AFRL-AFOSR-JP-TR-2017-0039

Learning in the context of distribution drift

Geoff Webb
MONASH UNIVERSITY

05/09/2017
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
11-05-2017	Final	23 Apr 2015 to 22 Apr 2017

4. TITLE AND SUBTITLE
Learning in the context of distribution drift

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-15-1-4007

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Geoff Webb

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MONASH UNIVERSITY
WELLINGTON RD
CLAYTON, 3800 AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2017-0039

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The increasing ubiquity of data and its ever-increasing use to deliver tangible value raises the need for ever more effective technologies for data analysis. Many online data sources are subject to distribution drift: the frequency of different factors and the relationships between them change over time. This is problematic for machine learning because almost all algorithms assume that distributions are constant. This project investigates new technologies for learning in the context of distribution drift, guided by the insight that different subgroups will change in different ways, at different speeds and at different times. The results are leading towards robust and reliable data analytics, able to make more effective use of big data under real-world conditions of change.

The key developments in this project have been the creation of:
- a sound and applicable theoretical framework for analyzing concept drift,
- efficient and effective techniques for analyzing, understanding and describing concept drift observed in real world data,
- efficient and effective algorithms for learning from time varying data sequences,
- efficient and effective algorithms for classifying high-dimensional data,
- efficient and effective algorithms for handling ordinal data, and
- efficient and effective algorithms for learning in the context of concept drift

These new algorithms and techniques greatly improve the communitys capacity to learn under the demanding circumstances of concept drift.

15. SUBJECT TERMS
Machine Learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			KNOPP, JEREMY
Unclassified	Unclassified	Unclassified	SAR	17	**19b. TELEPHONE NUMBER** *(Include area code)* 315-227-7006

Final Report for AOARD Grant FA2386-15-1-4007

"**Learning in the context of distribution drift**"

5 May 2017

Name of Principal Investigators (PI and Co-PIs): Geoffrey Webb

- e-mail address : geoff.webb@monash.edu
- Institution : Monash University
- Mailing Address : Faculty of Information Technology, 25 Exhibition Walk, Monash University, Victoria 3800, Australia.
- Phone : +61 3 99053296
- Fax :

Period of Performance: 04/23/2015 – 04/22/2017

Abstract:

The increasing ubiquity of data and its ever-increasing use to deliver tangible value raises the need for ever more effective technologies for data analysis. Many online data sources are subject to distribution drift: the frequency of different factors and the relationships between them change over time. This is problematic for machine learning because almost all algorithms assume that distributions are constant. This project investigates new technologies for learning in the context of distribution drift, guided by the insight that different subgroups will change in different ways, at different speeds and at different times. The results are leading towards robust and reliable data analytics, able to make more effective use of big data under real-world conditions of change.

The key developments in this project have been the creation of

- a sound and applicable theoretical framework for analyzing concept drift,
- efficient and effective techniques for analyzing, understanding and describing concept drift observed in real world data,
- efficient and effective algorithms for learning from time varying data sequences,
- efficient and effective algorithms for classifying high-dimensional data,
- efficient and effective algorithms for handling ordinal data, and
- efficient and effective algorithms for learning in the context of concept drift

These new algorithms and techniques greatly improve the community's capacity to learn under the demanding circumstances of concept drift.

Introduction:

Industry, commerce, government, and science increasingly rely on big data to support their core processes [1]. Machine learning is one of the primary means of interpreting this data. However, the world is dynamic, in a constant state of flux. Despite this changing environment, conventional machine learning algorithms derive static models from historical data. When these fixed models of the past are used to predict the future, their failure to address change severely degrades performance. For example, Google Flu Trends was a major initiative that could accurately estimate the current incidence levels of flu on a region to region basis by learning from Google search query data [2]. This enabled public authorities, other organisations and the public to respond to each flu season in real time rather than waiting the 2 weeks it takes to compile official flu incidence reports. This system

was remarkably accurate throughout the period from 2009 to 2013, but in 2014 became wildly inaccurate, at one stage estimating that flu incidence was twice the level that official statistics subsequently revealed it to have been [3]. Due to this remarkable decline in performance, Google decommissioned this extremely valuable public resource. The cause of its malfunction was change – flu strains change, resulting in changes to the signs and symptoms suffered and the subsequent searches conducted. Search behaviours change as the public become more expert users of online search. Public health interventions are also dynamic. For example, actions such as retargeted vaccination campaigns will affect how a given season's flu epidemic unfolds.

In general, such change can arise from many factors. These are as varied as price growth due to inflation, the needs of an aging population, ebbs and flows in fashion, global warming, and equipment wear and tear. Improvements in technology and standards result in ever improving precision in measurements and change the systematic errors that inevitably affect recorded values. Some changes are cyclic or seasonal, while others progress primarily in a single direction. The technical name for observations that are subject to such change is a *non-stationary distribution.*

The change in a distribution that results from it being non-stationary is called *concept drift.* To maintain or improve accuracy in the face of this phenomenon, models must be continually revised or replaced.

To conquer concept drift it is essential first to understand it. The previous state-of-the-art analysis of concept drift resorted to subjective qualitative labels, such as *abrupt* and *gradual* [4]. However, to enable detailed analysis of concept drift it is essential to provide quantitative measures instead. Otherwise interpretation remains subjective, depending on arbitrary choices such as a threshold value t for determining whether to label a given drift 'abrupt' or 'gradual.' Only quantitative measures allow analysis of how learning mechanisms are affected as factors such as drift magnitude increase or decrease. In this project, we have defined the first such generic measures (Webb, et. al., 2016). These include the key measures of drift *magnitude*, *duration* and *rate*.

While there is a long history of research on concept drift generally [4, 5], armed with our new quantitative tools for describing and analysing concept drift, this project has investigated these phenomena for the first time using reliable approaches founded in rigorous theory.

We have made fundamental advances across a wide range of technical challenges that must be overcome if we are to conquer the problem of learning in the context of concept drift. These advances can be summarized as follow:

- development of a sound and applicable theoretical framework for analyzing concept drift,
- development of efficient and effective techniques for analyzing, understanding and describing concept drift observed in real world data,
- development of efficient and effective algorithms for learning from time varying data sequences,
- development of efficient and effective algorithms for classifying high-dimensional data,
- development of efficient and effective algorithms for handling ordinal data, and
- development of efficient and effective algorithms for incremental learning in the context of concept drift.

