Theoretical and Empirical Analysis of Diversity in Non-Stationary Learning

Richard Stapenhurst
School of Computer Science
The University of Manchester
Manchester, UK
Email: richard.stapenhurst@cs.man.ac.uk

Gavin Brown
School of Computer Science
The University of Manchester
Manchester, UK
Email: gavin.brown@cs.man.ac.uk

***Abstract*—In non-stationary learning, we require a predictive model to learn over time, adapting to changes in the concept if necessary. A major concern in any algorithm for non-stationary learning is its rate of adaptation to new concepts. When tackling such problems with ensembles, the concept of *diversity* appears to be of significance. In this paper, we discuss how we expect diversity to impact the rate of adaptation in non-stationary ensemble learning. We then analyse the relation between voting margins and a popular measure of diversity, KW variance, and use the similarities between them to draw some useful conclusions regarding ensemble adaptivity.**

I. Introduction

Non-stationary problems represent a considerable challenge for machine learning algorithms. 'Non-stationary' refers to the distribution of the data with respect to some unmeasured value (often related to time); a common scenario involves receiving a constant stream of data, where each example must be classified and a true label is provided only after classification. At some point, the procedure that generates the data will change - indicating a change in *concept* - necessitating a change to the predictive model. These problems bring two challenges: learning from data that can be viewed only once and must then be discarded, and adapting to the changes of concept.

We encounter concept change in many real-world applications such as financial market prediction, weather forecasting and spam email detection. Žliobaitė provides an excellent contemporary survey of the field [1], indicating the major types of non-stationarity (e.g. gradual drift, sudden changes, recurring concepts), and some of the approaches that are used to deal with them, such as change detection ('trigger based'), or continuous adaptation ('evolving').

Some popular approaches to non-stationary learning utilise *ensembles* [2] . Ensembles are committees of *base learners* that are trained together and have their predictions combined. With such an approach, we hope to distill the learning problem into multiple base learners in such a way that, following concept change, some base learners will remain relevant to the new concept. Dynamic Weighted Majority [3] works on this basis, evaluating the performance of each learner and updating its combination weight accordingly, adding or removing classifiers based on ensemble performance. Learn^{++} [4] is similar, but keeps an ensemble of static base learners, adding a new learner with each new batch of data, and reweighting existing learners using boosting principles. Both of these techniques monitor the error rate to implicitly detect change, and update the ensemble to accommodate this.

An important consideration for non-stationary learning algorithms is *"How quickly does the model adapt to a new concept?"*; appreciating the factors that determine this rate of adaptation is crucial to understanding and developing techniques. In this paper we consider the impact of *diversity* on the capability of an ensemble to learn new concepts. The hypothesis is that high diversity will be indicative of *quicker adaptation*.

978-1-4244-9931-1/11/$26.00 ©2011 IEEE

II. Background

In this section, we introduce some fundamental concepts that are essential for understanding the contributions of the paper.

A. Margins

Popular ensemble training techniques such as Bagging [5] and Boosting [6] combine their base learners using uniform or weighted voting; in two class problems ($y \in \{-1, 1\}$), for base learners $h_1 \dots h_L$, the ensemble prediction H of example x is: $H(x) = sign(\sum_{l=1}^{L} \alpha_l h_l(x))$, where α_l is the weight of the l^{th} base learner (α is normalised so that $\sum_{l=1}^{L} \alpha_l = 1$, and $\alpha = \frac{1}{L}$ for uniform voting). The *voting margin*[1] is defined as the 'confidence-weighted correctness' of the prediction on an example x_i:

$$m_i = y_i \sum_{l=1}^{L} \alpha_l h_l(x_i). \tag{1}$$

The sign of a margin indicates whether x_i is correctly predicted: $m_i > 0 \implies H(x_i) = y_i$. Margins are always in the range $m_i \in [-1, 1]$, with their magnitude indicating the confidence of the prediction. $|m_i| = 1$ occurs only when the whole ensemble makes the same prediction.

