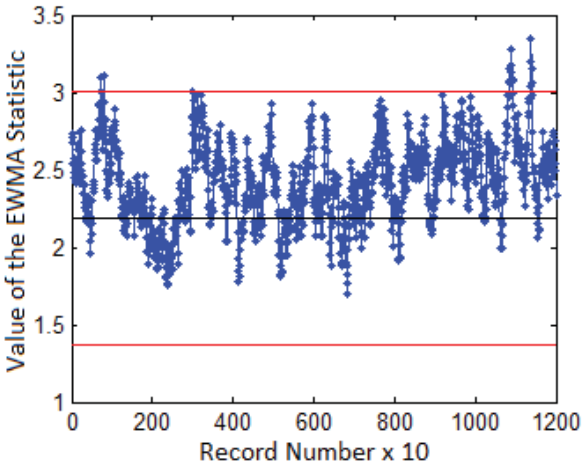

Figure 3. EWMA plot obtained using test data where the labels for the records of two out of the 396 companies (involving around 6% of all test records) were exchanged. The classifier ensemble, center line, LCL and UCL are the same as those in Figure 2.

Figure 3 shows that the EWMA statistic crosses the UCL early on, signaling that the process has drifted significantly. Figure 3 also shows a very clear upward shift in the level of the EWMA, with a large majority of the points in the plot falling above the centerline, providing more evidence that a significantly larger than expected number of misclassifications per set of ten records is occurring.

In practice, predictions by the optimal ensemble would be stopped immediately after the UCL has been crossed, since this is an indication that the process has drifted. At that point, an analyst would determine if a cause for the signal can be found (erroneous record keeping, for example), or if this behavior represents the new state of the commercial network. Only if the latter is true, a new ensemble of classifiers would need to be trained under the new conditions, from which new LCL and UCL values would be calculated.

4. Conclusions and Future Work

We have presented an approach for modeling the performance of a classification ensemble and used a measure of that performance to detect process drift. The example presented involves application of a classification ensemble to a set of commercial records involving a group of companies in a common line of business. The case considered is challenging because of the relatively large number of records involved, the variety and coarseness of the predictors and the relative lack of information available for some of the companies in the network.

Performance of four different classifiers, alone and in combination, was investigated and it was found that, in this case, the most complex classification ensemble is not optimal. Several choices of classifiers, including use of some single classifiers, produce optimal or nearly optimal predictions. This is an indication that the combinatorial space available when multiple classifiers are used should be explored in a systematic way, and that practical considerations, such as the time needed to train and evaluate different ensemble combinations, should be considered as part of the overall ensemble design strategy.

The approach presented to evaluate classification performance involves monitoring a meaningful measure of the misclassification rate, in this case errors in every ten new consecutive records in the data stream. We have shown that the use of an exponentially weighted moving average statistic measuring this proportion of misclassifications is an effective and relatively simple way to detect when significant changes in the behavior of the data stream have occurred. In the example presented, the EWMA is able to detect when a change impacting fewer than 10% of the records in the test set has occurred, suggesting this as a promising tool for detecting concept drift, minimizing the number of interruptions and effort involved in re-training a new classification ensemble.

In the near future, we plan to apply this methodology to data streams from other technical and business areas with the goal of developing a general approach for detecting concept drift in the context of small classification ensembles.

Acknowledgment

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

5. References

[1] Ruta, D. and Gabrys, B. (2005). Classifier selection for majority voting. Information fusion 6(1), pp. 63-81.

[2] Wang, H., Fan, W., Yu, P.S., Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. Proceedings of the ninth ACM SIGKDD internationalconference on Knowledge discovery and data mining. pp. 226-235.

[3] Kolter, J. Zico, and Marcus A. Maloof (2007). Dynamic weighted majority: An ensemble method for drifting concepts. The Journal of Machine Learning Research 8, pp. 2755-2790.

[4] PIERS Trade Database (2014). https://www.piers.com/ JOC Group Inc. Newark, NJ.

[5] Pagell, R.A., and Halperin, M (1998). International business information: how to find it, how to use it. Greenwood Publishing Group.

[6] Kennedy, A. (1994). International Business Information Sources and Their Utilization in Export/Import Research. Journal of Teaching in International Business, Vol. 6(2), 83-101.

[7] MatLab.R2014a (2014). The Mathworks Inc., Natick, MA.

[8] Myers, R.H., Montgomery, D.C., Vining, G.G. (2002). Generalized Linear Models. John Wiley & Sons, Inc. New York, NY.

[9] Tulyakov, S., Jaeger, S., Govindaraju, V. and Doermann. (2008). Review of Classifier Combination Methods in Machine Learning in Document Analysis and Recognition. Springer Berlin Heidelberg. pp. 361-386.

[10] Montgomery, D.C. (1991). Introduction to Statistical Quality Control, 2nd Ed. John Wiley & Sons, Inc. New York, NY.

[11] Testik, M. and Borror, C. (2004). Design strategies for the multivariate exponentially weighted moving average control chart. Quality and Reliability Engineering International 20(6), pp. 571-577.

[12] Qin, Qin et al. (2014). Application of EWMA and CUSUM Models to School Absenteeism Surveillance for Detecting Infectious Disease Outbreaks in Rural China. Online Journal of Public Health Informatics 6(1):e14.