We describe below the methods used and key outcomes in each of these areas.

Methods and outcomes

We describe approaches taken under each of the major areas outlined above.

Development of a sound and applicable theoretical framework for analyzing concept drift

To master concept drift it is first necessary to understand it. This requires conceptual tools for defining and analyzing it. The previous state-of-the-art was qualitative analysis, describing drift in terms such as *gradual* and *abrupt*. We have argued for a more powerful framework to support quantitative description and analysis.

This framework defines a concept as a probability distribution in effect at a given time and quantifies concept drift using two primitives. The first is *magnitude*, which can be instantiated by any standard measure for measuring distance between probability distributions. The second primitive is *duration*, which is instantiated as elapsed time. This provides the first quantitative framework for drift analysis. A detailed theoretical analysis of this problem has been published in the leading data mining journal, *Data Mining and Knowledge Discovery* (Webb et. al., 2016)[1].

We have shown that the previous qualitative descriptors of drift are subjective, reliant on arbitrary thresholds such as the maximum duration over which a drift can occur while still being considered abrupt. Perhaps even more critical is that these qualitative descriptors rely on an implicit assumption that real world domains have periods with no drift and that drift occurs for discrete periods between these periods of stability. Such an assumption appears implausible in many real-world contexts.

In contrast, our quantitative measures are objective and make no assumptions as to whether domains will ever be without drift.

We have demonstrated that our quantitative measures can reveal insights that are impossible to derive from qualitative measures, such as that some learners recover from abrupt drift in a fixed period of time irrespective of the magnitude of that drift. This is illustrated in Figure 1, taken from (Webb et. al., 2016), which shows the error over time of a Hoeffding Tree when it is exposed to drift of differing magnitudes at time point 100,000. The larger the magnitude of the drift the larger the immediate spike in error, but irrespective of the magnitude of the drift the error converges to the same rate at time step 200,000. This observation is impossible to derive without a quantitative notion of drift magnitude.

This new theoretical framework is informing the technologies that we are developing to understand and manage concept drift.

Development of efficient and effective techniques for analyzing, understanding and describing concept drift observed in real world data

We have developed a suite of techniques that utilize our quantitative measures of drift to provide useful descriptions of the drift inherent in sample data. This has required overcoming significant obstacles. The first of these results from our proof that drift magnitude increases monotonically with dimensionality. That is, each time a new variable is added to data subject to drift, the magnitude of the drift increases. This means that for high dimensional data, drift magnitude will tend toward 1.0 due to accumulation of minor quantities of drift across every dimension. This holds for both of the metrics that are commonly used to measure distances between distributions – *Hellinger Distance* and *Total Variation Distance*.

Further, a single value expressing total drift magnitude provides only a very gross description of a complex drift phenomenon. It fails to recognize or to describe the details of how drift differs across the subspaces defined on different variables of the data. In the real world, drift is often not uniform. For example, not all factors are subject to inflation and those that are may inflate at varying rates. In many real world applications, it is likely

[1] We refer to papers included in the List of Publications using the APA name, date format. We refer to papers by others using the number format.

to be useful to understand which variables and combinations of variables are drifting in which manners at any particular time.

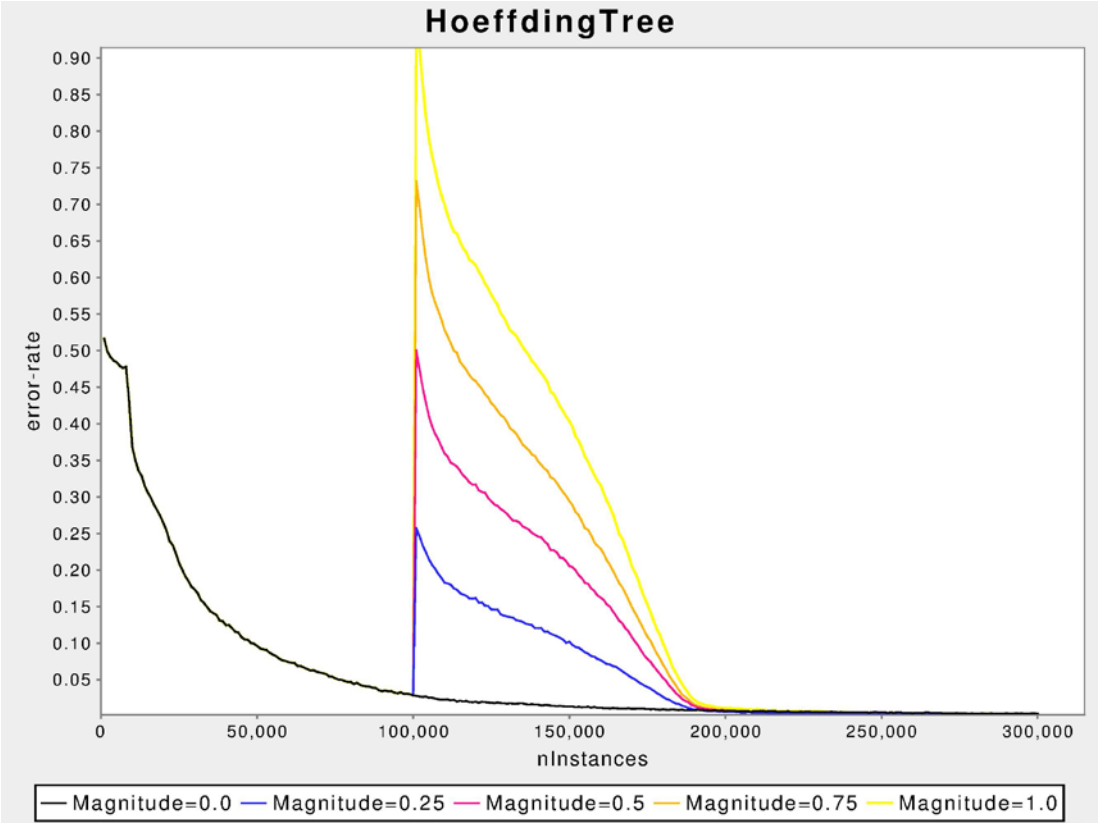
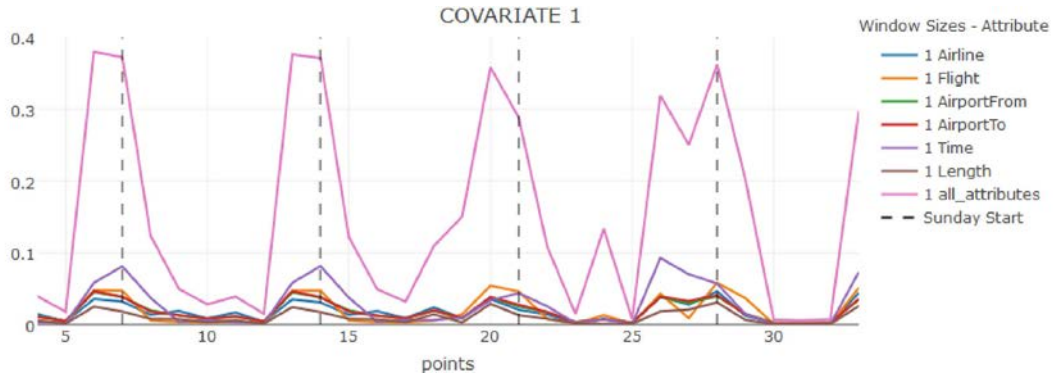