B. Diversity

Diversity is intuitively an important property of an ensemble; there have to be some differences between the predictions of the base learners if we are to gain anything by combining them. Kuncheva and Whitaker [8] analyse several measures of diversity, recommending the Q-statistic [9] because of its ease of calculation and meaningful interpretation. All the diversity measures can either be considered to be 'symmetric' or 'asymmetric', where asymmetric measures are of the form $f(h_1 \dots h_l, y)$, depending on the ensemble *and* the target variable, while symmetric measures are of the form $f(h_1 \dots h_l)$ and are independent of the target.

Tang [10] expresses several diversity measures in terms of average base learner accuracy and the number of incorrect predictions, using this reformulation to show that maximising diversity is equivalent to maximising the minimum margin *if average base learner error remains constant*. Brown [11] and Chen [12] present two equivalent decompositions of the misclassification rate of an ensemble, e_{maj}, showing that it comprises an average individual error term, e_{ind}, and a diversity term, Δ, that can be either positive or negative; for example, in [11, Equation 12]:

$$\begin{aligned} e_{maj} &= e_{ind} - \Delta_i, & (2) \\ \Delta_i &= y_i H(x_i) \frac{1}{L} \sum_{l=1}^{L} \frac{1}{2}(1 - h_l(x_i) H(x_i)). & (3) \end{aligned}$$

When the ensemble prediction is incorrect, the Δ term is additive, when the ensemble is correct, the Δ term is subtractive. These situations are described as "good" and "bad" diversity respectively; "good" diversity is when there is high disagreement but the ensemble is correct, "bad" diversity is high disagreement when the ensemble is incorrect. Again, this is with the constraint that *average base learner error remains constant*; alternatively, one could think of the diversity term as offsetting changes in average base learner accuracy when they do not affect the ensemble prediction.

In the remainder of this paper, we show how margins and diversity are related, and how we can apply this knowledge to non-stationary learning.

III. Margins and Diversity in Stationary Learning

In this section, we introduce Kohavi-Wolpert variance [13] - a measure of diversity - and discuss its application in *stationary* learning.

A. Kohavi-Wolpert Variance

We start with a symmetric measure, KW variance, which captures the variability in the predictions on a single example:

$$\kappa_i = \frac{1}{2}(1 - [\hat{P}(y_i|x_i)^2 + \hat{P}(-y_i|x_i)^2]), \tag{4}$$

this is derived from a bias-variance decomposition of 0/1 classification error.

[1] Voting margins are shown by Fruend and Schapire [7] to be useful in determining generalisation error. Larger margins reduce an upper bound on generalisation error

Theorem 1. *The diversity measure, Kohavi-Wolpert variance, can be expressed in terms of the margins on the data.*

$$\kappa_i \propto -m_i^2. \tag{5}$$

Proof - See Appendix A.

This is a pleasantly simple relationship. Kuncheva [14, Appendix 10A] also shows that is κ related to the 'disagreement' measure [15], which counts the number of pairs of classifiers that disagree on each example, as $\kappa = \frac{L-1}{2L} dis$. KW variance seems more appropriate than the Q-statistic because of these links; unlike the Q-statistic, it can be written in terms of margins, is related to the intuitive disagreement measure, and has a meaningful value even on single examples.

B. The Role of Diversity in Stationary Learning

There is well accepted theory showing that some level of diversity is required to achieve the best performance on training data [16] [17]. This indicates the role of diversity in stationary learning problems; a set of suitably diverse base learners is required for the ensemble is to be expressive enough to correctly classify more of the training data than a single model.

Theorem 2. *The diversity term Δ from the decomposition of ensemble error (Equation 3) has a sign that is determined by $y_i H(x_i)$ (ensemble correctness) and a magnitude that is determined by the absolute value of the margin $|m_i|$:*

$$\Delta_i \quad = \quad \frac{1}{2} y_i H(x_i)(|m_i| - 1). \tag{6}$$

Proof - See Appendix B.

Given that $|m_i| = \sqrt{m_i^2}$, there is clearly a close relationship between this quantity and KW variance. This theorem shows that KW variance is involved in determining the *training performance* of an ensemble. In fact, with the exception of Chen [12], who showed some interesting empirical correlation between diversity and generalisation error, a relationship between diversity and generalisation error is not proposed anywhere in the literature.

