Analysis of the Evolution of Features in Classification Problems with Concept Drift: Application to Spam Detection

Márcia Henke, Eduardo Souto, Eulanda M. dos Santos
Institute of Computing, Federal University of Amazonas – Manaus, Brazil
{henke,esouto, emsantos}@icomp.ufam.edu.br

***Abstract*— Machine Learning solutions for concept drift detection problems try to decide to what extent a particular set of examples still represents the current concept rather than treating all data equally. Monitoring the set of relevant features used to generate the classification model may be an effective strategy for concept drift detection. This paper focuses on analyzing the possibility of detecting drifts through feature evolution monitoring in the spam detection problem. Results of the experiments show that the relevant features of the target domain are significantly different from the relevant features of the source domain. This offers a new possibility for analyzing the relationship between feature evolution and misclassification rate. The experiments were conducted using two databases: a public database composed of samples collected between 2003 and 2004; and a new private database composed of samples collected between 2012 and 2013.**

Keywords—feature evolution; concept drift; spam.

I. INTRODUCTION

Several real-world applications present learning context (target environment) which changes over time. Weather predictions, fraud detection, customer preferences, environmental monitoring and anti-spam filters are examples of problems whose environments are dynamic. This problem is named as concept drift in the Machine Learning literature.

For anti-spam filters, for instance, most filters employ classification models generated from an initial training period, in which a dataset composed of legitimate messages and also spam messages is learnt and a spam classification model is created based on this knowledge. After this training period, the program moves to monitor new messages and classifies each incoming message as spam or not, based on the initial training. Thus, these solutions use mostly features extracted in a static way, i.e., features representing the problem in an instant of time t. As a result, these solutions cannot adapt their model to deal with changes [1].

According to Harries and Sammut [2], the concept drift may be the consequence of context change, which is directly related to the features. Therefore, changes may be detected by monitoring the set of relevant features used in the construction of the classification model. Given this context, this paper presents a study that analyzes the evolution of the feature space of two spam detection problems to demonstrate the difficulty of generalization of a classifier obtained in a source domain (time *t*) to a target domain (time *t+1*). Such difficulty involves at least two factors: behavior changing in spam attacks, and lack of stability of the strategies of feature selection for ranking the most relevant features. Results of the experiments show that the feature space of the target domain is significantly different from the feature space of the source domain. This offers a new possibility for analyzing the relationship between feature evolution and misclassification rate.

Two databases were used in the experiments: (1) ECUE (Email Classification Using Examples), which is a public database; and (2) ETSS (Emerging Technologies and Security System Group Spam Dataset), which is a new and private database. ETSS database is composed of a sample set almost 10 times larger than the sample set of ECUE database. This fact leads to a more stable feature evolution analysis, as it is explained in Section III.

The paper is organized as follows: section II presents a brief definition, methods and approaches to handling concept drift. The results obtained in two series of experiments are presented in Section III. Finally, Section IV presents conclusions and future work.

II. CONCEPT DRIFT PROBLEM

Informally, concept drift refers to changes in the class (concept) definitions over time [3] or generally speaking, concept drift could be changes in the probabilities of classes or conditional probability distributions of classes [1]. We can distinguish the following sources of changes: class prior probabilities; class-conditional probability distribution; and posterior probabilities.

Formally, given a continuous stream of examples $x_1, x_2, ...$, each example is an m-dimensional vector in a pre-defined vector space R^m. At every time point t, the n most recent examples are put together to form a set X of examples. Given a set $\bar{X}$ containing the $\bar{n}$ examples that appeared prior to those in X, in dynamic environment problems, examples in X are not generated by the same distribution in $\bar{X}$ [4].

As mentioned before, the evolution of relevant features may be used to perform concept drift detection [5]. This work focuses on monitoring the sets of relevant features in order to determine the time a classification model should keep its knowledge effective. Here, relevant features are assumed to be pointed out by a feature selection process, as described in the next sub-section.