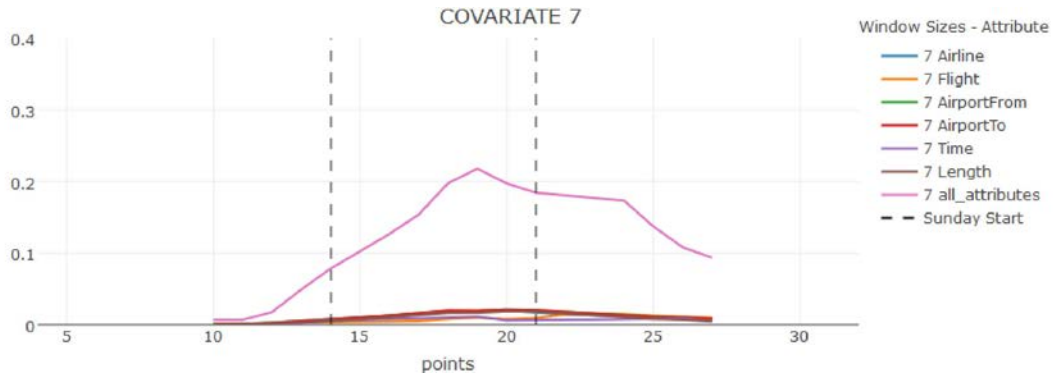
Figure 1. Illustration that Hoeffding Trees converge to the same error rate at the same time irrespective of the magnitude of an abrupt drift to which they are subjected.

For these reasons, we investigate the introduction of concept drift maps, methods for describing the drift affecting different subspaces of the data. Details are provided in Webb, *et. el.* (2017). Figures 2a and 2b (copied from Webb, *et. el.*, 2017) show drift maps showing the drift of each single variable (covariate) and for all variables (labelled *all_attributes* in the plots). Figures 2c and 2d show drift maps for the class. The maps in Figures 2a and 2c measure drift between successive days while the maps in Figures 2b and 2d measure drift between successive 7 day periods. The data in question is the benchmark airlines dataset. The drift magnitude is measured in Total Variation Distance in these and all the following examples, but the results are very similar if the other suitable metric, Hellinger Distance is used in its place.

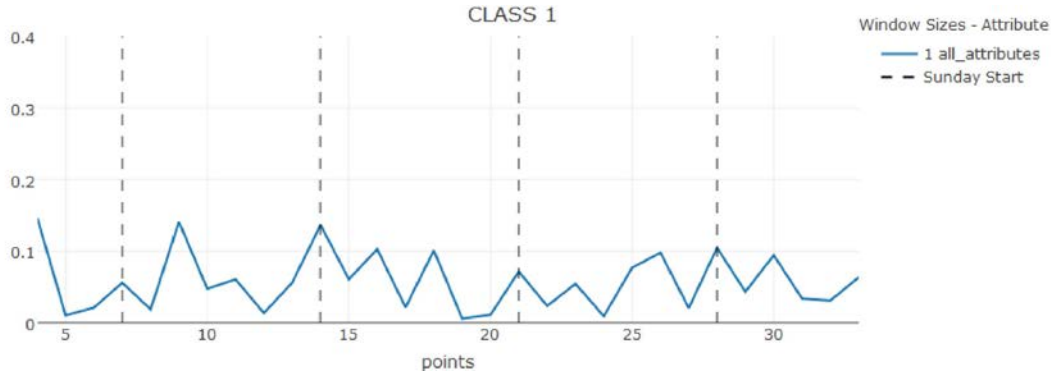
Figure 2a shows that there is initially a clear weekly cycle to the drift, with great change between the week and the weekend, and less change day to day during the week. Figure 2b shows that the regularity of the first two weekly cycles is due to stability between the weeks – there is little drift between the first two weeks but as the daily pattern breaks down the inter week drift increases. This illustrates that different insights are provided by different granularities of analysis. These two sub figures also illustrate how the drift across all variables is greater than the drift across any of the individual variables.