IV. The Role of Diversity in Non-Stationary Learning

In the previous section, we have seen that there is a close relationship between a symmetric diversity measure (KW variance) and voting margins. Furthermore, we have seen exactly how diversity appears in a decomposition of ensemble training error. The main question that we address in this paper is *What is the role of diversity in non-stationary learning?*.

To some extent, Equation 6 still applies. However, since the coefficient $y_i H(x_i)$ depends on the target y_i, we can see that the diversity term will be perturbed by a change in concept; therefore, this equation alone does not suggest any principled way of dealing with concept change.

However, we can make an observation regarding the quantity $|m_i|$: for uniformly weighted ensembles, this value determines the number of base learners that must change their predictions in order to 'swing' the vote on x_i.

Theorem 3. *The number of base learner predictions that must change to 'swing' the vote on x_i is:*

$$\left\lceil \frac{L|m_i| + 1}{2} \right\rceil. \tag{7}$$

Proof - In Appendix C.

This number can easily be related to the rate of adaptivity of the ensemble, since it indicates that fewer base learners need to change their predictions to affect the ensemble prediction.

Considering our previous connection between margins and diversity (Theorem 1), we have shown that *reducing the absolute value of the margins is equivalent to increasing the amount of diversity*. Therefore, our two hypotheses - that diverse ensembles should adapt quickly, and that the absolute margin is instrumental in determining the rate of adaptivity - converge on the same prediction, which we now test empirically.

V. Experiments

In this section we investigate the following hypothesis:

Hypothesis - *When two ensembles, with identical error rates, but different amounts of diversity, are introduced to a new concept, the ensemble with higher diversity will adapt more quickly.*

A. Experimental Setup

The goal of these experiments is to examine the effect of the diversity of an ensemble when it encounters a new concept. To this end, we will generate ensembles of varying diversity that have similar performance on an initial concept. We will then present data from a new concept in an online learning fashion to these ensembles, and evaluate the rate at which they adapt to the new concept.

1) Data: Since we are only concerned with abrupt concept changes, we simply need two datasets with the same number of features; one for the initial concept and one for the new concept. Within each concept, the data is independent and identically distributed. The severity of the difference between the two concepts is important, since we would expect the impact of diversity to be more pronounced when severity is higher.

Figure 1 shows how we generate concepts of varying severity. We add some features to a stationary dataset, and fill them with noise. For a new concept, we swap some of the original features with the new features, requiring a learning algorithm to forget about some of the old features and learn to use the new ones. The number of features that we change in this way can be varied to affect the severity of the drift.

In our experiments, where multiple repetitions are performed, we choose different features to swap at every repetition, thereby mitigating the impact of datasets where some features are more important than others. Figure 2 lists the datasets that were used and their associated statistics.

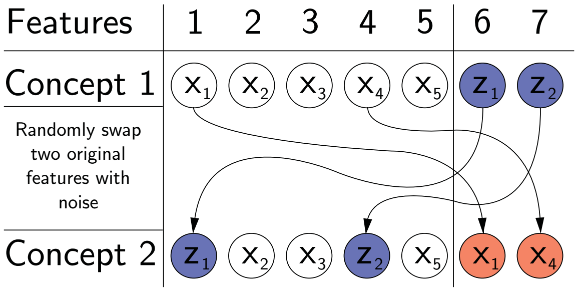

Fig. 1. We take an original stationary problem with 5 features, $x_1 \dots x_5$. For the first concept, we append two noise features, z_1, z_2, to this. The ensemble is then trained on this concept. To produce a second concept, we swap z_1, z_2 with two random original features (in this case, x_1, x_4). This gives us two 7-feature concepts, both with 2 irrelevant features, where 3 features (x_2, x_3, x_5) are consistent between concepts. Increasing the number of noise features that we swap will decrease the similarity between concepts. Since the concepts both contain all the original features (the difference being in the *order* of the features), they will share common properties such as the Bayes error rate.

