The Influence of a Classifiers' Diversity on the Quality of Weighted Aging Ensemble

Michał Woźniak[1], Piotr Cal[1], and Bogusław Cyganek[2]

[1] Department of Systems and Computer Networks
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{michal.wozniak,piotr.cal}@pwr.wroc.pl
[2] AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
cyganek@agh.edu.pl

Abstract. The paper deals with the problem of data stream classification. In the previous works we proposed the WAE (*Weighted Aging Ensemble*) algorithm which may change the line-up of the classifier committee dynamically according to coming of new individual classifiers. The ensemble pruning method uses the diversity measure called the *Generalized Diversity* only. In this work we propose the modification of the WAE algorithm which applies the mentioned above pruning criterion by the linear combination of diversity measure and accuracy of the classifier ensemble. The proposed method was evaluated on the basis of computer experiments which were carried out on two benchmark databases. The main objective of the experiments was to answer the question if the chosen modified criterion based on the diversity measure and accuracy is an appropriate choice to prune the classifier ensemble dedicated to data stream classification task.

Keywords: classifier ensemble, data stream, incremental learning, ensemble pruning, forgetting.

1 Introduction

One of the main challenge of the modern computer classification systems is to propose efficient approach to analyze data stream. Such tool should take into consideration the following characteristics which separate the data stream classification task from the traditional canonical classification model:

- the statistical dependencies among the input features described given objects and their classifications may be changing,
- data can come flooding in the classifier what causes that it is impossible to label all incoming objects manually by human experts.

The first phenomena is called *concept drift* and according to [9] we can propose its taxonomy: gradual drift (smooth changes), sudden drift (abrupt changes), and

N.T. Nguyen et al. (Eds.): ACIIDS 2014, Part II, LNAI 8398, pp. 90–99, 2014.

reoccurring drift (changes are either periodical or unordered). When we face with the pattern classification task with the possibility of concept drift appearance then we can consider the two main approaches which strictly depend on the drift type:

- Detecting concept drift in new data and if these changes are significant, then retrain the classification model.
- Adopting a classification model to changes.

In this work we will focus on the last issue. The model is either updated (e.g., neural networks) or needs to be partially or completely rebuilt (as CVFDT algorithm [3]). Usually we analyze the data stream using so-called data chunks (successive portions of incoming data). The main question is how to adjust the data chunk size. On the one hand, a smaller chunk allows focusing on the emerging context, though data may not be representative for a longer lasting context. On the other hand, a bigger chunk may result in mixing the instances representing different contexts. One of the important group of algorithms dedicated to stream classification exploits strength of ensemble systems, which work pretty well in static environments [6]. An assumed strategy for generating the line-up of the classifier ensemble should guarantee its diversity improvement and consequently accuracy increasing.

The most popular ensemble approaches, as the *Streaming Ensemble Algorithm* (SEA) [13] or the *Accuracy Weighted Ensemble* (AWE)[14], keep a fixed-size set of classifiers. Incoming data are collected in data chunks, which are used to train new classifiers. If there is a free space in the ensemble, a new classifier joins the committee. Otherwise, all the classifiers are evaluated based on their accuracy and the worst one in the committee is replaced by a new one if the latter has higher accuracy. The SEA uses a majority voting strategy, whereas the AWE uses the more advanced weighted voting strategy. A similar formula for decision making is implemented in the *Dynamic Weighted Majority* (DWM) algorithm [5]. Nevertheless, unlike the former algorithms, the DWM modifies the weights and updates the ensemble in a more flexible manner. The weight of the classifier is reduced when the classifier makes an incorrect decision. Eventually the classifier is removed from the ensemble when its weight falls below a given threshold. Independently, a new classifier is added to the ensemble when the committee makes a wrong decision. Some evolving systems continuously adjust the model to incoming data, what is called implicit drift detection [7] as opposed to explicit drift detection methods that raise a signal to indicate change. In this work we propose the modification of the previously developed dynamic ensemble model called WAE (*Weighted Aging Ensemble*) which can modify the line-up of the classifier committee on the basis of the linear combination of diversity measure called *Generalized Diversity* and accuracy. Additionally the decision about object's label is made according to weighted voting, where weight of a given classifier depends on its accuracy and time spending in an ensemble. The detailed description of the algorithm is presented in the next section. In this work we would like to study how the method of individual classifier selection to

