Automated Selection of Training Data and Base Models for Data Stream Mining Using Naïve Bayes Ensemble Classification

Patricia E.N. Lutu

Abstract—**A data stream is a continuous, time-ordered sequence of data items. Data stream mining is the process of applying data mining methods to a data stream in real-time in order to create descriptive or predictive models for the process that generates the data stream. The characteristics of a data stream typically change with time. These changes give rise to the need to continuously make revisions to, or completely rebuild predictive models when the changes are detected. Rebuilding and changing the predictive models needs to be a fast process because stream data may arrive at a high speed. Manual labeling of training data before creating new models may not be able to cope with the speed at which model revision needs to be performed. This paper presents experimental results for the performance of methods for the automated selection of high quality training data for predictive model revision for predictive modeling using Naïve Bayes ensemble classification. The experimental results demonstrate that the use of two base model combination algorithms for the ensembles, results in a high level of confidence in the predictions. Secondly, the periodic revision of the ensemble using the base models created from the automatically selected training data produces revised ensemble models with high predictive performance.**

***Index Terms*—data mining, data stream mining, Naïve Bayes classification, ensemble classification, instance labeling**

I. Introduction

A data stream is a continuous, time-ordered sequence of data items [1]. Data stream mining is the process of applying data mining methods to a data stream in real-time in order to create descriptive or predictive models for the process that generates the data stream [1], [2], [3]. The characteristics of a data stream typically change with time. The changes are categorised as distribution changes and concept changes [2], [4]. These changes give rise to the need to continuously make revisions to, or completely rebuild predictive models when the changes are detected. Rebuilding and changing the predictive models needs to be a fast process because stream data may arrive at a high speed. Traditional methods of manually labeling training data before creating new models may not be able to cope with the speed at which model revision needs to be performed.

Lutu [5] has proposed a predictive modeling framework for automated instance labeling, in stream mining. The purpose of this paper is to report on extensions of the work reported by Lutu [5]. Experimental studies are reported on methods for the periodic revision of classification ensemble models in order to ensure that the stream instances that are classified by ensemble models are assigned class labels with a high level of confidence. This is necessary in order to ensure that high quality automatically labeled and selected training data is used for the creation of new base models which can be used for ensemble revision. The experimental results demonstrate that the use of two base model combination algorithms for the ensembles, results in a high level of confidence in the predictions. Secondly, the periodic revision of the ensemble using the base models created from the automatically labeled and selected training data produces revised ensemble models with high predictive performance. The rest of the paper is organised as follows: Section II provides background to the reported research. Section III presents the proposed ensemble model revision approach. Sections IV and V respectively provide the experimental methods and experiment results. Section VI concludes the paper.

II. Background

A. *Modeling activities for predictive data stream mining*

Predictive data stream mining involves the creation of classification or regression models. For classification modeling, training data with class labels is used to create the model. A real-life data stream does not have a finite size or static behaviour. This means that a data stream cannot be stored in its entirety, and the instance classes cannot be correctly predicted by a model that is created as a once-off activity. This makes it necessary to continuously or periodically revise the predictive model using training data that is reasonably recent.

Two major approaches that are used for predictive data stream mining are the 'sliding window' approach and the ensemble approach. An ensemble model consists of several base models whose predictions are combined into one final prediction using a combination algorithm [6]. The initial modeling activities for either approach involve selecting training data instances, selecting relevant predictive features, creating the initial model, and testing the model. Thereafter, the model is used to provide predictions for data stream instances as they arrive, and the predictive performance is monitored to determine when the model should be revised or

Manuscript received February 03, 2017; revised March 06, 2017.