Dataset	Examples	Features
Heart Disease	270	13
Breast Cancer	569	30
Pima Indians Diabetes	768	8
STAGGER	1500	3

Fig. 2. Dataset properties.

Generalisation errors are computed using holdout samples. The amount of holdout data varies between datasets; for UCI data, we use 200 examples for training on the second concept, and the remaining examples for computing generalisation error. For STAGGER, we use 500 examples to compute generalisation error.

2) Experimental Protocol: We briefly give an algorithmic description of the experimental protocol.

This process is repeated 10000 times, and we compute the average error rates from Step 9. Most of the graphs we present in the next section are the result of several experiments. We vary the severity parameter, repeating the experiment for every possible value. The method we use to generate H is not critical; we ensured experimentally that there was a reasonable spread of diversity values (typically giving mean KW variance in [0.05, 0.25]), and that the diversity was independent of training error within the specified range. All error ranges were $[min, min + 0.02]$, with min values of 0.03, 0.2 and 0.23 for Breast Cancer, Heart Disease and Dia-

Algorithm 1 Experimental Protocol

```
Require: dataset, severity, min/max errors.
Generate two concept datasets, $C_1$ and $C_2$ using severity parameter to determine the number of random overlapping features.
repeat
    Generate an ensemble $H$.
    $err \leftarrow$ Training error of $H$ on $C_1$.
until $err \in [min, max]$
Compute diversity of $H$ on $C_1$.
for each example in $C_2$ do
    Train $H$ on the next example from $C_2$.
    Compute generalisation error of $H$ on $C_2$ hold-out data.
end for
```

betes datasets respectively; these values depended on what predictive performance was possible on each dataset.

The ensembles consisted of 40 perceptrons, and when learning the second concept they were trained using the Online Bagging algorithm [18]. Perceptrons were used as base learners because their learning rate is dependent only on a parameter (learning rate, which we held at 0.05 for all experiments), while other algorithms (e.g. classification trees, naive bayes, decision stumps) will learn at different rates depending on the amount of data and noisiness of data seen previously. Similarly, Online Bagging makes no provision for concept change, so adaptation will be entirely determined by the diversity of the ensemble and the learning rate of the perceptrons. In all experiments, we calculated 95% confidence intervals, but found that (due to the number of repetitions) they were too small to serve any purpose in the graphs we present.

Minku [19] carries out a similar investigation, although with a more pragmatic emphasis than ours. A λ parameter in Online Bagging is used as a proxy for diversity, and diversity is maintained throughout the experiment, while we only enforce certain amounts of diversity on the initial concepts. The results are broadly consistent, although ours are specifically tailored to address our hypotheses regarding margins, diversity and adaptivity.

B. Results

We show results on 4 datasets: heart disease, breast cancer, diabetes and STAGGER [20]. The first three are UCI datasets where we produce concepts using the technique described previously, while STAGGER is a popular artificial concept change problem. We vary the number of overlapping features to investigate how the severity of change affects performance.

We make use of two kinds of plot. Where the problem is already non-stationary (STAGGER), we show the performance *difference* between high and low diversity ensembles as data is received (Figure 3). When the difference is positive, this indicates an advantage for diverse ensembles, and when it is negative, an advantage for non-diverse ensembles. For the other problems, we generate concepts; since severity is a parameter to this process (the proportion of features to swap), we investigate varying amounts of severity. These plots show performance difference using brightness (light is an advantage for diverse ensembles, dark for non-diverse), with time on the x-axis as before, and severity on the y-axis. Low values of severity indicate a small difference between the two concepts, while high values indicate a large difference.