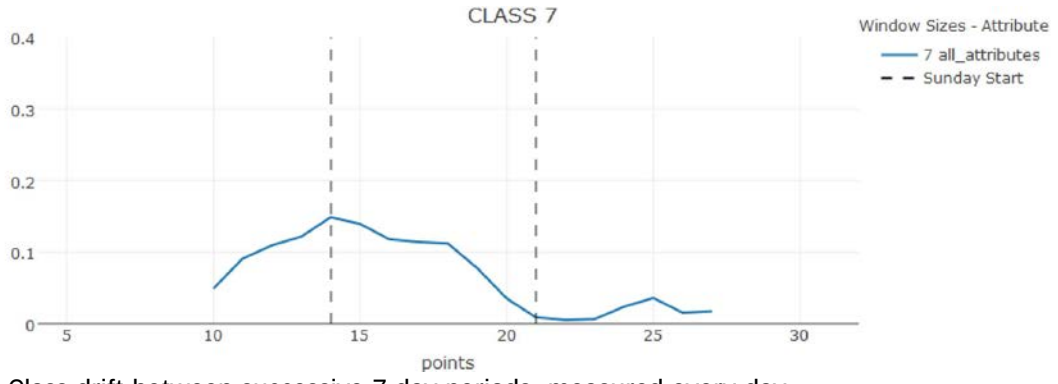
(a) Covariate drift between successive days, measured every day



(b) Covariate drift between successive 7 day periods, measured every day



(c) Class drift between successive days, measured every day



(d) Class drift between successive 7 day periods, measured every day

Figure 2: Drift in each individual variable (covariate), all variables, and the class for the airlines dataset

The class for this dataset is whether a flight arrives on time or late. It can be seen that there is little correlation between the drift in the class and the drift in the covariates. Indeed, the inter-week drift follows the reverse pattern, being high between the first three weeks and reducing between the last two weeks. This may be quite important for a data analyst to understand, as different drift mechanisms may best address each of drift in the covariates and drift in the class.

Figure 3 shows a heatmap of the pairwise drift in the joint distribution on the Landsat-8 French land usage satellite data. This data represents 10 meter square areas in a satellite image. The class, *id*, describes the land usage of that location and the covariates (also called *variables* or *attributes*) contain the various direct and derived measures from the satellite image. The drift presented here is between 5 May 2013 and 29 November 2013.

Each cell in the heat map, except the diagonals, show the magnitude of the marginal drift for two variables between the two periods. For example, the top-left cell shows the Total Variation Distance between the joint distribution of values of NDWI and band1 on 5 May and the joint distribution of those values on 29 November.

These dates in Spring and Fall were chosen as ones between which there should be expected to be substantial changes. May is generally just before the harvest of winter crops, e.g. wheat, canola and barley. These drift magnitudes reveal that this results in substantial drift in the satellite images.

The diagonal in Figure 3 represents univariate drift. For example, the cell at the intersection of the row and column labelled *id* gives the magnitude of the drift for the class variable, *id*.

Figure 3: Pairwise drift in the joint distribution for the satellite data.

As the land usage assigned to each point does not change over the period, the drift magnitude is 0.0. The largest univariate drift is for band 5, which corresponds to near-infrared. This is explained by the fact that chlorophyll reflects near-infrared; in May, a lot of surfaces are covered by growing crops, which leads to a large amount of near-infrared being reflected. On the other hand, most crops have been harvested late November. More generally, it can be seen that each of these univariate drifts is lower than any of the bivariate drifts involving that same variable, as our monotonicity proof demonstrates they must.

Development of efficient and effective algorithms for learning from time varying data sequences

One of the problems confronting learning in the context of concept drift is how to make use of information about the way the world is changing. For example, with the Landsat data discussed above, the task is to identify land usage from satellite images. However, it is not possible to distinguish many crops, such as wheat from corn, from a single image alone. Both will appear similar when the crop is ripening, and when the crop has been harvested. It is only by analysis of the development of the crop over time that it is possible to distinguish the two.

We have worked on two types of analysis for time varying data. The first is highly scalable techniques for classification of time series data (Petitjean et. al., 2016). Nearest neighbor classification under the Dynamic Time Warping (DTW) distance measure is the most accurate known approach to classification of many time series problems. However, it is extremely computationally demanding.

Our new techniques exploit the recent development of methods for generating an average time series from a group of time series under DTW. This enables standard clustering algorithms to be adapted to clustering of time series under DTW.

Our new learning method involves first clustering the time series belonging to each class and then creating a new dataset containing the average time series for each cluster. This can be used to reduce the original dataset to any target size. This reduced dataset can then be used for nearest neighbor classification, speeding the nearest neighbor classification process by many orders of magnitude with only minor losses in accuracy. Such speed up is essential to scale time series classification from the current benchmark datasets used in the field containing at most few thousands of series to modern applications such as satellite land usage which require classification of many billions of series in near real time.

The second type of analysis is pattern discovery. This can be important for understanding how data is evolving over time. We have developed a new measure of how surprising a sequential pattern is given example data and efficient algorithms for finding patterns that maximize that measure. This measure differs from previous measures by considering all partitions of the sequential pattern into subsequences. For example, suppose that buying a shirt is often followed by buying a tie and that buying shoes is often followed by buying socks. Then all of the following sequences should appear frequently, <shirt, tie, shoes, socks>, <shirt, shoes, tie, socks>, <shirt, shoes, socks, tie>, <shoes, shirt, tie, socks>, <shoes, shirt, socks, tie > and <shoes, socks, shirt, tie>. However, unless one or more of those six sequences is more frequent than the others, none of them should be considered interesting as they are just a consequence of the two original sequences. Our new measure is the difference between the frequency of a pattern and the maximum for any partition of the sequence into two subsequences of the average frequency of all rearrangements of those subsequences. We show that this approach is very effective at identifying surprising sequential patterns.

However, a naïve approach to applying it would be extremely computationally expensive, as it would require creating all arrangements of all partitions of every pattern considered. We have developed a variant of the OPUS search algorithm that can search this large space

of potential patterns very efficiently (Petitjean, et. al., 2016).

Development of efficient and effective algorithms for classifying large quantities of data

One of the main contexts in which it is important to take account of concept drift is for online learning, where new data are constantly arriving and it is desirable to continually update the learned model. Many online learning applications involve massive data streams. The quantity of data is very large and new data arrives with high frequency. In consequence, highly efficient algorithms for learning from large quantities of data are required. In this line of research, we have investigated how to best achieve such processing while minimizing any negative impact on accuracy.