a classifier committee could influence the compound classifier quality. Then we present preliminary results of computer experiments which were carried out on SEA and Hyper Plane Stream datasets. The last section concludes our research.

2 WAE - Classifier Ensemble for Data Stream Classification

Let's propose the idea of the WAE (*Weighted Aging Ensemble*), which was firstly presented in [15], then its modification will be presented. We assume that the data stream under consideration is given in a form of data chunks denotes as $\mathcal{DS}_k$, where k is the chunk index. The concept drift could appear in the incoming data chunks. We do not detect it, but we try to construct self-adapting classifier ensemble. Therefore on the basis of the each chunk some individuals are trained using different classifier models and we check if they could form valuable ensemble with the previously trained classification models. Because we assume the fixed size of the ensemble therefore we should select the most valuable classifier committee line-up on the basis of the exhaustive search (the number of the possible ensembles is not so high). In the previous versions our algorithm we proposed to use the *Generalized Diversity* (denoted as $\mathcal{GD}$) proposed by Partridge and Krzanowski [10] as the search criterion to assess all possible ensembles and to choose the best one. $\mathcal{GD}$ returns the maximum values in the case of failure of one classifier is accompanied by correct classification by the other one and minimum diversity occurs when failure of one classifier is accompanied by failure of the other.

$$\mathcal{GD}(\Pi) = 1 - \frac{\sum_{i=1}^{L} \frac{i(i-1)p_i}{L(L-1)}}{\sum_{i=1}^{L} \frac{ip_i}{L}} \tag{1}$$

where L is the cardinality of the classifier pool (number of individual classifiers) and p_i stands for the probability that i randomly chosen classifiers from Π will fail on randomly chosen example.

Lets $P_a(\Psi_i)$ denotes frequency of correct classification of classifier Ψ_i and $itter(\Psi_i)$ stands for number of iterations which Ψ_i has been spent in the ensemble. We propose to establish the classifier's weight $w(\Psi_i)$ according to the following formulae

$$w(\Psi_i) = \frac{P_a(\Psi_i)}{\sqrt{itter(\Psi_i)}} \tag{2}$$

and the final decision returned by the compound classifier Ψ is given by the following formulae

$$\Psi(x) = \arg\max_{j \in \mathcal{M}} \sum_{k=1}^{L} [\Psi_k(x) = j]\, w(\Psi_k), \tag{3}$$

where $\mathcal{M}$ denotes the set of possible labels, x is feature values, and [ ] stands for Inverson's bracket.

This proposition of classifier aging has its root in object weighting algorithms where an instance weight is usually inversely proportional to the time that has passed since the instance was read [4] and Accuracy Weighted Ensemble (AWE)[14], but the proposed method called Weighted Aging Ensemble (WAE) incudes two important modifications:

1. classifier weights depend on the individual classifier accuracies and time they have been spending in the ensemble,
2. individual classifier are chosen to the ensemble on the basis on the non-pairwise diversity measure.

In our work we propose tho replace $\mathcal{GD}$ (1) as the ensemble pruning criterion by the linear combination of the ensemble accuracy and the mentioned above measure

$$\mathcal{Q}(\Pi) = a\mathcal{GD}(\Pi) + (1-a)P_a(\Psi), \quad \text{where} \quad a \in [0,1] \tag{4}$$

where Ψ is classifier ensemble using pool of individual classifiers Π, P_a denotes its accuracy, and a stands for arbitrary chosen factor.