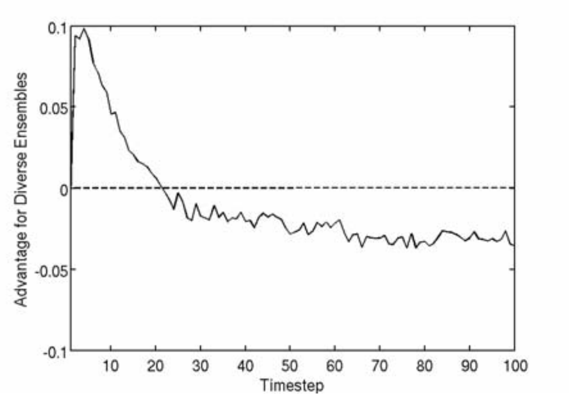

Fig. 3. **STAGGER dataset (transition from concept 1 to concept 2)**. The difference in generalisation error rate between the least and most diverse ensembles on the first and second concepts from the STAGGER dataset. The x-axis indicates number of examples since the concept change. The y-axis indicates the difference in error rate (positive means that less diverse ensembles have higher error).

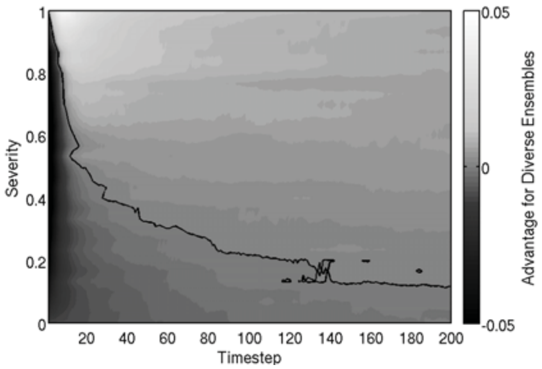

Fig. 4. **Breast Cancer dataset**. The difference in generalisation error rate between the least and most diverse ensembles on concepts generated from the Breast Cancer dataset; White areas show the biggest gains for high diversity, while dark areas indicate that low diversity is more desirable. The x-axis indicates number of examples since the concept change. The y-axis indicates the severity of concept change (the proportion of the 30 features that were swapped with random noise). The dark line indicates no performance difference between diverse and non-diverse. The intensity of the background colour indicates the percentage advantage for diverse ensembles for the given timestep and severity.

Figure 4 shows the effect of diversity on generalisation error for the heart disease dataset. On the leftmost side of the plot, we see that initial performance is usually similar, or that less diverse ensembles are better when severity is low. The intense white patch indicates a huge advantage to highly diverse ensembles, where they adapt to the new concept quickly; the difference is more noticeable when severity is higher. As more data from the new concept arrives, we see that the difference shrinks, with the less diverse ensembles doing better when severity is low.

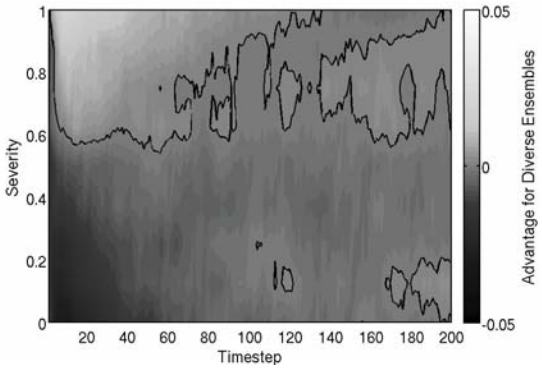

Fig. 5. **Pima Indians Diabetes dataset**. See Figure 4 for explanation.

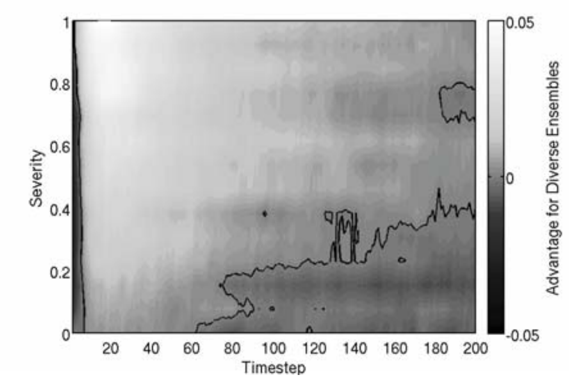

Fig. 6. **Heart Disease dataset**. See Figure 4 for explanation.