A. Feature Selection

Feature selection techniques are often used in the Machine Learning literature aiming to reduce the feature space and to define the most relevant feature subset. However, there does not exist in the literature a formal definition of feature relevance [6][7]. In [6], we can find the

following definition: a f_i feature is strongly relevant if the high performance classification (given all the features) decreases significantly when f_i is removed from the feature space. On the other hand, a feature is considered weak relevant when the contribution of this feature in the classification accuracy is variable or non-existent [7].

Even though feature relevance is not well defined, feature selection can be considered a key issue in dynamic problems, since, with time, some features may become less relevant, while others may acquire relevance. Even though, the feature selection techniques do not show stability in the task of selecting the most relevant features. Salem [8] demonstrated empirically that stability is dependent on the data set. The dimensionality of the feature space and the amount of samples strongly affect the stability of feature selection. The higher the amount of samples, the greater the positive impact, and, the higher the dimensionality of the feature space, the greater the negative impact on the stability of selected relevant features.

In order to select the subset of the most relevant features, we can use filter, wrapper and hybrid methods [8]. The most popularly used method is the Information Gain (IG) filter method, which is employed in this work.

III. EXPERIMENTS

Our experiments analyze the effectiveness of a classification model for spam attacks over time. The goal is to determine the time during which the classification model generated remains efficient when tested on a target domain. In particular, two issues will be investigated: (1) The model performance considering the error rate for time windows after the training period; and (2) the relation between the model error rate and the feature evolution problem.

To deal with these two issues, we will use a public database, ECUE Spam Dataset [9], and a private database, ETSS spam dataset. Two series of experiments were performed. First, the model is built in the training phase and evaluated statically in the test phase. Next, the model is built and is tested in monthly time windows.

A. Datasets

The two datasets investigated are divided into monthly time windows for training/validation and test. The ECUE dataset is composed of three types of features: (1) Word features - represent a sequence of characters separated by White space or html tag delimiters; (2) character features - represent the occurrence of a single character in the email; and (3) structural features - representing the structure of the email, e.g. the proportions of the following characters: white space, punctuation, uppercase, lowercase, and total number of characters in the email. In this work, we investigate the ECUE Concept Drift Dataset, whose samples were obtained between February 2003 and January 2004. It contains near 166.610 features and more than 10,000 samples.

The ETSS dataset is composed of only one type of features: word features, which represent a sequence of characters separated by White space or html tag delimiters. This dataset is composed of samples collected between June 2012 and October 2013. The original feature space of ETSS database is composed of 105.756 features.

The databases are organized as shown in Tables I, II, and III. Table I shows that the training sets were generated using samples from the first three months of the databases. Each training set consists of 1000 samples, 500 spam and 500 legitimate emails for both datasets. Table II shows ECUE test set distribution divided into 12 months. The samples are 87% spam and 13% legitimate emails. Table III shows ETSS test set distribution divided into 14 months. The samples are 80% spam and 20% legitimate email.

TABLE I. SAMPLE DISTRIBUITION OF THE TRAINING DATASETS

Source Domain – ECUE						**Source Domain – ETSS**				
Message	***2002***		***2003***	***N/Date***	**Total**	**Message**	**2012**			**Total**
	Nov	***Dec***	***Jan***				***Jun***	***Jul***	***Aug***	
Spam	96	367	37	0	500	Spam	167	166	167	500
Nonspam	83	84	84	249	500	Nonspam	167	167	166	500
Total	179	451	121	249	1000	Total	334	333	333	1000

TABLE II. SAMPLE DISTRIBUITION OF THE TEST DATASET: ECUE

Target Domain – ECUE													
Message	**2003**											**2004**	***Total***
	Feb	***Mar***	***Apr***	***May***	***Jun***	***Jul***	***Aug***	***Sep***	***Oct***	***Nov***	***Dec***	***Jan***	
Spam	142	391	405	459	406	476	582	1849	1746	1300	954	746	9456
Nonspam	151	56	144	234	128	19	30	182	123	113	99	130	1409
Total	293	447	549	693	534	495	612	2031	1869	1413	1053	876	10865

