Kalpa Publications in Computing

Volume 12, 2019, Pages 227–240

Proceedings of 4th International Conference on the Internet, Cyber Security and Information Systems 2019

The impact of ensemble diversity on learning big data in dynamic environments

Tinofirei Museba
University of Johannesburg
Department of Applied Information Systems
Johannesburg, South Africa
tmuseba@uj.ac.za

Abstract

For many classification tasks, data is collected over an extended period of time and the predictive model learns over time, adapting to changes in the underlying distribution of the data if necessary. To optimize generalization performance, margin distribution is considered to be an important factor. A major concern posed by nonstationary learning for any algorithm is the rate of adaptation to new concepts and the volume of the data. Tackling the problem of learning in nonstationary environments associated with drifting concepts with ensembles of classifiers makes the concept of diversity to be of paramount significance in optimizing the rate of adaptation to new concepts for classification tasks. In this paper, we investigate the impact of ensemble diversity on the rate of adaptation to new concepts in nonstationary learning. The rate of adaptation is analyzed by exploiting the correspondence that exists between voting margins and the double fault measure, a popular diversity measure strongly linked to the margin. We utilize the Adaptive Classifier Ensemble Boost algorithm (AceBoost) to generate diverse base classifiers and optimize margin distribution to exploit different amounts of diversity to generate an optimal ensemble capable of handling different kinds of drift. The experimental results confirm that AceBoost outperforms other state of the art algorithms that exploit ensemble diversity to handle concept drift.

Keywords- ensemble, diversity, margin, support vector machine, concept drift

1 Introduction

The assumption with many machine-learning algorithms is that the classification landscape is static. However, this assumption is not valid for many real world problems. In the case of classification problems, the decision boundaries may change over time because of the changes in the underlying distribution of the data. This phenomenon is not as concept drift (A.Tsymbal, 2004).

Such nonstationary problems represent a considerable challenge for machine learning algorithms designed for classification tasks. In dynamic environments, a learning model receives a constant stream of data, where an example has to be classified and a true label has to be provided only after the classification process. At some point, the underlying data generation mechanism or the concept that we are trying to learn from data is constantly evolving (Haixun Wang, Wei Fan, 2003), indicating a change in concept that necessitates a change of the prediction model. Online learning involves learning from data that is viewed ony once and must then be discarded and adapting to the changes of concept.

Real world applications associated with concept drift include weather forecasting, financial market predictions and spam email detection. One of the most promising and popular approaches to nonstationary learning is ensemble learning (L.I. Kuncheva, 2008). A classifier ensemble is a group of classifiers or learners, models or hypotheses whose prodictions are combined with the goal of obtaining higher classification accuracy (Michael J,Procopio, 2007). Even though ensembles are genuinely capable of incorporating new data by either introducing a new component to the base learners or update existing components to adapt to drift, the literature does not contain any deep study of why they can be helpful for that and which of their features contributes significantly towards adaptation to drift. A better understanding of the behavior of ensembles in the presence of concept drift allows better exploitation of the ensemble features to accurately handle drifting concepts. An important consideration for learning in dynamic environments associated with concept drift is to know how quickly the learning algorithm adapts to new concepts. For an ensemble of classifiers, appreciating the features that determine the rate of adaptation is crucial in the understanding and development of an effective online ensemble. The success of an ensemble classifier in adapting to drift has been hugely attributed to both the accuracy and the diversity among the base learners (Thomas G. Dietterich, 1997). Despite the popularity of the diversity feature of ensemble classifiers, the role it plays has not been visible in research in the presence of concept drift. In this paper, we investigate the impact of diversity on the potential of an ensemble to learn new concepts. The paper provides a deeper understanding of how the diversity of ensemble classifiers can optimize adaptation to new concepts in nonstationary environments associated with drifting concepts. The paper is organized as follows: section 2 looks at the background of ensemble diversity and voting margins. Section 3 provides a preview of related work conducted in the area of exploiting diversity to optimize rate of adaptation.

Section 3.1 and section 3.2 provides a description of our proposed algorithm. Section 3.3 provides a description of the diversity measure used in the formulation of the algorithm. Section provides a description of the drift detection method and a description of the datasets used in the experiments is provided in section 3.5. Section 4 provides a description of the experiments conducted and the results obtained from the experiments are provided in section 5.