Patricia. E. N. Lutu is a Senior Lecturer in the Department of Computer Science, University of Pretoria, Pretoria 0002, Republic of South Africa, phone: +27124204116; fax:+27123625188; e-mail: Patricia.Lutu@up.ac.za; web: http://www.cs.up.ac.za/~plutu;

replaced in its entirety. Data streams typically exhibit the phenomena of concept change and distribution change. Concept change refers to changes in the description of the class variable values and may be classified as concept drift which is a gradual change of the concept, or concept shift which is a more abrupt change of the concept [1], [2], [3], [7]. Distribution change, on the other hand, refers to changes in the data distribution of the data stream. Concept and distribution change strategies typically involve continuous monitoring of model performance and data characteristics [7]. For predictive data stream mining, concept drift and distribution changes are typically handled through continuous or periodic revision of the predictive model. Concept shift (sudden concept change) is handled through complete replacement of the current model.

A parallel activity to model usage and monitoring is the labeling of new training data which is then used to test the model performance, and to create new models for model revision or model change. Masud et. al [3], Zhu et. al [8] and Zhang et. al [9] have all observed that, for predictive stream mining, manual labeling of data is a costly and time consuming exercise. In practice it is not possible to manually label all stream data, especially when the data arrives at a high speed. It is therefore common practice to label only a small fraction of the data for training and testing purposes. Masud et. al [3] have proposed the use of ensemble classification models based on semi-supervised clustering, so that only a small fraction (5%) of the data needs to be labeled for the clustering algorithm. Zhu et. al [8] have proposed an active learning framework for solving the instance labeling problem. Active learning aims to identify the most informative instances that should be manually labeled in order to achieve the highest accuracy.

B. Ensemble models for stream mining

Several ensemble classification methods for stream mining have been reported in the literature. Examples of ensemble frameworks that have been reported in the literature are the streaming ensemble algorithm (SEA) [10], the accuracy-weighted ensemble (AWE) [11], and the dynamically weighted majority (DWM) ensemble [12]. Lutu [5] has proposed an ensemble framework for stream mining based on Naïve Bayes ensemble classification [13],[14]. The framework consists of an online component and an offline component. The online component uses Naïve Bayes ensemble base models to make predictions for newly arrived data stream instances. The offline component consists of algorithms to combine base model predictions, determine the reliability of the ensemble predictions, select training data for new base models, create new base models, and determine whether the current online base models need to be replaced. Two objectives of this framework are (1) to remove the need for manual labeling of training instances, and (2) to determine when model revision should be performed.

C. Naive Bayes ensemble classification

Naïve Bayes classification has been reported in the literature as one of the 'ideal' algorithms for stream mining, due to its incremental nature [15]. The Naïve Bayes classifier assigns posterior class probabilities for the query instance $\boldsymbol{x}$ based on Bayes theorem. The training dataset for a classifier is characterised by d predictor variables $X_1,...,X_d$ and a class variable C. The training dataset consists of n training instances, and each instance may be denoted as $(\boldsymbol{x},c_j)$ where $\boldsymbol{x}=(x_1,...,x_d)$ are the values of the predictor variables, and $c_j \in \{c_1,...,c_K\}$ is the class label. Given a new query instance $\boldsymbol{x}=(x_1,...,x_d)$ Naïve Bayes classification involves the computation of the probability $Pr(c_j \mid \boldsymbol{x})$ of the instance belonging to each class c_j as [13], [14]

$$Pr(c_j \mid \boldsymbol{x}) = \frac{P_r(c_j)P_r(\boldsymbol{x} \mid c_j)}{P_r(\boldsymbol{x})} \qquad (1)$$

where $Pr(\boldsymbol{x} \mid c_j) = \prod Pr(x_i \mid c_j)$

and $Pr(\boldsymbol{x}) = \sum_{j=1}^{K} P_r(c_j)P_r(\boldsymbol{x} \mid c_j)$

The values of continuous-valued variables are typically converted into categories through the process of discretisation. The quantities $Pr(c_j)$ for the classes, and $Pr(x_i \mid c_j)$, for the predictor variables, are then estimated from the training data. For zero-one loss classification, the class c_j with the highest probability $Pr(c_j \mid \boldsymbol{x})$ is selected as the predicted class for instance $\boldsymbol{x}$.