The WAE pseudocode is presented in Alg.1 [15].

Algorithm 1. Weighted Aging Ensemble (WAE) based on heterogenous classifiers

Require: input data stream,
data chunk size,
k classifier training procedures,
ensemble size L

$i := 1$
$\Pi = \emptyset$
repeat
  collect new data chunk DS_i
  for $j := 1$ **to** k **do**
    $\Psi_{i,j} \leftarrow$ classifier training procedure (DS_i,j)
    $\Pi := \Pi \cup \{\Psi_{i,j}\}$ to the classifier ensemble Π
  end for
  if $|\Pi| > L$ **then**
    choose the most valuable ensemble of L classifiers using (4)
  end if
  for $j = 1$ **to** L **do**
    calculate $w(\Psi_i)$ according to (2)
  end for
until end of the input data stream

3 Experimental Investigations

The aims of the experiment were:

- assessing if the proposed method of weighting and aging individual classifiers in the ensemble is valuable proposition compared with the methods which do not include aging or weighting techniques,
- establishing the dependency between the a factor value used in (4) and quality of the WAE algorithm

3.1 Set-Up

All experiments were carried out on two syntectic benchmark datasets:

- the **SEA** dataset [13] where each object belongs to the on of two classes and is described by 3 numeric attributes with value between 0 and 10, but only two of them are relevant. Object belongs to class 1 (TRUE) if $arg_1 + arg_2 < \phi$ and to class 2 (FALSE) otherwise. ϕ is a threshold between two classes, so different thresholds correspond to different concepts (models).Thus, all generated dataset is linearly separable, but we add 5% noise, which means that class label for some samples is changed, with expected value equal to 0. We simulated drift by instant random model change.
- **Hyper Plane Stream** [16] where each object belongs to one of the 5 classes and is described by 10 attributes. The dataset is a synthetic data stream containing gradually evolving (drifting) concepts. The drift is appeared each 800 observations.

For each of the experiments we decided to form heterogenous ensemble i.e., ensemble which consists of the classifier using the different models (to ensure its higher diversity) and we used the following models for individual classifiers:

- Naïve Bayes,
- decision tree trained by C4.5 [12],
- SVM with polynomial kernel trained by the sequential minimal optimization method (SMO) [11]
- nearest neighbour classifier,
- classifier using a multinominal logistic regression with a ridge classfier [8],
- OneR [2].

During each of the experiment we tried to evaluate dependency between data chunk sizes (which were fixed on 50, 100, 150, 200) and overall classifier quality (accuracy and standard deviation) and the diversity of the best ensemble for the following ensembles:

1. *simple* - an ensemble using majority voting without aging.
2. *weighted* - an ensemble using weighted voting without aging, where weight assigned to a given classifier is inversely proportional to its accuracy.

3. *aged* - an ensemble using weighted voting with aging, where weight assigned to a given classifier is calculated according to (2).

Method of ensemble pruning was the same for each ensembles and presented in (4). We run the experiments for different a values ($a \in \{0.0, 0.1, ..., 1.0\}$).

All experiments were carried out in the Java environment using Weka classifiers [1]. The new individual classifiers were trained on a given data chunk. The same chunk was used to prune the classifier committee, but the ensemble error was estimated on the basis on the next (unseen) portion of data.

3.2 Results

The results of experiment are presented in Fig.1-2 and in Tab. 1-2. The figures show the accuracies and diversity for different types of ensembles and different values of a factor and chunk size. Tab.1-2 present overall accuracy and standard deviation for the tested methods and how they depend on data chunk size. Unfortunately, because of the space limit we are not able to presents all extensive results, but they are available on demand.