2 Background on ensemble diversity and Voting margins

Diversity and voting margins are two important aspects of ensemble methods that are theoretically connected. The voting margin can be defined as the confidence-weighted correctness of the prediction on a given example (Richard Stapenhurst, 2011). Voting margins are useful in determining generalization error. Large margins reduce an upper bound on generalization error. The sign of a margin indicates whether a training instance is correctly classified or not. The magnitude of the margin is an indication of the confidence of the prediction. Diversity is intuitively an important feature of an ensemble classifier as there has to be some differences between the predictions of the base learners .Kuncheva and Whitaker et al (2001) provided an analysis of several diversity measures. Tang et al (2006) proved that maximizing the diversity among the base classifiers is equivalent to optimizing the margin of an ensemble on the training samples if the average classification accuracy is constant and maximal diversity is achievable. As a result, increasing the diversity among base classifiers is an effective approach of improving the margin distribution of ensemble classifiers.

Several diversity measures were expressed in terms of average base learner accuracy and the number of incorrect predictions. Richard Stapehurst et al (2011) proved that there is a connection between margins and diversity for classification tasks. He concluded that reducing the absolute value of the margins is equivalent to increasing the amount of diversity. This led to the fact that diverse ensembles can adapt quickly to any concept and that the absolute margin is instrumental in determining the rate of adaptivity and optimize convergence to new concepts. The next section looks at the related work that exploited ensemble diversity to obtain small generalization eror via optimizing the margin distribution.

3 Related Work

The success of ensemble learning algorithms for learning in nonstationary environments is hugely attributed to both the accuracy and diversity among base learners (T.G Dietterich, 1997) and more empirical studies revealed that there is a positive correlation between accuracy of the ensemble and the diversity among its members (L.I Kuncheva, C.J.Whitaker, 2003). Amongst the first research efforts that investigated the role of ensemble diversity for nonstationary learning was the work of Leandro Minku (2012). The author proposed the Diversity for Dealing with Drifts (DDD), an ensemble algorithm that works on the assumption that diverse ensembles are able to adapt more quickly to concept change. The algorithm maintains various ensembles that have varying amounts of diversity. The algorithm is derived from online bagging and uses a standalone change detector algorithm to detect changes. A significant part of the success of DDD is related to the amount of data that base learners are trained on rather than the diversity of the ensembles. The proxy parameter λ in Online Bagging does not control diversity in isolation but rather affects several factors within the model. DDD does not highlight the correspondence that exists between diversity and voting margins for nonstationary learning. Richard Stapenhurst (2012) opened new possibilities for studying the impact of diversity in nonstationary learning by exploiting the correspondence between diversity and voting margins and how margins interact with nonstationary learning and derived further theoretical predictions about what role diversity plays in nonstationary learning. Ensembles of varying diversity that have similar performance on the initial concept were generated to examine the effect of the diversity of an ensemble when it encounters a new concept. The Online DivBoost manages diversity so as to facilitate fast adaptation to new concepts in nonstationary learning. Online DivBoost uses the Kohavi-Wolpert variance as a diversity measure although the diversity measure is not strongly linked to the margin. A Double Rotation algorithm that generates diverse base classifiers and optimize the margin distribution to exploit the diversity of an ensemble to produce an optimal ensemble was proposed by Qinghua Hu et al (2014). The goal was to obtain an ensemble system with small generalization error via optimizing the margin distribution as a good margin distribution means that most examples have large margins. The algorithm generates a set of diverse decision tree classifiers and exploits the diversity to construct an optimal ensemble. The algorithm was designed for static domains and it uses the disagreement measure to measure the diversity of the base classifiers. In the next section, we provide a description of our proposed algorithm that produces an ensemble system with small generalization error via optimizing the margin distribution. The Adaptive Classifier Ensemble Boost (AceBoost) uses the Double Fault Measure as the diversity measure is strongly linked to the voting margins.