Feature selection involves the identification of features (predictor variables) that are relevant and not redundant for a given prediction task [16]. Liu and Motoda [16], Kohavi [17], John et al. [18], and Lutu [19] have discussed the merits of conducting feature selection for Naïve Bayes classification. One straight-forward method of feature selection, is the use the Symmetrical Uncertainty (SU) coefficient as discussed by Lutu [19]. The SU coefficient for a predictor variable X and the class variable C is defined in terms of the entropy function as [16], [20]

$$SU = 2.0(E(X)+E(C)-E(X,C)/(E(X)+E(C)). \qquad (2)$$

where $E(X) = -\sum_{i=1}^{I} Pr(x_i) \, log_2 \, Pr(x_i)$

and $E(C) = -\sum_{j=1}^{J} Pr(c_j) \, log_2 \, Pr(c_j)$

are respectively the entropy for X and C and

$$E(X,C) = -\sum_{i=1}^{I} \sum_{j=1}^{J} Pr(x_i,c_j) \, log_2 \, Pr(x_i,c_j).$$

is the joint entropy for X and C [16], [20]. The SU coefficient takes on values in the interval [0,1] and has the same interpretation as Pearson's product moment correlation coefficient for quantitative variables [16]. White and Liu [20], and Lutu [19] have observed that the entropy functions and the joint entropy function in (2) can be computed from a contingency table.

A 2-dimensional contingency table is a cross-tabulation which gives the frequencies of co-occurrence of the values of two categorical variables X and Y. For Naïve Bayes classification and feature selection, X is the feature and the second variable is C, which is the class variable. Various statistical measures can be derived from a contingency table. The main reason why Naïve Bayes (NB) classification was selected for the framework reported by Lutu [5] is because model creation is a simple and fast activity. For each predictive feature, a single contingency table of counts for the feature values and class labels provides all the data needed for the computation of the quantities for the terms in (1) and (2). Additionally, the class entropy measure for making decisions on base model selection, as discussed in Section III, can be easily computed from the contingency tables.

III. Proposed Approach to Ensemble Revision

A. Automated instance labeling and selection for Naïve Bayes classification

The framework proposed by Lutu [5] uses three measures for assessing the performance of the ensemble base models. These measures are: *Certainty, Reliability*, and *Incoherence. Certainty* measures the frequency that all base models in the ensemble have predicted the same class. *Reliability* measures the frequency that one class is predicted by the majority of base models in the ensemble. *Incoherence* measures the frequency that each base model in the ensemble has predicted a different class from the other base models. The studies reported in this paper are an extension of the studies reported by Lutu [5]. The extensions involve refinements to the prediction categories and the criteria for making decisions on ensemble model revision. For the studies reported by Lutu [5], the ensemble predictions were based on the majority vote by the base models. A major refinement is to use two methods to determine the class label for a query instance. Given a query instance $\boldsymbol{x}$, each base model of the ensemble provides a probabilistic score for each class. The scores provided by the base models are typically stored in a decision matrix with one row for each base model and one column for each class [6]. Various combination methods are available for determining the ensemble prediction. For the studies reported in this paper, the *mean score* for each class is computed and the class selected based on the score is the one with the highest *mean score* value. Additionally, each base model provides a prediction based on the class scores. This is the class with the highest score. The class with the *majority vote* (by the base models) is also used to determine the prediction category. A prediction is treated as *valid* if the class selected by the *mean score* value is the same as the one selected by the *majority vote*.

The information provided by the score-based predictions and by the majority vote predictions, for the *valid* predictions, is used as a basis for determining the prediction categories. Table I provides a summary of the revised prediction categories, which are: *Certain*, *Reliable*, *WeaklyReliable*, and *NotReliabl*e. Only those predictions that are *valid* are assigned the categories *Certain, Reliable* or *WeaklyReliable*. For the proposed automated instance labeling approach, the instances that are assigned the categories *Certain* and *Reliable* are selected as the training data for building NB base models that can be used to replace the current base models when the need arises. The contingency tables for the new base model can be incrementally created off-line as the training data is being generated. At the end of a time period two major activities are conducted. The first activity is to determine if the newly created base model has any predictive value. This assessment is based on the class entropy of the training data used to create the contingency tables for the model, and the number of selected (relevant) features. The second activity is to assess the predictive performance of the current ensemble on the data for the time period, and then make a decision on whether the new base model should replace one of the current base models.