The high-level strategy we followed was to develop methods for coupling the computational efficiency of generative learning with the accuracy of discriminative learning. We developed two major lines of investigation. The first learned discriminative models, but used generative parameters learned with minimal computation to both speed up the discriminative learning process and to regularize the discriminative models towards more sensible defaults then the arbitrary 0 or 1 usually used in regularization. Thus, we can learn slightly more accurate models with much less computation (Zaidi et. al, 2016; Zaidi, Petitjean, & Webb, 2016; Liu & Zaidi, 2016; Zaidi et. al., in press).

The second strategy creates large families of nested generative models, such that the computation of each model involves just a minor addition to the computation of another model. Discriminative leave-one-out cross validation is then used to very efficiently choose the best performing of all this large number of models. The result is very efficient discriminative selection between a large range of efficient generative models (Martinez et. al, 2016; Chen et. al., 2016; Chen et. al., 2017).

Development of efficient and effective algorithms for handling ordinal data

It is critical that learning algorithms be robust to vagaries in how data are represented. For example, the fuel efficiency of a vehicle might be expressed as Miles per Gallon or as Gallons per Mile. There is a linear relationship between the number of miles travelled for each gallon and the first of these measures, but a nonlinear relationship with the second. Hence, any linear classifier will necessarily learn different models depending on which representation is used. This is undesirable given that they both express the same underlying phenomenon and it is arbitrary which expression is employed. As linear classifiers are extremely efficient and efficiency is important in online learning, we have developed robust techniques for addressing this problem by creating efficient distance measures that are invariant under monotone transformations of numeric variables. That is, they make no stronger assumption than that the data are ordinal. These measures use forests of stochastic trees developed from extremely small samples to assess similarity between items. Details can be found in Fernando & Webb (2017).

Development of efficient and effective algorithms for learning in the context of concept drift

Changing data distributions imply that a learner should pay greater attention to more recent examples and less attention to older examples. A number of mechanisms have been developed for this purpose. *Windows* maintain only a fixed number of the most recent examples [6]. *Aging*, *discounting*, or *decay* maintains a weight for each example which decreases over time, thus giving more weight to recent examples and less weight to older examples [6]. An alternative strategy is to *regularise* the learner in such a way that the learned model tends back toward a prior state unless new examples reinforce elements that have been previously learned [7]. For ease of discussion we will refer to these three mechanisms as *data treatments* and refer to a small window or high decay or regularisation as a *short-term sample* and a larger window or lower decay or regularisation rate as a *longer-term sample*.

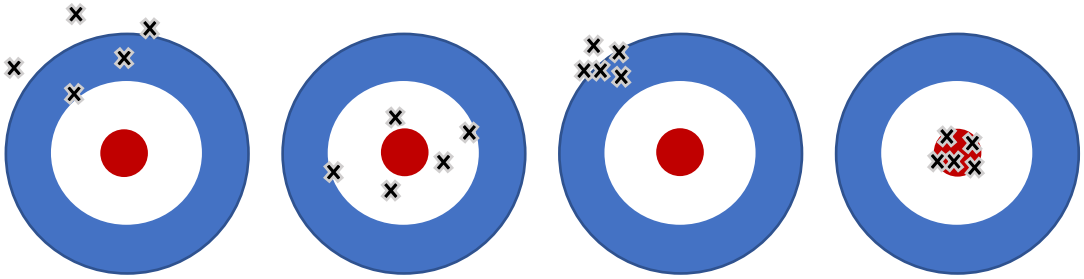


Figure 4: Bias and variance in shots at a target.

Distribution drift may vary from rapid to slow and may be interspersed with periods of stability. If distribution drift is fast we want to use a short-term sample, as the historical information is no longer relevant and we do not want to be misled by it. In contrast, if distribution drift is slow or the distribution is stationary then we want to learn from a longer-term sample, as we want to collect as much detailed information about the current situation as we can and the long term historical information is relevant and hence should ideally be brought to bear. Numerous schemes have been developed for detecting the rate of concept drift and adjusting the learning system in response [8].

The error of a learning system can be decomposed into two components, bias and variance [9]. These measure how the predictions of learned classifiers differ when the learner learns from different data samples. Bias relates to the distance of the central tendency of the predictions to the true value. Variance relates to the dispersion of predictions. Figure 4 illustrates this concept using an analogy with target shooting, where the centre of the target represents the true value and the 'shots' represent predictions made by classifiers learned from different samples. When either bias or variance is high, error will be high.

The theoretical foundations of this project are underpinned by the concept of bias and variance and the *low bias hypothesis* [10]. To illustrate this hypothesis, consider Figure 5, which shows a typical pair of learning curves for two classification learning algorithms, in this case naïve Bayes [11] and k-Dependence Bayes (KDB) [12]. A learning curve plots the

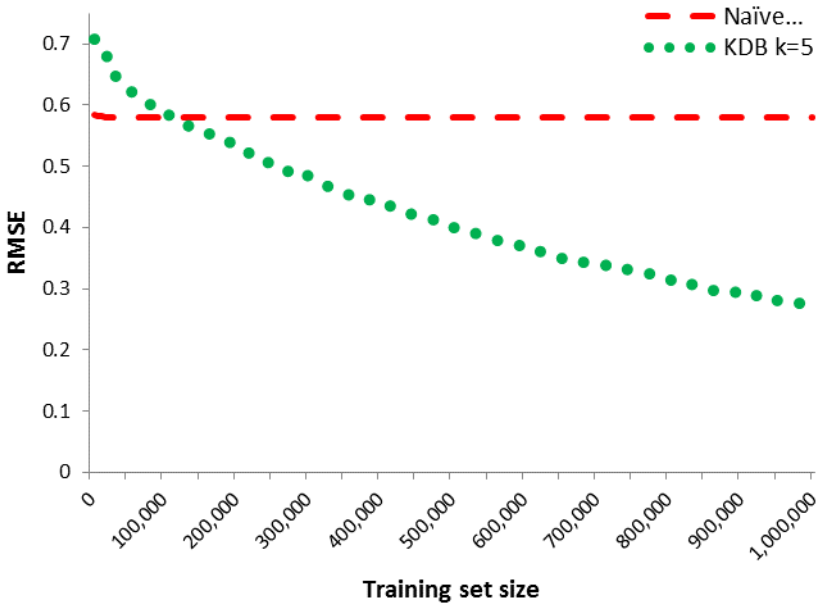


Figure 5: Learning curves for naïve Bayes and k-Dependence Bayes on the poker-hand dataset. *This illustrates how low bias algorithms, such as k-Dependence Bayes, which can create very detailed models of the data often overfit small data but are more accurate on large data than low variance algorithms that cannot form such detailed models.*