3.1 Online Adaptive Classifier Ensemble Boost (AceBoost)

The Online Adaptive Classifier Ensemble Boost (AceBoost) exploits the correspondence between ensemble diversity and margins to manage diversity explicitly using a proxy parameter. The algorithm enables a principled study of the rate of ensemble diversity that is required for each concept in dynamic environments associated with concept drift for classification tasks. AceBoost learns in an online incremental fashion using techniques analogous to Online AdaBoost. The algorithm improves the margin distribution by minimizing the margin reduced classification loss making it capable of selecting the appropriate amount of diversity suitable for the currently occurring concept for adaptation regardless of the amount the magnitude of change or severity. AceBoost mimics the sampling with replacement of AdaBoost and uses a Poisson distribution instead of a weighted sampling to optimize computational time of normalizing example weights. Each training instance is presented to the training algorithm n times where n is drawn form a Poisson (λ) and λ is a weight derived from a previous performance on that sample. Since the margin of the training samples have an underlying relationship with the diversity of the base classifiers because of the margin loss function, the Poisson distribution λ is equated to the margin loss function μ. Since the algorithm is analogous to Online Adaboost, the property of increasing the margin is inherent in the algorithm. If the margin or separating decision boundary is increased, the generalization performance is improved as well.

3.2 Adaptive Classifier Ensemble Algorithm

The following section provides a description of how the algorithm works by interchanging high and low diversity levels for each concept encountered.

Algorithm : Adaptive AceBoost

Require: : Poisson distribution
∈ : margin loss
H_{low} ← AceBoost (low diversity)
H_{high} ← AceBoost (high diversity)
Let $\lambda = \mu$
For $k \in 1 \ldots N$ do
Train H_{low} on (x_t, y_t)
If change is detected then
Quantify magnitude of change with DDM
Train (x_t, y_t) with H_{high}
Update H_{low}
Endif
End for

3.3 The Double Fault Measure

To control diversity through enlarging and decreasing margins through a proxy parameter, the algorithm uses the Double Fault measure. The diversity of an ensemble learning machine can only be appropriately increased if it is represented by a diversity measure that is strongly related to the average loss and hence the average margin since Tang et al (2006) proved that enlarging the margins can be equated to increasing diversity. Existing diversity measures are more related to the variance of the outputs and do not consider the performance of individual classifiers. The Double Fault Measure is closely related to the generalization error and tends to produce a stable behavior because if strong classifiers are available as it seeks to decrease the probability of identical errors. The Double Fault measure is symmetric and works well in linking it with enlarging margins. The connection between double fault diversity measure and exponential loss creates a good relationship between the double fault diversity measure and ensemble accuracy.
The Double Fault diversity measure is a quadratic function of the margin expressed as:

$D_{DF} = \frac{1}{L(L-1)} \sum_{j=1}^{L} \sum_{k \neq j}^{L} \frac{N_{j,k}^{00}}{N}$ and the margin is interpreted as:

$$D_{DF} = \frac{1}{2}(1\text{-}\bar{m}) - \frac{L}{4(L-1)}\ (1\text{-}\overline{m^2})$$

The term (1-$\bar{m}$) tends to give preference to more accurate ensembles and gives the Double fault measure the impetus of asymmetry and the term (1- $\overline{m^2}$) stands for the notion of diversity. Above all, the double faulty measure takes into consideration the set of margins generated by the whole ensemble instead of just the behaviour of pairs of classifiers (Richard Stapenhurst, 2012)

3.4 Drift Detection Method

The AceBoost algorithm uses the Drift Detection Method (DDM) (J.Gama, 2004) as a change deection mechanism to properly investigate the role of diversity of an ensemble classifier when it encounters a new concept. The use of DDM enables us to see how quickly AceBoost adapts new concepts given perfect change detection information when concept drift is detected, the change detection algorithm quantifies its severity and the low level diversity ensemble is replaced by the high diversity ensemble. The severity of change is varied to avoid scenarios where AceBoost has an advantage. The use of DDM to explicitly handle concept drift allows the change detector to provide useful information about the occurrence, the severity and the width of the encountered new concepts. Ensemble classifiers with a drift detection mechanism combine the flexibility of the ensemble classifier to cope with different types of drifts and has the capacity to provide useful descriptions about location, width and severity or magnitude of the drift.