TABLE I
Prediction categories

Mean score range for predicted class	Vote for predicted class	category
0.8 to 1.0 (score >= 0.8) and (score <= 1.0)	3 out of 3	Certain
	2 out of 3	Certain
	1 out of 3	NotReliable
0.6 to 0.8 (score >= 0.6) and (score < 0.8)	3 out of 3	Reliable
	2 out of 3	Reliable
	1 out of 3	NotReliable
0.5 to 0.6 (score >= 0.5) and (score < 0.6)	3 out of 3	WeaklyReliable
	2 out of 3	WeaklyReliable
	1 out of 3	NotReliable
0.0 to 0.5 (score >= 0.0) and (score < 0.5)	3 out of 3	NotReliable
	2 out of 3	NotReliable
	1 out of 3	NotReliable

B. Base model selection for ensemble revision

Various measures of ensemble performance are defined in terms of the instance counts for the prediction categories, and the total number of predictions for a given time period. The measures are defined as follows:

Measures of ensemble performance:

Surrogate Accuracy = (count (Certain) + count(Reliable)) / (TotalPredictions)

Certainty = count (Certain) / TotalPredictions

Reliability = count (Reliable) / TotalPredictions

WeakReliability = count (Weakly Reliable) / TotalPredictions

NonReliability = count (Not Reliable) / TotalPredictions

Agreement = count(predictions where base modelprediction is the same as ensemble prediction)

The proposed approach handles distribution changes, concept drift and sudden concept change proactively. Additionally, the approach aims to ensure that frequent model revisions result in high confidence predictions. A major time period T (e.g. after 30,000 instances are received) is used for making decisions about ensemble model revision to handle possible distribution changes and concept drift. Additionally, a minor time period t (e.g. after 3,000 instances) is used to monitor the possibility of sudden concept change. Rule 1 is used at the end of each minor time period, to check for sudden concept change.

```
Rule 1:
If  (Certainty < specified value) AND
    (Surrogate Accuracy < specified value)  AND
    (WeakReliability > specified value) AND
    (Agreement for at least one base model < specified value)
then flag sudden concept change
```

When sudden concept change is detected, then the current ensemble should be abandoned, and a new ensemble should be created using manually labeled data. Ensemble performance is assessed at the end of each major time period T_i, to determine whether the most recently created base model should replace one of the current base models. If the new base model has zero class entropy (i.e. all training instances are of the same class) then no replacement decision is made, otherwise, Rule 2 is used to determine the need for base model replacement.

```
Rule 2:
If  (class entropy for the new base mode is greater than 0) then
    If   (base model has the lowest Agreement)   AND
         (base model has lowest no. of selected features)  AND
         (base model has lower class entropy than new base model)
         AND  (base model has fewer or equal number of selected
         features as new base model)
then replace base model with the new base model.
```

IV. Experimental Methods

The KDD Cup 1999 dataset available from the UCI KDD archive [21] was used for the studies. The dataset consists of a training dataset and a test dataset. The 10% version of the training dataset was used for the experiments. This dataset consists of 494,021 instances, 41 features and 23 classes. The 23 classes may be grouped into five categories: NORMAL, DOS, PROBE, R2L and U2R which can then be used as the classes [22]. A new feature (called ID) was added to the dataset with values in the range [1, 494021] as a pseudo timestamp. Fig. 1 shows a plot of the class distribution for this data stream. The data stream exhibits an extreme imbalance of the class distribution over time. It is clear from Fig. 1 that the classes DOS and NORMAL are the majority classes.