The other plots have similar properties, although Figure 5 only shows an advantage for diverse ensembles when severity is very high. Figure 3 shows the difference in performance on the second concept in the STAGGER dataset; since STAGGER concepts are already defined, we did not vary severity for this plot. We see again that the diverse ensembles have a large advantage initially, because they adapt more quickly, but that less diverse ensembles soon match their performance.

The STAGGER dataset exhibits unusual behaviour in that the *initial* generalisation error (at timestep 0) is lower for diverse learners. This is due to the high severity of concept drift between concepts 1 and 2 in STAGGER; of the 27 equally probable values for x, the true label differs between the concepts in 14 cases, meaning that applying a classification rule from concept 1 to concept 2 gives worse performance than randomly guessing the labels.

VI. Conclusions

We draw three conclusions from the experimental data:

1) *Low diversity ensembles typically perform better on new concepts* - High diversity makes

an ensemble perform closer to random guessing on a new concept. Unless the severity is very high (so that the new concept is 'opposite' of the old concept; Minku [19] investigates such concept changes more thoroughly), this means that low diversity ensembles perform better immediately after concept change.

2) *High diversity ensembles always adapt faster* - The generalisation error of the more diverse ensembles is always reduced faster, although its initial value may be higher when severity is low (for example, in Figure 5 the diverse ensembles are only successful when more than half of the features change in the second concept).
3) *Ensembles of any diversity converge to a similar final performance* - With the exception of the STAGGER dataset (due to, we expect, the discrete nature of STAGGER input space and imposed ordinality from Perceptron base learners), all the ensembles finally achieve similar generalisation error.

In this paper, we studied the problem of non-stationary learning, and the diversity property of an ensemble. We then showed some intuitive measurements of plasticity, which turned out to be closely related to diversity. Next, we explained our intuitions regarding the role of diversity in non-stationary problems, and how this differs from its role in stationary problems. Finally, we performed some experiments which confirmed that higher diversity was beneficial when adapting to a new concept.

APPENDIX

A. Derivation of KW Variance

We start from [13, Equation 3], which gives classification error in terms of bias, variance and noise:

$$err = \sum_x P(x)(bias_x^2 + variance_x + \sigma_x^2). \quad (8)$$

This is a decomposition of the 0/1 loss on a single classifier, where we consider the variance caused by possible training datasets. The $variance_x$ term is:

$$\kappa_i = \frac{1}{2}(1 - [\hat{P}(y_i|x_i)^2 + \hat{P}(-y_i|x_i)^2]), \quad (9)$$

where $\hat{P}(y_i|x_i)$ indicates the predicted probability of the correct label on example x_i (i.e. the confidence with which the ensemble assigns label y_i to example x_i). We can write this using margins, by using the value of m_i as:

$$\hat{P}(y_i|x_i) = \frac{1}{2}(1 + m_i), \quad (10)$$

$$\hat{P}(-y_i|x_i) = \frac{1}{2}(1 - m_i). \quad (11)$$

This gives:

$$\begin{aligned} \kappa_i &= \frac{1}{2}(1 - \frac{1}{4}(1+m_i)^2 - \frac{1}{4}(1-m_i)^2), \quad (12) \\ &= \frac{1}{4}(1 - m_i^2). \quad (13) \end{aligned}$$

Finally, since we will be concerned with how KW variance is maximised or minimised, we can remove constants and scale by $\frac{1}{4}$ to give a convenient form:

$$\kappa_i \propto -m_i^2. \quad (14)$$

This concludes the proof to Theorem 1.

B. Reformulation of Classification Ensemble Error Decomposition

Brown [11] gives the difference between ensemble error and average base learner error on example x_i to be:

$$\Delta_i = -y_i H(x_i) \frac{1}{L} \sum_{l=1}^{L} \frac{1}{2}(1 - h_l(x_i) H(x_i)). \quad (15)$$

We can rearrange to give:

$$\Delta_i = \frac{1}{2} y_i H(x_i) [H(x_i) \frac{1}{L} \sum_{l=1}^{L} h_l(x_i) - 1]. \quad (16)$$

Now, observing that $H(x_i) = sign(\sum_{l=1}^{L} h_l(x_i))$, we can see that multiplying the summed term by $H(x_i)$ is equivalent to taking its absolute value. Progressing from here, using the definition of the margin and the fact that $y_i \in \{-1, 1\}$, this is also the absolute value of the m_i.

$$\Delta_i = \frac{1}{2} y_i H(x_i)[|\frac{1}{L}\sum_{l=1}^{L} h_l(x_i)| - 1], \quad (17)$$

$$= \frac{1}{2} y_i H(x_i)[|m_i| - 1]. \quad (18)$$

This concludes the proof to Theorem 2.