Table 1. Classification accuracies and diversities for different sizes of data chunk for SEA dataset

chunk size	ensemble type	accuracy	sd	diversity	sd
50	*simple*	0.895	0.0059	0.476	0.0175
	weighted	0.893	0.0064	0.481	0.0165
	aged	0.895	0.0047	0.480	0.0136
100	*simple*	0.902	0.0048	0.466	0.0211
	weighted	0.904	0.0063	0.450	0.0170
	aged	0.906	0.0054	0.456	0.0196
150	*simple*	0.907	0.0075	0.437	0.0162
	weighted	0.910	0.0040	0.448	0.0306
	aged	0.908	0.0047	0.447	0.0297
200	*simple*	0.904	0.0046	0.459	0.0230
	weighted	0.899	0.0110	0.451	0.0355
	aged	0.904	0.0028	0.429	0.0268

3.3 Discussion of the Results

SEA dataset:

– The overall accuracies of the tested ensembles are stable according to the chunk sizes. We observed a slight accuracy improvement, but it is statistical significant for the chunk sizes 50 and 150 only (t-test). The standard deviation of the accuracies is unstable, but it is smallest for the chunk size 150. The observation is useful because the bigger size of data chunk means that effort dedicated to building new models is smaller because they are being built rarely.

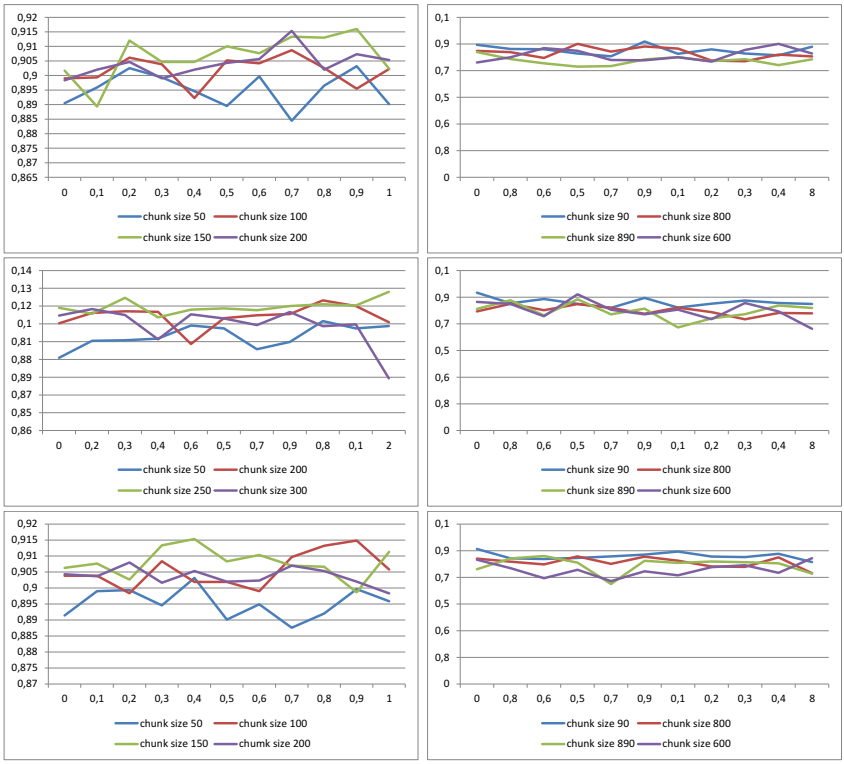

Fig. 1. The computer experimental results for the SEA dataset. Dependencies between a factor used in the pruning criterion (4) and ensembles' accuracies (left) diversities (right) for three type of classifier ensemble: *simple* (top), *weighted* (middle), and *aged* (bottom), and for 4 different sizes of data chunk.

- The overall diversities do not depend strongly on chunk size.
- Taking into consideration the mentioned above observations we may suggest that the best choice of chunk size is ca. 150, especially for *weighed* and *aged* ensemble.