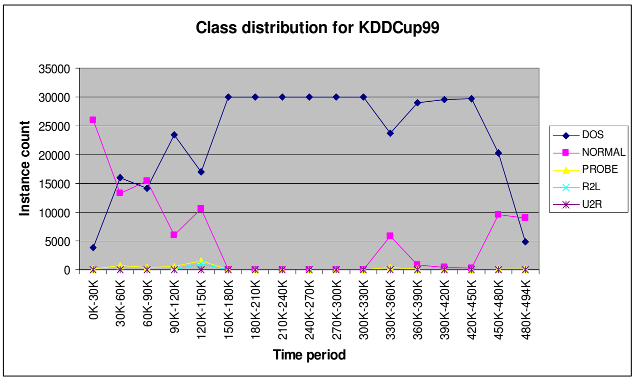

Fig. 1. Class distribution for the KDD Cup 1999 data stream (adopted from [5])

The algorithms for feature selection, Naïve Bayes classification, and base model prediction combination, were implemented as a GNU C++ application. Details of the required data structures have been presented by Lutu [19]. The decisions on concept change detection and model revision were made through manual inspection of the model performance measures followed by the application of Rule 1 or Rule 2 presented in Section III. These decisions can easily be automated in C++ program code.

V. Experiments for Stream Mining

A. *Objectives of the experiments*

Exploratory experiments were conducted to answer the following questions: (1) Do the *Certain* and *Reliable* categories lead to the selection of a high percentage of correctly labeled training data? (2) Are the *Certainty* and *Reliability* measures good estimators of predictive accuracy? (3) Do the proposed methods for ensemble model revisions result in improvements in *Certain* and *Reliable* predictions? (4) Are the *Certainty, Reliability*, *WeakReliability*, *NonReliability* measures useful for forecasting possible concept change?

B. *Creation of the initial base models*

The top 90,000 instances of the KDD Cup 1999 data stream were used for the creation of the three base models MW1, MW2 and MW3 for the initial ensemble. The instances were divided into three batches of training instances for each Naïve Bayes base model. The base models were tested on the data for the time period W4 (90K-120K). Table II shows the properties of the training data, and the test results for the initial ensemble base models.

TABLE II
PROPERTIES OF THE TRAINING DATA FOR THE INITIAL ENSEMBLE BASE MODELS

Model name	Time period	Training set size	Class Entropy	Relevant features	Testing accuracy on W4 data
MW1	0K-30K	30,000	0.604	12	60.0
MW2	30K-60K	30,000	1.139	17	97.9
MW3	60K-90K	30,000	1.099	14	61.2

C. *Experimental results*

The initial ensemble model {MW1, MW2, MW3} was used to provide predictions for the data stream instances starting at time period W4 (90K-120K). For each prediction period W_i, the instances that were assigned the *Certain* or *Reliable* category were selected as training data and the base model MW_i was created from this training data. Additionally, a decision was made whether or not to replace one of the current base models, using the logic of Rule 2. Tables III and IV show the description of the ensemble models and the ensemble performance, for the time periods W4,…,W17 (90K to 494K). For the time periods W6 to W11, the base models are exactly the same. This is because all the instances for these time periods belong to the DOS class. The entries in column 3 of Table III show that base model replacements were done at the beginning of time periods W5, W6 and W13. The results of columns 3, 4 and 5 of Table IV show that, for the periods W5 to W15, the values for *Certainty*, *Surrogate Accuracy* and 'real' *Accuracy* are consistently high. Additionally, the *Surrogate*

Accuracy values are generally very close to the (real) *Accuracy* values. For the period W16, the values for *Certainty, Surrogate Accuracy* and 'real' *Accuracy* suddenly plummet to much lower levels. This is an indication of sudden concept change. Based on the above observations, the answer to the question: '*Are the Certainty and Reliability measures good estimators of predictive accuracy*?' is 'yes'. The reader will recall that *Surrogate Accuracy* is defined in terms of *Certainty* and *Reliability*.