3.5 Datasets

The exploitation of ensemble diversity by AceBoost to improve adaptation to different types of drifts is validated using one artificial dataset and one real world dataset. We hypothesize that ensembles of varying amounts of diversity are better positioned to accurately adapt to novel concepts and a proper selection of a base learner capable of easily adjusting margins influenced by different amounts of diversity coupled with the efficacy of a change detection mechanism have a significant impact on the accurate adaptation to new concepts. The artificial dataset used in this empirical experiment is the Waveform dataset which is available on the UCI repository. It is 3 class classification problem that is based on 3 waveforms each of which is sampled at one interval and 100 000 instances are generated. Each class is a random convex combination of two of the waveform. The choice of the waveform dataset is motivated by the existence of dataset generators at UCI repository. The real

world dataset used in this experiment is the Forest Covertype. The dataset is contains 581012 instances and 54 attributes. The dataset is composed of cartographical and geological data gathered from over 581012 forests and 30 by 30 meter square cells of undistributed forest cover from cartographical data.

4 Experiments

The goal of these experiments is to examine the impact of diversity of an ensemble when it encounters different kinds of concepts in huge streaming data. To achieve the goal, we generate ensembles of varying diversity whose performance on the initial concept is similar. A drift detection mechanism is used to detect drift and quantify the magnitude or severity of change. Data is presented in an online learning fashion to the ensembles of varying diversity that have been generated and evaluate the rate at which the ensembles adapt to new concepts in nonstationary environments that exhibit concept drift and huge volumes of data. We use the Massive Online Analysis (MOA) to perform the experiments since it is suitable for huge streaming data associated with concept drift. Support Vector Machines is used as the base learner since it provides a clear decision boundary that can easily be increased or decreased to reduce generalisation error. Generalisation error is computed using the hold out samples. The proxy parameter is varied and experiment repeated for every possible value. We ensured that there is a reasonable spread of diversity values ranging from 0.05 to 0.25 and the diversity was independent of training error within the specified range but equated to the margin value. Support Vector Machines is used as a base learner because its learning rate is not dependent on a parameter such as learning rate for all experiments. The algorithm learns at different rates depending on the amount of data and noisiness of data previously encountered.

5 Results

We show results on both the Waveform dataset and the Covertype dataset. Overlapping features are varied in order to properly investigate how new concepts encountered affects the performance of the algorithm. We plot the performance of the algorithm and make comparisons with other existing state of the art algorithms that exploit the aspect of diversity to optimize generalisation performance when learning big streaming data associated with concept drift.

The following plots indicate the performance of AceBoost against other two state of the algorithms on the Waveform dataset.

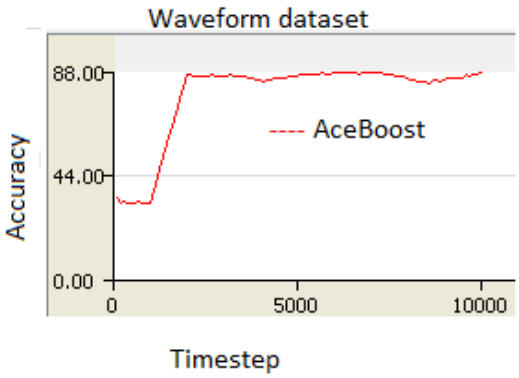

Figure 1: Average accuracy achieved by AceBoost on the Waveform dataset

For the synthetic dataset, Waveform, AceBoost performs well as evidenced by the accuracy obtained. In stable environments, the algorithm maintains its predictive accuracy. When the change of concept is severe, the algorithm learns with low diversity first and increases the diversity levels in subsequent steps. The effect of diversity on streaming data is evident. As the diversity is varied, generalization error is significantly reduced and adaptation to new concepts is optimized.

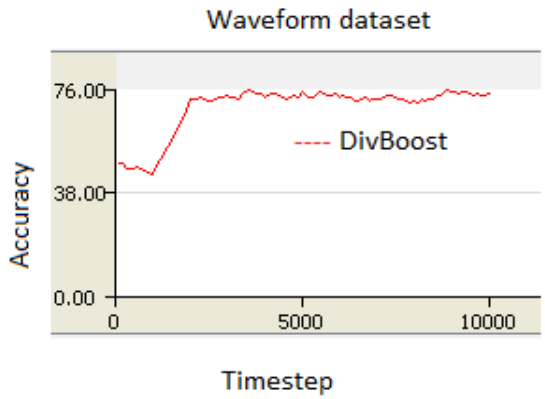

Figure 2: Average accuracy achieved by DivBoost on the Waveform dataset

At each step of the learning process, DivBoost struggles to achieve stability. Adjusting the margins using the diversity measure limits its rate of adaptation to new concepts.