B. Evaluation Metrics

The performance for spam detection can be calculated based on the following measures: True positive (TP), is the number of spam correctly classified as spam; False Negative (FN), is the number of spam incorrectly classified as legitimate; False Positive (FP), is the number legitimate email falsely classified as spam; and True Negative (TN), the number of legitimate email correctly classified as legitimate. In order to evaluate the period during which a classifier has its knowledge effective, the error rate is used as a metric to evaluate its performance on the source domain and target domain. So, the error rate equation can be obtained by:

$$Error = \frac{FP+FN}{Total\ samples} \qquad (1)$$

C. Classifier

The classification method chosen in this work was SVM. This choice is especially due to the fact that SVM classifier is stable, i.e., small changes in data do not significantly affect the performance of the classifier. This characteristic is important to reduce the possibility of false detection of change in dynamic environment problems.

TABLE III. SAMPLE DISTRIBUITION OF THE TEST DATASET: ETSS

Target Domain – ETSS															
Message	**2012**			***2013***											***Total***
	Sep	***Oct***	***Nov***	***Dec***	***Jan***	***Feb***	***Mar***	***Apr***	***May***	***Jun***	***Jul***	***Aug***	***Sep***	***Oct***	
Spam	4776	5496	4924	4280	10503	4808	6597	4133	8064	3838	11198	8795	7590	5293	90295
Nonspam	1356	1729	1476	1261	1293	1503	1615	1927	2014	1686	1936	1722	1432	1454	22404
Total	6132	7225	6400	5541	11796	6311	8212	6060	10078	5524	13134	10517	9022	6747	112699

Two initial parameters need to be defined for SVM: kernel function and the regularization parameter *C*. Conditional on the kernel function selected, other parameters must be defined, like the gamma value for the RBF (Radial Basis Function). In this work, training set was used to define the parameters through 10-fold cross-validation. The best results were obtained with the kernel, y=0.01 and C = 5 for ECUE dataset. For ETSS dataset, the RBF kernel, with y=0.1 and C=100, achieved the lowest error rates.

D. Features Selection Strategy

In this paper, the feature selection strategy used was IG. For the feature space reduction in ECUE dataset, 700 of the most relevant features were chosen. According to Delany et al. [29], this is the number of features that best describes the problem on ECUE dataset. In terms of the ETSS dataset, our experiments using 10-fold cross-validation showed that 300 of the most relevant features is the ideal number of features to describe this database. The experiments performed in this work are divided into the following two series:

1) Static Training and Test

The first series has been applied in order to reduce the dimensionality of the feature space and to select the most relevant features of the problem, as it is usually done in static problems, i.e., features are selected on the training dataset, then, the test samples are represented by the same feature space as in the training set. SVM was trained and validated on the source domain (training set) using 10-fold cross-validation, and then evaluated on the target domain (test set).

2) Monthly Training and Test

The second series aims to evaluate features dynamically. To achieve this objective, IG is applied in each monthly time window on the target domain. Then, the classifier is trained and tested in each time window using 10-fold cross-validation. Here, the objective is to demonstrate that features found in each time period of the target domain are different from the features of the source domain and are more representative for the problem as time goes on. Moreover, this series of experiments simulates an implicit concept drift detection method, since the system updates its knowledge base with the new pieces of information.

E. Results

We start by evaluating the effectiveness of the classification model through testing. Figure 1 shows the classifier performance considering the error rate obtained from equation 1 on ECUE, Figure 1 (a), and ETSS, Figure 1 (b), for Series1. As it can be seen in Figure 1, the classifier presents an expected behavior on a dynamic environment problem, especially in ECUE database, series 1, i.e., a gradual increase in the error rate as time goes on. However, some behaviors require closer examination. For instance, in the months of February in Figure 1 (a) and September_12 in Figure 1 (b), the months following the training, the classifier has high error rates (3.07% and 6.64% respectively) for series 1. This fact leads to the following question: the samples that make up training set have not representative features for the problem?