TABLE III
DESCRIPTION OF THE CONTINUOUSLY REVISED ENSEMBLE

Prediction Window	Prediction time period	Ensemble	
		base models	name
W4	90K-120K	MW1, MW2, MW3	E1
W5	120K-150K	MW2, MW3, MW4	E2
W6, ..,W11	150K-330K	MW2, MW4, MW5	E3
W12	330K-360K	MW2, MW4, MW5	E3
W13	360K-390K	MW2, MW5, MW12	E4
W14	390K-420K	MW2, MW5, MW12	E4
W15	420K-450K	MW2, MW5, MW12	E4
W16	450K-480K	MW2, MW5, MW12	E4
W17	480K-494K	MW2, MW5, MW12	E4

TABLE IV
PREDICTIVE PERFORMANCE OF THE CONTINUOUSLY REVISED ENSEMBLE

Prediction window	ensemble name	Certainty %	Surrogate Accuracy %	'real' Accuracy %
W4	E1	19.7	99.4	99.4
W5	E2	62.1	97.2	90.4
W6,..,W11	E3	100	100	100
W12	E3	91.2	98.0	95.5
W13	E4	97.9	98.8	97.7
W14	E4	99.9	99.9	99.9
W15	E4	99.9	99.9	99.9
W16	E4	22.1	68.1	22.3
W17	E4	87.5	95.2	86.7

Table V shows the values of the performance measures for the time periods W5 to W17. The values for the *Certain* and *Reliable* categories for the periods W5 to W15 (120K to 450K clearly indicate that the class labels assigned to instances in these categories are generally correct (can be trusted). This implies that the automatically labeled and selected training data for the ensemble is of good quality.

TABLE V
ASSESSMENT OF ENSEMBLE PREDICTION CATEGORIES

Prediction period name	Certain		Reliable	
	Certainty %	Correct%	Relia-bility%	Correct %
W5	62.1	56.5	35.0	33.5
W6,..,W11	100	100	0	-
W12	91.2	90.7	6.9	4.6
W13	97.9	97.6	0.9	0.13
W14	99.9	99.9	0.05	0.04
W15	99.9	99.9	0.03	0.01
W16	22.1	21.9	46.1	0.4
W17	87.5	86.0	7.8	0.5

However, the values for the period W16 (450K-480K) indicate that the assigned class labels for the *Reliable* category cannot be trusted as they have a very low level of 'correctness' and so, the instances for this time period should not be selected as training data. This situation is detected as concept change which is discussed below. So, do the *Certain* and *Reliable* categories lead to the selection of a high percentage of correctly labeled training data? Based on the foregoing observations, the answer to this question is 'yes'. The percentage of correctly labeled training data is in the region of 95%, which implies a noise level in the region of 5%.

Table VI shows the properties and prediction behaviour of all the base models that were used in the periodically revised ensemble. It can be deduced from Table III that at the end of W4, the decision was made to replace MW1 with MW4, to obtain the ensemble {MW2, MW3, MW4}. Based on the values in Table VI and Rule 2 of Section III, the reason for the replacement is because MW1 has the smallest number of selected features and lowest *Agreement* level. Additionally, MW1 has lower class entropy than MW4, and fewer selected features than MW4. At the end of W5, the decision was made to replace MW3 with MW5, to obtain the ensemble {MW2, MW4, MW5}. At the end of W12, the decision was made to replace MW4 with MW12, to obtain the ensemble {MW2, MW5, MW12}. The reasons for these replacements can be easily deduced from the values in Table VI and Rule 2.

TABLE VI
PROPERTIES OF THE BASE MODELS USED IN THE PERIODICALLY REVISED ENSEMBLE

Model Name	Time period for training data	Training set size	Class Entropy	Relevant features	Agreement% with ensemble prediction in time period:		
					W4	W5	W12
MW1	0K-30K	30,000	0.604	12	59.7	-	-
MW2	30K-60K	30,000	1.139	17	98.4	97.0	91.7
MW3	60K-90K	30,000	1.099	14	61.5	72.7	-
MW4	90K-120K	29,815	0.831	22	-	91.7	94.1
MW5	120K-150K	29,147	1.214	41	-	-	96.4
MW12	330K-360K	29,403	0.844	22	-	-	-

Table VII provides a comparison of the initial ensemble and the periodically revised ensemble in terms of *Certainty* and *Reliability.* The results of columns 2 and 4 indicate that *Certainty%* is much higher for the periodically revised ensemble, starting with the first revision in W5 (120K-150K). Based on this observation, the answer to the question: '*Do the proposed methods for ensemble model revisions result in improvements in certain and reliable predictions*?', is 'yes'.