error of the learned classifier on previously unobserved data as the quantity of training data from which it learns increases. Typically, as the training data quantity increases, error decreases. However, different algorithms have different learning curves. Naïve Bayes typically has lower error than KDB when trained on small data, but soon the curves cross and KDB has much lower error on big data. This is due to the relative expressiveness of the classifiers that the two algorithms learn. Naïve Bayes learns a very simple form of classifier that can only accurately describe distributions in which the variables are conditionally independent of one another. In contrast, KDB can accurately describe a broad spectrum of types of multivariate distribution. Naïve Bayes is more accurate for small training data because it cannot overfit the data. It is incapable of creating overly detailed classifiers based on too little evidence. In contrast, KDB is more accurate for big data because it can more accurately describe the intricate multivariate relationships that the data reveal.

This well understood phenomenon is known as the *bias-variance* trade-off [9]. Algorithms that can accurately describe a wide range of distributions, such as KDB, typically have low bias but high variance, while algorithms such as naïve Bayes have high bias but low variance.

The low bias hypothesis [10] holds that low bias algorithms will derive the most accurate classifiers when the training data quantity is high while low variance algorithms will be most accurate when data quantity is low.

APPROACH

The low-bias hypothesis has profound implications for how classifiers should best handle distribution drift. If the rate of drift is high we want a data treatment that provides a short-term sample. In consequence, we should use a learner with low variance and high bias. On the other hand, if the rate of drift is low we want a data treatment that provides a longer-term sample and so we should use a learner with low bias and high variance which can best take advantage of the more detailed information available in large data. This is illustrated in Figure 6 which shows the high-bias low-variance Naïve Bayes (NB), mid-bias

and variance A1DE and low-bias high-variance A2De with varying decay rates on the benchmark electricity dataset. The lower the variance of the algorithm, the higher the decay rate at which it achieves its lowest error. This complex task has rapid drift (Webb, et. el., submitted for publication) which is why the very simple NB is able to perform relatively well with a high decay rate.

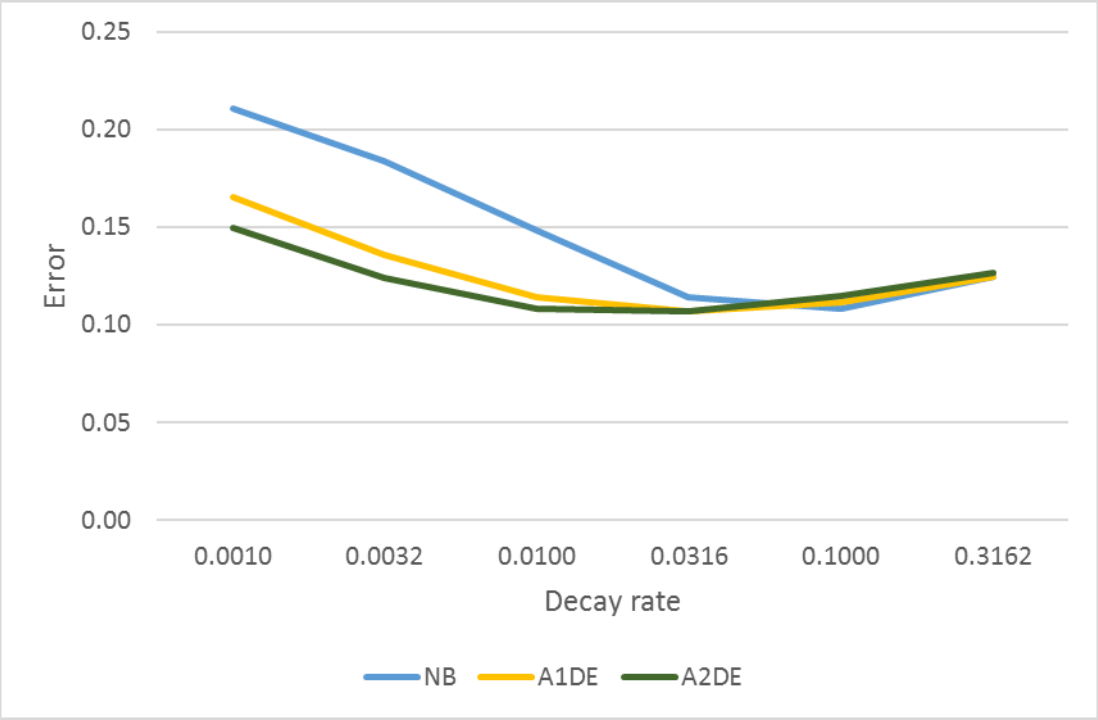


Figure 6: Error of NB, A1De and A2DE with varying decay rates on the electricity dataset. *The lowest bias algorithm (A2DE) performs best with low decay rates when it can learn complex interactions from long-term data while the lowest variance algorithm performs best with high decay rates where it can adjust quickly to drift as it occurs.*

This interaction between bias-variance performance, desirable sample size and drift rate implies that when learning in the presence of variable or unknown rates of distribution drift we want to be able to access multiple configurations of data manipulation and learner. This is illustrated in Figure 7.

We have developed a novel learning system based on this architecture that we call *Stacked Bayesian Network Classifiers with varying rates of decay*. The learners of varying levels of variance are taken from the AnDE family of classifiers. These classifiers are incremental, allowing models to be refined as new data is obtained, and hence well-suited to streaming data. A single parameter n controls a bias/variance trade-off. Choosing n=0 gives NB with low variance but high bias. Increasing values of n decreases bias at the expense of increased variance.