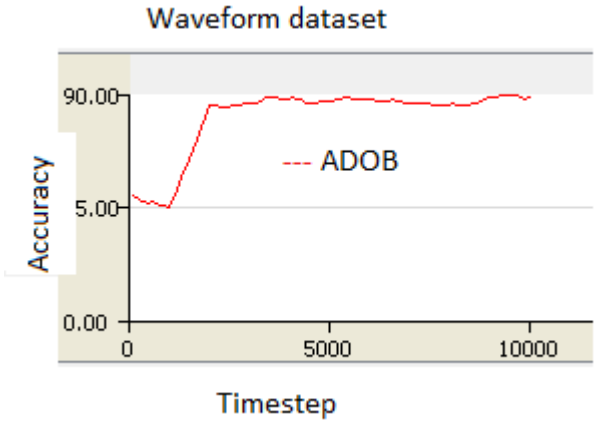

Figure 3: Average accuracy achieved by ADOB on the Waveform dataset

The ADOB algorithm performs comparably well despite slow adaptation to new concepts. The algorithm recovers slowly each time it encounters a new concept. The affinity of a diversity to the margin affects the performance of the algorithm.

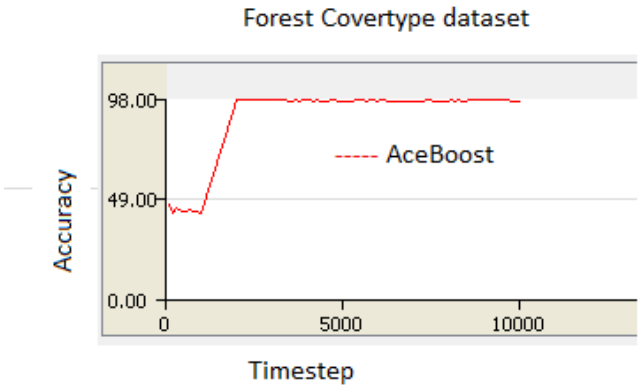

Figure 4: Average accuracy achieved by the AcBoost algorithm on the Forest Covertype dataset

Each time AceBoost encounters a new concept, it selects an ensemble that matches the current concept making adaptation to new concepts faster than anticipated. Regardless of the type of drift currently occurring, AceBoost varies the ensemble diversity to adapt accurately to the type of drift currently occurring.

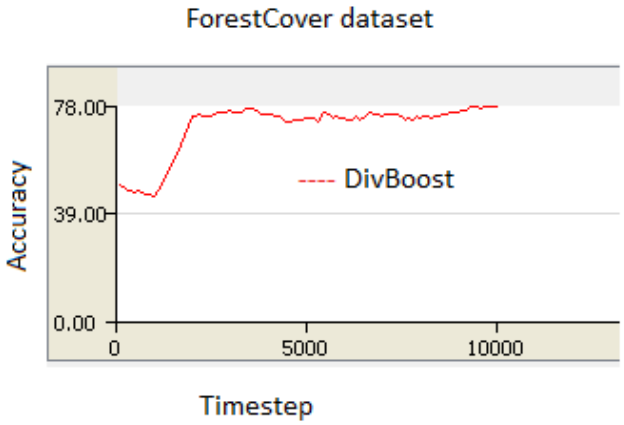

Figure 5; Average accuracy achieved by DivBoost on the Forest Covertype dataset

The DivBoost algorithm performs comparably well to other state of the algorithms although it shows instability and slow recovery and adaptation to new concepts. The use of a diversity measure not strongly linked to the margin makes the recovery to new concepts slow and the process of enlarging the margin to increase generalization performance is hindered greatly.

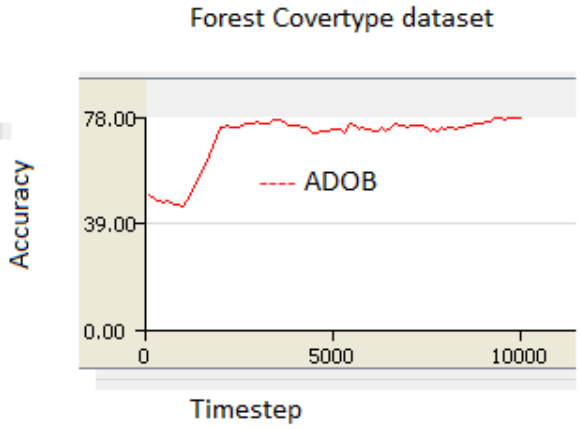

Figure 6: Average accuracy achieved by ADOB algorithm on the Forest Covetype dataset

The average accuracy of the ADOB algorithm on the Forest Covertype dataset reached 77.40%. The algorithm adapts to new concepts slowly and is unstable in situations where the severity of change is significant. The algorithm recovers slowly from a drift as the volume of the data increases, predictive accuracy of the algorithm decreases.