Additional analysis was conducted on the prediction results of the ensemble {MW2, MW5, MW12} for the W16 (450K-480K) prediction period. Tables VIII and IX show the details of the incremental performance of the ensemble for minor time periods consisting of 3,000 instances each. Results are shown for the first six minor time periods. The results show that (1) *Certainty, Surrogate Accuracy* and 'real' *Accuracy* decrease steadily until concept change occurs in the minor time period 459K-462K, (2) *WeakReliability* suddenly increases when concept change occurs, and (3) *Agreement* for one of the base models (MW2) suddenly decreases significantly when concept change occurs. So, to answer the question: "*Are the Certainty, Reliability, WeakReliability, NonReliability measures good at forecasting possible concept change*?", the answer is "yes". The above observations support the claim that the measures are good indicators of concept change.

TABLE VII
COMPARISON OF INITIAL AND PERIODICALLY REVISED ENSEMBLES

Prediction period	Initial ensemble performance		Continuously revised ensemble performance	
	Certainty %	Reliability %	Certainty %	Reliability %
W4	19.7	79.7	19.7	79.7
W5	30.9	65.2	62.1	35.0
W6,..,W11	0	100	100	0
W12	11.7	85.7	91.2	6.9
W13	2.1	96.0	97.9	0.9
W14	1.3	98.6	99.9	0.05
W15	0.6	99.3	99.9	0.03
W16	18.8	73.7	22.1	46.06
W17	55.3	39.6	87.5	7.77

TABLE VIII
INCREMENTAL PERFORMANCE FOR W16 PREDICTIONS: CERTAINTY AND RELIABILITY

Prediction time period	Certainty %	Reliability%	Weakly Reliabiliy%	Non Reliability%
450K-453K	69.2	21.0	2.0	7.9
453K-456K	58.4	25.1	1.4	15.1
456K-459K	64.47	23.9	2.4	9.2
459K-462K	**28.5**	41.9	**27.8**	1.9
462K-465K	0.0	56.6	41.5	1.9
465K-468K	0.0	60.2	37.7	2.0

TABLE IX
INCREMENTAL PERFORMANCE FOR W16 PREDICTIONS: ACCURACY AND AGREEMENT

Prediction time period	Surrogate Accuracy %	Accuracy %	Agreement%		
			MW2	MW5	MW12
450K-453K	90.1	70.53	90.5	71.8	90.8
453K-456K	83.5	58.47	83.5	60.2	84.6
456K-459K	88.40	65.37	89.4	67.5	89.5
459K-462K	**70.4**	28.6	**28.5**	98.1	98.1
462K-465K	56.6	0.0	0.0	98.1	98.1
465K-468K	60.2	0.0	0.0	98.0	98.0

VI. CONCLUSIONS

The objectives of the research reported in this paper were to assess the usefulness of the proposed methods for automated selection of high quality training data for Naïve Bayes base model creation and ensemble revision for predictive data stream mining. The experimental results reported in Section V have demonstrated that , for the KDD Cup 1999 dataset, the use of two base model prediction combination algorithms for the ensembles, results in a high level of confidence in the class labels for the automatically selected training data. Secondly, the periodic revision of the ensemble using the base models created from the selected training data produces revised ensemble models with high predictive performance. Thirdly, the proposed measure of *Surrogate Accuracy* provides meaningful information about ensemble model predictive performance.

REFERENCES

[1] C.C. Aggarwal, *Data Streams: Models and Algorithms*, Boston: Kluwer Academic Publishers, 2007.