C. Proof of Number of Votes Required to Swing the Ensemble Prediction

In an ensemble, the difference in the number of votes between the most popular class and the other class, d_i, is:

$$d_i = H(x_i) \sum_{l=1}^{L} h(x_i). \quad (19)$$

this quantity is always positive due to the definition of H. Hence, we can write it as a multiple of the absolute margin:

$$d_i = L|m_i|. \quad (20)$$

Now we consider what happens when one base learner where $h(x_i) \neq H(x_i)$ changes its prediction. This means that we have one fewer vote in favour of the majority, and one more in favour of the minority; $d_i \rightarrow d_i - 2$. To 'swing' the vote, d_i must be < 0. This gives the number of votes required to 'swing':

$$\#swing = \left\lceil \frac{L|m_i| + 1}{2} \right\rceil. \quad (21)$$

This concludes the proof to Theorem 3.

References

[1] I. Zliobaite, "Learning under concept drift: an overview," *CoRR*, vol. abs/1010.4784, 2010.

[2] L. I. Kuncheva, "Classifier ensembles for detecting concept change in streaming data: Overview and perspectives," *Supervised and Unsupervised Ensemble Methods and Their Applications*, vol. 2, pp. 5–9, 2008.

[3] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts," *J. Mach. Learn. Res.*, vol. 8, pp. 2755–2790, December 2007.

[4] M. D. Muhlbaier, A. Topalis, and R. Polikar, "Learn^{++}.nc: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 152–168, 2009.

[5] L. Breiman, "Bagging predictors," in *Machine Learning*, 1996, pp. 123–140.

[6] R. Schapire and Y. Freund, "Decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and Systems Sciences*, 1997.

[7] R. Schapire, Y. Freund, P. Bartlett, and W. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistics*, 1998.

[8] L. Kuncheva and C. Whitaker, "Ten measures of diversity in classifier ensembles: Limits for two classifiers," in *In Proc. of IEE Workshop on Intelligent Sensor Processing*, 2001, pp. 1–10.

[9] G. U. Yule, "On the association of attributes in statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 194, pp. pp. 257–319, 1900.

[10] E. K. Tang, P. N. Suganthan, and X. Yao, "An analysis of diversity measures," *Machine Learning*, vol. 65, pp. 247–271, 2006.

[11] G. Brown and L. I. Kuncheva, ""good" and "bad" diversity in majority vote ensembles," in *MCS*, 2010, pp. 124–133.

[12] H. Chen, "Diversity and regularization in neural network ensembles," Ph.D. dissertation, University of Birmingham, 2008.

[13] R. Kohavi and D. H. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *MACHINE LEARNING: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL*. Morgan Kaufmann Publishers, 1996, pp. 275–283.

[14] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[15] P. D. Turney, "Technical note: Bias and the quantification of stability," pp. 23–33, 1995.

[16] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connect. Sci.*, vol. 8, no. 3, pp. 385–404, 1996.

[17] R. R. Bouckaert, M. Goebel, and P. J. Riddle, "Generalized unified decomposition of ensemble loss," in *Australian Conference on Artificial Intelligence*, 2006, pp. 1133–1139.

[18] N. C. Oza, "Online ensemble learning," Tech. Rep., 2001.

[19] L. L. Minku, A. P. White, and X. Yao, "The impact of diversity on on-line ensemble learning in the presence of concept drift," 2010.

[20] J. C. Schlimmer and R. H. Granger, Jr., "Incremental learning from noisy data," *Mach. Learn.*, vol. 1, no. 3, pp. 317–354, 1986.