5.1 Tables

The table below provides a summary of the performances of the algorithms on both the synthetic and real world dataset.

Table 1: Accuracy of the algorithms on both synthetic and real world datasets

Benchmark dataset	AceBoost	DivBoost	ADOB
Waveform	84.30	73.70	81.80
Forest Covertype	98.50	97.80	95.60

6 Conclusions

In this paper, we presented the Adaptive Classifier Ensemble Boost (AceBoost), a new ensemble algorithm that exploits diversity to learn big data in dynamic environments associated with drifting concepts. Since the algorithm is based on the premises of the equality that exists between maximizing the margin and increasing ensemble diversity, generalisation performance for classification is easily optimized by increasing or decreasing the voting margins. AceBoost uses the Double Fault measure as the diversity measure and the values are controlled through a proxy parameter. This makes the algorithm suitable to adapt to any type of drift encountered in huge streaming data. The results achieved by AceBoost indicate that we could benefit from an ensemble classifier designed to learn big data by varying the amounts of diversity. The performance of AceBoost on big streaming data wast compared with other two state of the algorithms. AceBoost performed comparably well against the two state of the algorithms. The problem of learning big data in nonstationary environments associated with concept drift can be solved using machine learning methods such as ensembles that can vary their diversity levels. In some scenarios, low diversity ensembles typically adapt accurately on new concepts and high diverse ensembles always adapt faster. For future work, various diversity measures can be tested their suitability and their link to voting margins.

References

L.I. Kuncheva, “Classifier ensembles for detecting concept drift in streaming data: An Overview and perspectives” , Supervised and Unsupervised Ensemble Methods and their Applications, Volume 2, pages 5-9, 2008

Michael J.Procopio, “ An experimental analysis of classifier ensembles for learning drifting concepts”, PhD, University of Florida, 2007.

Thomas G. Ditterich, “Machine Learning Research”, AI Magazine, 18(4), pages 97-136, 1997.

Richard Stapenhurst, Gavin Brown, “Theoretical and Empirical Analyis of Diversity in Non-Stationary Learning”, IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments, Paris, France, April 2011.

L.Kuncheva, C.Whitaker, “Ten measures of diversity in classifier ensembles: Limits for two classifiers” In Proceedings of the IEE Workshop on Intelligent Sensor Processing, pages 1-10, 2001.

E.K Tang, P.N. Suganthan, X.Yao, “An analysis of diversity measures”, Machine Learning 65(2006), pages 247-271.

L.I Kuncheva, C.J Whitaker, “Measures of diversity in Classifier Ensembles and their relationship with the ensemble accuracy”, Machine Learning, Volume 51, pages 181-207, 2003.

Leandro Minku, X.Yao, “DDD: A New Ensemble Approach for dealing with concept drift” IEEE Transactions on Knowledge and Data Engineering, 24(4), pages 619-633, 2012.

Qinghua Hu, Leijun Li, Xiangqian Wu, Gerald Schaefer, Daren Yu, “Exploiting diversity for optimising margin distribution in ensemble learning”, Knowledge Based Systems, Volume 67, pages 90-104, 2014.

Joao Gama, Pedro Medas, Gladys Castillo, Pedro Rodriguez, “Learning with Drift Detection”. Brazilian Symposium on Artificial Intelligence, pages 286-295, 2004.

Haixum Wang, Wei Fan, Philip S.Yu, Jiawei Han, “Mining Concept-Drifting Data Streams using Ensemble Classifiers”, In KDD’03: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pages 226-235, 2003.

A.Tsymbal, “The problem of concept drift: definitions and related work”, Department of Computer Science, Trinity College Dublin, Ireland, Technical Report TCD-CS-2004-15, April 2004.