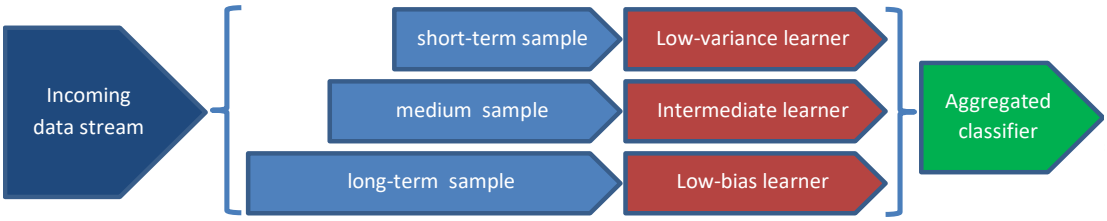


Figure 7: Architecture for learning from streaming data in the context of variable or unknown distribution drift. *The incoming data are processed in parallel by multiple learning configurations that each match data manipulations and learning capabilities to differing possible rates of drift. We illustrate here three configurations, but any number might be used in practice.*

When applied to classify an object, the base ANDE classifiers each output a predicted probability for each possible class. These predicted probabilities from all base classifiers are then passed to a logistic regression model which combines them to make a final prediction. The logistic regression model is learned incrementally using stochastic gradient descent. This allows it to dynamically adjust in response to concept drift, paying more attention to the outputs of the base models that are currently most accurate and less to those that are less accurate.

With this configuration, the combined classifier usually attains error close to the best of any single AnDE classifier with any decay rate, and often attains lower error than any single AnDE classifier across a wide range of drift scenarios. This is because of the manner in which it can interpolate between the base classifiers that are performing best at any particular time and hence can adjust effective decay rate and bias-variance profile as drift alters.

This Stacked ANDE classifier obtains error rates that are very competitive with all state-of-the-art alternatives on both real world and synthetic datasets containing concept drift. Figures 8 and 9 illustrate this on the benchmark real world airlines and electricity datasets. The algorithm is more flexible that the previous state-of-the-art – able to respond consistently to both gradual and rapid drift of a wide variety of types and is always competitive with the best of the previous state-of-the-art for each specific type of drift.

A paper is currently in preparation.

Summary

We have developed a wide-range of sophisticated algorithms for learning in the context of distribution drift. We have developed a new theoretical framework for analyzing distribution drift. This has been critical, as previous techniques were qualitative and did not support objective assessment of differences between cases of drift or of relative performance of alternative algorithms under different drift scenarios. We have new effective techniques that can analyze real-world data streams and create useful descriptions of the drift within them. This is critical if people are to understand the problems they confront in a specific application in order to plan how best to tackle it. We have also created new algorithms for learning from the way in which data drifts over time. As a result, drift need not always be just an obstacle to learning, it can also be an aide. To handle the computational demands of rapid learning in complex data we have developed new efficient and effective algorithms for classifying high-dimensional data. These are useful not only in the computationally demanding context of stream learning but also allow more complex classifiers to be constructed in conventional batch learning than might otherwise be feasible.

Finally, we have combined all of the above techniques to create new algorithms based on the insight that different forms of drift are each best addressed by using different windows of past data for learning and that different window sizes require learners with different bias-variance profiles. Fast drift requires that models are only learned from recent data and slow drift allows the use of more historical data to refine the models that are learned. When only small amounts of recent information is used, error will be minimized by a learner with low variance, while when more historical data is used, error will be minimized by a learner with low bias. We have built a proof of concept learner that demonstrates these principals and which shows the direction for powerful flexible learning systems of the future.

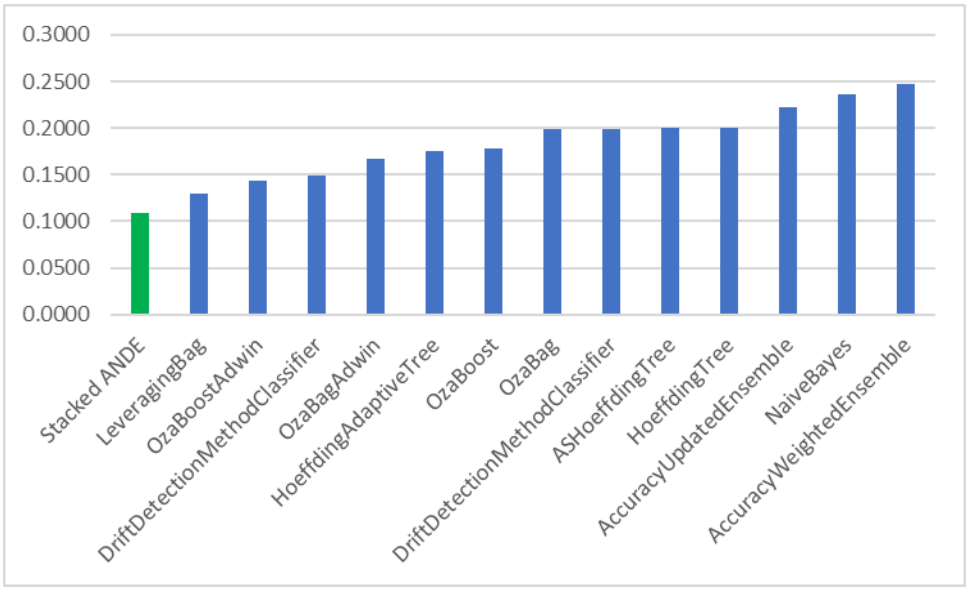


Figure 8: Error of Stacked ANDE and state-of-the-art competitors on the benchmark electricity dataset.