Hyper Plane Stream dataset:

- The overall accuracies of the tested ensembles increase according to chunk size. The differences are statistically significant between the following pais of chunk sizes: 50 and 150, 50 and 200, 100 and 200 (t-test).
- The standard deviations of all ensemble accuracies increase according the chunk size.
- The ensemble diversity is decreasing according to the chunk sizes but the standard deviation is increasing.
- Taking into consideration the mentioned above observations we may suggest that the best choice of chunk size is also ca. 150.

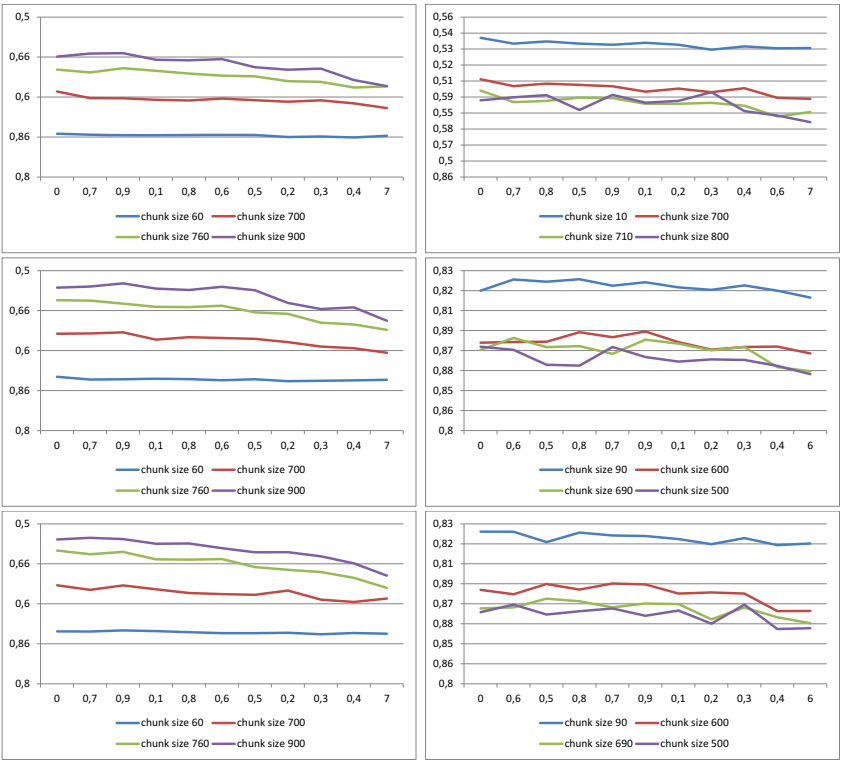

Fig. 2. The computer experimental results for the Hyper Plane Stream dataset. Dependencies between a factor used in the pruning criterion (4) and ensembles' accuracies (left) diversities (right) for three type of classifier ensemble: *simple* (top), *weighted* (middle), and *aged* (bottom), and for 4 different sizes of data chunk.

Table 2. Classification accuracies and diversities for different sizes of data chunk for Hyper Plane Stream dataset

chunk size	ensemble type	accuracy	sd	diversity	sd
50	*simple*	0.452	0.0014	0.371	0.0021
	weighted	0.463	0.0015	0.366	0.0028
	aged	0.463	0.0015	0.370	0.0025
100	*simple*	0.486	0.0051	0.339	0.0037
	weighted	0.497	0.0082	0.338	0.0034
	aged	0.507	0.0069	0.336	0.0048
150	*simple*	0.513	0.0083	0.331	0.0043
	weighted	0.526	0.0126	0.330	0.0052
	aged	0.520	0.0146	0.330	0.0039
200	*simple*	0.514	0.0133	0.324	0.0060
	weighted	0.537	0.0152	0.328	0.0043
	aged	0.535	0.0145	0.328	0.0043

The interesting observation may be made analyzing the dependency among a factor values, diversity, and accuracy of the ensembles. The clear tendencies were observed for Hyper Plane Stream dataset only. The accuracy and diversity were decreasing according to the a value. It is surprising, because if a is close to 1 then the diversity should play the key role in the pruning criterion (4), but the overall diversity is higher for the ensembles formed using the mentioned criterion for the small a (what means that accuracy plays the key role in this criterion).