[2] J. Gao, W. Fan and J. Han, "On appropriate assumptions to mine data streams: analysis and practice", in *Proc.7th IEEE Int. Conf. on Data Mining,* IEEE Computer Society, 2007, pp.143-152.

[3] M.M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Classification and novel class detection of data streams in a dynamic feature space", in *Proc. ECML PKDD 2010,* LNAI, Springer-Verlag, 2010, pp. 337-352.

[4] J. Gao, W. Fan, J. Han and P.S. Yu, "A general framework for mining concept-drifting data streams with skewed distributions", in *Proc. SDM Conference,* 2007.

[5] P.E.N. Lutu, "Naïve Bayes classification ensembles to support modeling decisions in data stream mining", in *Proc. IEEE SSCI 2015*, Cape Town, South Africa, pp. 335-340.

[6] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Hoboken, New Jersey: John Wiley & sons 2004.

[7] A. Bifet, G. Holmes, R. Kirkby and B. Pfahringer, *Data Stream Mining: a Practical Approach*, University of Waikato, 2011. Available:http://jwijffels.github.io/RMOA/MOA_2014_04/doc/pdf/StreamMining.pdf

[8] X. Zhu, P. Zhang, X. Lin and Y. Shi, "Active learning from data streams", in *Proc. 7th IEEE Int. Conf. on Data Mining*, 2007, pp. 757-762.

[9] P. Zhang, X. Zhu and L. Guo, "Mining data streams with labeled and unlabeled training examples", in *Proc. 9th IEEE International Conference on Data Mining,* 2009, pp. 627-636.

[10] W.N. Street and Y. Kim,"A streaming ensemble algorithm (SEA) for large-scale classification", in *Proc. 7th ACM Int. Conf. on Knowledge Discovery and Data Mining*,ACM Press,New York,2001,pp.377-382.

[11] H. Wang, W. Fan, P.S. Yu and J. Han, "Mining concept-drifting data streams using ensemble classifiers", in *Proc. ACM SIGKDD*, Washington DC, 2003, pp.226-235.

[12] J.Z. Kolter and M.A. Maloof, "Dynamic weighted majority: an ensemble method for drifting concepts", *Journal of Machine Learning Research,* vol. 8, pp. 2755-2790, 2007.

[13] T. M. Mitchell, *Machine Learning*, Burr Ridge, IL:WCB/McGraw-Hill, 1997.

[14] P. Giudici, *Applied Data Mining: Statistical Methods for Business and Industry*, Chichester: John Wiley and Sons, 2003.

[15] R. Munro and S. Chawla, *An Integrated Approach to Mining Data Streams*, Technical Report TR-548, School of Information Technologies, University of Sydney, 2004. Available: http://www.it.usyd.edu.au/research/tr/tr548.pdf

[16] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Boston:Kluwer Academic Publishers, 1998.

[17] R. Kohavi, "Scaling up the accuracy of Naïve Bayes classifiers: a decision tree hybrid", in *Proc. KDD 1996*, pp 202-207.

[18] G.H. John, R. Kohavi and K. Pleger, "Irrelevant features and the subset selection problem", in Proc. *11th Int. Conf. on Machine Learning*, 1994, pp 121-129.

[19] P.E.N. Lutu, "Fast feature selection for Naïve Bayes classification in data stream mining", in *Proc. World Congress on Engineering*, London, U.K., July 2013, vol. III, pp. 1549-1554.

[20] A.P. White and W.Z. Liu, "Bias in information-based measures in decision tree induction", *Machine Learning*, vol. 15, pp. 321-329. Boston : Kluwer Academic Publications, 1994.

[21] S.D. Bay, D. Kibler, M.J. Pazzani and P. Smyth, "The UCI KDD archive of large data sets for data mining research and experimentation", *ACM SIGKDD*, vol. 2, no. 2, pp. 81-85, 2000.

[22] S. W. Shin and C. H. Lee, "Using attack-specific feature subsets for network intrusion detection", in *Proc. 19th Australian conference on Artificial Intelligence,* Hobart, Australia, 2006.