To investigate such issue and to determine the period of time in which the classification model generated remains effective, a second series of experiments was carried out, where the classification model is built and tested in a monthly time window over a period of 12 months, for ECUE database, and 14 months for ETSS database. It is noteworthy that the selection of features is also done monthly in these experiments. Figure 1, shows the error rates achieved in this second series of experiments on ECUE (a) and on ETSS (b) databases. Figure 1, also, shows the error rates attained in the first series of experiments to better illustrate the comparison between the results achieved in each series of experiments.

Comparing the error rates of both series, Figure 1 (a) and (b), we can observe that building the classification models based on monthly feature selection, is always more effective than the "static" model used in the first series, since SVM attained lower error rates in Series 2 than in Series 1. In all cases, the error rates did not exceed 2% in Figure 1(a) and 3.6% in Figure (b). The average error in Series 1 was 7.90% and 5.57%, while in Series 2, 0.98% and 2.48%, for ECUE and ETSS databases respectively. In ECUE database, the error rate was 0.34% in the month of February (Series 2), while it was 3.07% in Series 1. For the ETSS database, the error rate was 3.56 in the month of September_12 (Series 2), while it was 6.64% in Series 1. These results allow us to observe that there is an important and considerable feature evolution between the source domain and the target domain.

It is important to note that the variability of error rates in the experiments of series 2 is due to two factors: the deficiency of the feature selection strategy and the unbalanced samples. The first factor is directly related to the high dimensionality of the feature spaces, i.e. 166.610 and 105.756 features for ECUE and ETSS databases respectively.

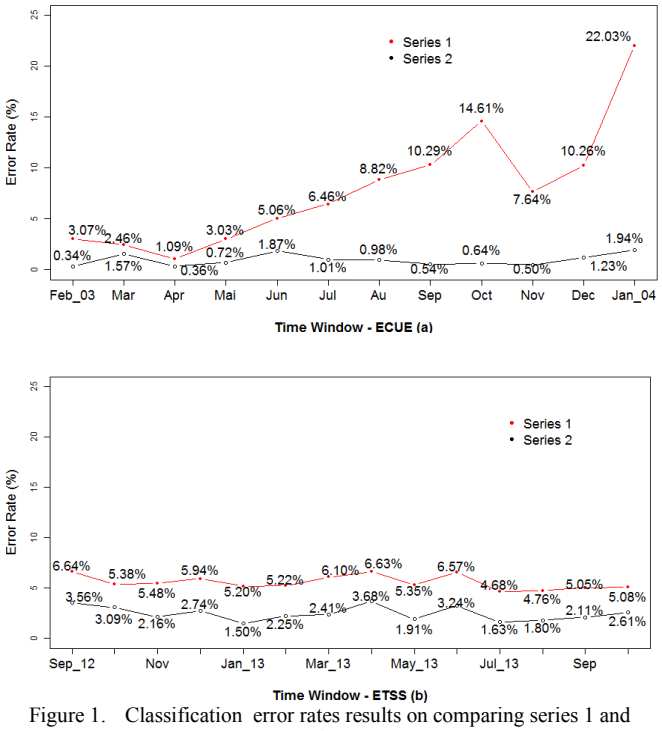

Figure 1. Classification error rates results on comparing series 1 and series 2.

The second factor is observed especially in ECUE database, as can be seen in Table 2. The low representation of nonspam class undermines a balanced distribution of samples, which is a problem for the cross-validation strategy. For example, the July time window has only 19 nonspam samples.