4 Conclusions

In this paper we discussed the aging ensemble classifier applied to data stream classification problem WAE (*Weighted Aging Ensemble*), which uses dynamic classifier ensemble line-up, which is formed when new individual classifiers trained on new data chunk are come and the decision which classifiers are chosen to the ensemble is made on the basis of the linear combination of the ensemble accuracy and the diversity measure. The decision is made according to weighted voting where weight assigned to a given classifier depends on its accuracy (proportional) and how long the classifier participates in the ensemble (inversely proportional). Formulating general conclusions from the experiments is risky because of their limited scope, but it is clearly visible that using the diversity measure dedicated for the static classification is not appropriate for the data stream classification task. We observed that the better accuracy, evaluated on unseen chunks, could be achieved using only accuracy as the pruning criterion and what was surprising such strategy caused that chosen ensemble had the highes diversity according to $\mathcal{GD}$. To formulate the strong conclusions on the basis of computer experiments their scope should be significantly extended. Additionally, the used diversity measure does not seem to be appropriate for the data stream classification tasks, therefore we would like to extend the scope of experiments by using another non-pairwise diversity measures and maybe to propose a new one which can evaluate diversity taking into consideration the nature of the discussed pattern classification task.

It is worth noting that classifier ensemble is a promising research direction for aforementioned problem, but its combination with a drift detection algorithm could have a higher impact to the classification performance.

Acknowledgment. The work was supported by EC under FP7, Coordination and Support Action, Grant Agreement Number 316097, ENGINE European Research Centre of Network Intelligence for Innovation Enhancement (http://engine.pwr.wroc.pl/).

References

1. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
2. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Mach. Learn. 11(1), 63–90 (1993)

3. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proc. of the 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 97–106 (2001)
4. Klinkenberg, R., Renz, I.: Adaptive information filtering: Learning in the presence of concept drifts, pp. 33–40 (1998)
5. Kolter, J.Z., Maloof, M.A.: Dynamic weighted majority: a new ensemble method for tracking concept drift. In: Third IEEE International Conference on Data Mining, pp. 123–130 (November 2003)
6. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
7. Kuncheva, L.I.: Classifier ensembles for detecting concept change in streaming data: Overview and perspectives. In: 2nd Workshop SUEMA 2008 (ECAI 2008), pp. 5–10 (2008)
8. Le Cessie, S., Van Houwelingen, J.C.: Ridge estimators in logistic regression. Applied Statistics, 191–201 (1992)
9. Narasimhamurthy, A., Kuncheva, L.I.: A framework for generating data to simulate changing environments. In: Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, AIAP 2007, Anaheim, CA, USA, pp. 384–389. ACTA Press (2007)
10. Partridge, D., Krzanowski, W.: Software diversity: practical statistics for its measurement and exploitation. Information and Software Technology 39(10), 707–717 (1997)
11. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods, pp. 185–208. MIT Press, Cambridge (1999)
12. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)
13. Street, W.N., Kim, Y.: A streaming ensemble algorithm (sea) for large-scale classification. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2001, pp. 377–382. ACM, New York (2001)
14. Wang, H., Fan, W., Yu, P.S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2003, pp. 226–235. ACM, New York (2003)
15. Woźniak, M., Kasprzak, A., Cal, P.: Weighted aging classifier ensemble for the incremental drifted data streams. In: Larsen, H.L., Martin-Bautista, M.J., Vila, M.A., Andreasen, T., Christiansen, H. (eds.) FQAS 2013. LNCS, vol. 8132, pp. 579–588. Springer, Heidelberg (2013)
16. Xu, X.: Stream data mining repository (2010), `http://www.cse.fau.edu/~xqzhu/stream.html`