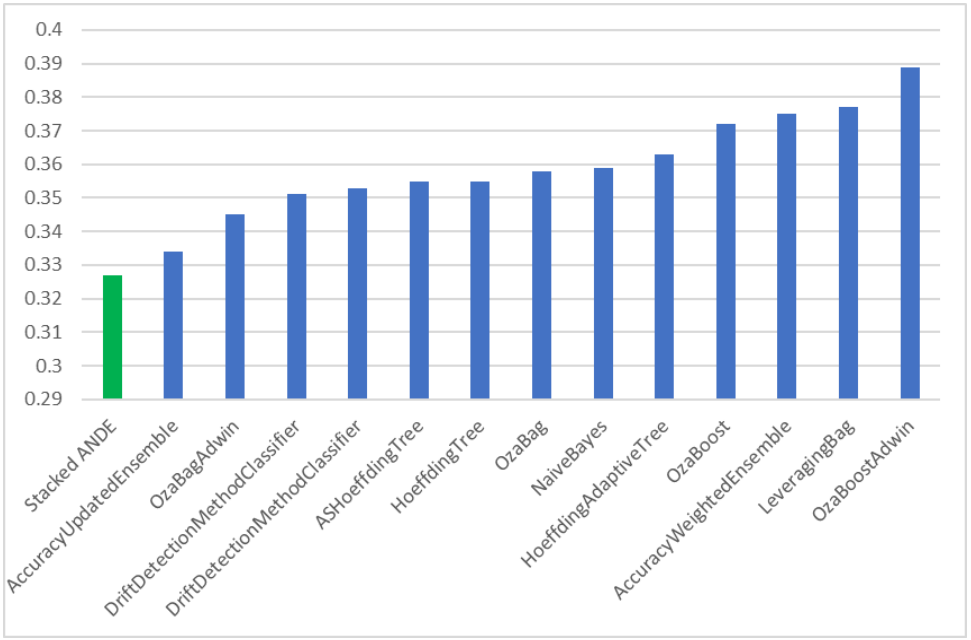


Figure 9: Error of Stacked ANDE and state-of-the-art competitors on the benchmark airlines dataset.

References

Note, these are references to work by others. Our work is referenced using the APA name, date format, and the papers are listed under the List of Publications.

1. White, C., *Using Big Data for Smarter Decision Making*. 2011, BI Research: Ashland, Or.

2. Cook, S., et al., *Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic.* PloS one, 2011. **6**(8): p. e23610.

3. Lazer, D., et al., *The parable of Google Flu: traps in big data analysis.* Science, 2014. **343**(14 March).

4. Gama, J., et al., *A Survey on Concept Drift Adaptation.* Acm Computing Surveys, 2014. **46**(4).

5. Sammut, C. and M. Harries, *Concept Drift*, in *Encyclopedia of Machine Learning*, C. Sammut and G.I. Webb, Editors. 2010, Springer: New York. p. 202-205.

6. Dietterich, T.G., *Machine learning for sequential data: A review*, in *Structural, syntactic, and statistical pattern recognition*. 2002, Springer. p. 15-30.

7. Hosmer Jr, D.W. and S. Lemeshow, *Applied Logistic Regression*. 2004: John Wiley & Sons.

8. Schaul, T., S. Zhang, and Y. LeCun, *No more pesky learning rates.* arXiv preprint arXiv:1206.1106, 2012.

9. *Bias-Variance Decomposition*, in *Encyclopedia of Machine Learning*, C. Sammut and G.I. Webb, Editors. 2010, Springer: New York. p. 100-101.

10. Brain, D. and G. Webb, *On the effect of data set size on bias and variance in classification learning*, in *Proceedings of the Fourth Australian Knowledge Acquisition Workshop (AKAW99)*, D. Richards, et al., Editors. 1999, University of New South Wales: Sydney. p. 117-128.

11. Lewis, D.D., *Naive Bayes at forty: The independence assumption in information retrieval*, in *Proceedings of the Tenth European Conference on Machine Learning (ECML-98)*. 1998, Springer: Berlin. p. 4-15.

12. Sahami, M., *Learning limited dependence Bayesian classifiers*, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. 1996, AAAI Press: Menlo Park, CA. p. 334-338.

List of Publications and Significant Collaborations that resulted from your AOARD supported project:

a) papers published in peer-reviewed journals

Fernando, T. L., & Webb, G. I. (2017). SimUSF: an efficient and effective similarity measure that is invariant to violations of the interval scale assumption. *Data Mining and Knowledge Discovery*, **31**(1), 264-286.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). *Characterizing Concept Drift. Data Mining and Knowledge Discovery*, **30**(4), 964-994.

Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., & Keogh, E. (2016). Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. *Knowledge and Information Systems*, **47**(1), 1-26.

Zaidi, N. A., Webb, G. I., Carman, M. J., Petitjean, F., & Cerquides, J. (2016). ALRn: Accelerated higher-order logistic regression. *Machine Learning*, **104**(2), 151-194.

Petitjean, F., Li, T., Tatti, N., & Webb, G. I. (2016). Skopus: Mining top-k sequential patterns under leverage. *Data Mining and Knowledge Discovery*, **30**(5), 1086-1111.

Martinez, A. M., Webb, G. I., Chen, S., & Zaidi, N. A. (2016). Scalable Learning of Bayesian Network Classifiers. *Journal of Machine Learning Research*, **17**(44), 1-35.

Chen, S., Martínez, A. M., Webb, G. I., & Wang, L. (2016). Selective AnDE for large data learning: a low-bias memory constrained approach. *Knowledge and Information Systems*, **50**(2), 475-503.

Chen, S., Martinez, A., Webb, G., & Wang, L. (2017). Sample-based Attribute Selective AnDE for Large Data. *IEEE Transactions on Knowledge and Data Engineering*, **29**(1), 172-185.

Zaidi, N., Webb, G. I., Carman, M., Petitjean, F., Buntine, W., Hynes, H., & De Sterck, H. (in press). Efficient Parameter Learning of Bayesian Network Classifiers. *Machine Learning*.

b) papers published in peer-reviewed conference proceedings

Zaidi, N. A., Petitjean, F., & Webb, G. I. (2016). Preconditioning an Artificial Neural Network Using Naive Bayes. *Proceedings of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2016*, pp. 341-353.

Liu, N., & Zaidi, N. (2016) Artificial Neural Network: Deep or Broad? An Empirical Study. *AI2016: Advances in Artificial Intelligence*, pp. 535-541.

c) papers published in non-peer-reviewed journals and conference proceedings

d) conference presentations without papers

e) manuscripts submitted but not yet published

Webb, G. I., Lee, L. K., Petitjean, F., & Goethals, B. (submitted for publication). *Understanding Concept Drift*. https://arxiv.org/abs/1704.00362

f) provide a list any interactions with industry or with Air Force Research Laboratory scientists or significant collaborations that resulted from this work.

Attachments: Publications a), b) and c) listed above if possible.