As discussed before, this paper focuses on monitoring the relevant features set to determine if a classification model maintains its effectiveness or not. Thus, two questions arise from these experiments: is there a direct relationship between feature evolution and misclassification rate? Is it possible to use this relationship to determine a threshold and the time for which a classification model remains effective? To investigate this relationship, a comparative analysis was undertaken of the feature vectors generated during the training and testing on the second series of experiments. The purpose of this comparison is to calculate the number of target domain features that are absent in the source domain.

Figure 2 shows the missing features as a percent on the source domain from each monthly time window of the target domain. The results show the first month after the training period (February 2003 and September 2012) with 43.43% and 33.00% of features missing, i.e., 43.43% and 33.00% of the features in target domain are absent in the source domain. This feature evolution has a direct impact on the error rate of the classification model, as can be observed in Figure 1 (a) and (b), where the classification error rates in the first month of testing were 3.07% and 6.64%, in ECUE and ETSS databases respectively. However, this behavior is not linear; April 2003 shows a feature evolution higher (55% features missing), and a lower error rate with 1.09% when compared to February 2003 on ECUE dataset. The same way, August 2013 shows a feature evolution higher (37% features missing), and a lower error rate with 4.76% when compared to the September 2012 month on ETSS dataset.

Therefore, to find the effectiveness of a classification model based on feature evolution, a more detailed analysis is required. Besides percent of the absence of features, in order to use feature evolution as pointers for updating the classification model, other issues should be taken into account such as frequency of new features. Even though, this analysis allow us to conclude that feature evolution monitoring may be an effective strategy for drift detection.

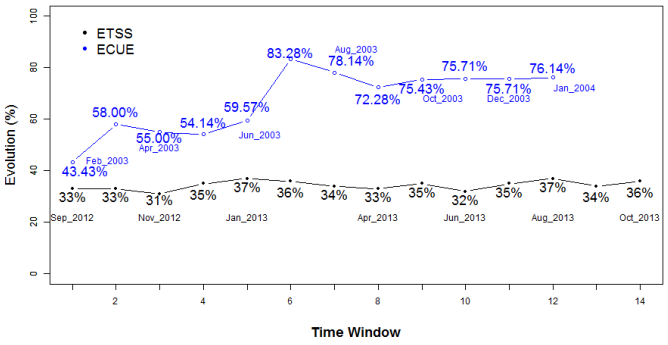

Figure 2. Percent of Target Domain Features Absent on Source Domain.

F. Conclusion

This paper presented a study on the feature evolution for the spam detection problem over time. The analysis showed that the percent values of presence or absence of features may be good indicators to update the classification model. However, other issues, such as stability, frequency and relevance of features on source and target domains must also be observed.

REFERENCES

[1] B. Kurlej and M. Wozniak, "Active learning approach to concept drift problem," *Log. J. IGPL*, vol. 20, no. 3, pp. 550–559, 2011.

[2] M. Kubat, "Extracting Hidden Context," vol. 126, no. 1998, pp. 101–126, 2000.

[3] R. Elwell and R. Polikar, "Incremental learning of concept drift in nonstationary environments.," *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1517–31, Oct. 2011.

[4] A. Dries and U. RÃ¼ckert, "Adaptive concept drift detection," *Stat. Anal. Data Min.*, vol. 2, no. 5â ’6, pp. 311–327, Dec. 2009.

[5] Zhu Qiuming, "Pattern Classification in Dynamic Environments: Tagged Feature-Class Representation and the Classifiers," *IEEE T. Syst. Man. Cybern. B.*, vol. 19, no. 5, pp. 1203–1210, 1989.

[6] D. A. Bell, "A Formalism for Relevance and Its Application in Feature Subset Selection," pp. 175–195, 2000.

[7] L. Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: A survey and experimental evaluation," *Data Mining, 2002. ICDM ...*, 2002.

[8] S. Alelyani, "On Feature Selection Stability: A Data Perspective," no. May, 2013.

[9] "DIT Applied Intelligence Research Centre» Resources." [Online]. Available: http://www.dit.ie/computing/staff/sjdelany/datasets/. [Accessed: 23-Jan-2015].