Analysis of Concept Drift and Temporal Inductive Transfer for Reuters2000

George Forman
Hewlett-Packard Labs
1501 Page Mill Rd.
Palo Alto, CA 94304
ghforman@hpl.hp.com

Abstract

The success of machine learning classification pales for real-world, time-varying streams of data. We define three subtypes of concept drift, and confirm that *recurrent themes* appear in the benchmark dataset Reuters2000. To encourage research in this difficult area, we define a 'daily classification task' (DCT) problem formulation, in which a few random iid training samples are provided each day. Ideally, past training data could be leveraged to improve the current day's classifier. Empirical results for Reuters2000 show that two likely methods are not successful: (1) the popular idea of a sliding window incorporating recent past training data, and (2) inductive transfer of the previously learned classifiers to provide additional predictive features for the current learning task. The former provides a method of characterizing the degree of concept drift. The latter excels if all *past* labels are given: 'hindsight DCT.'

1 Introduction

Machine learning research typically assumes training cases are random samples, independently and identically distributed (iid) from a stationary test distribution. In contrast, commercial applications of machine learning often desire to apply trained classifiers to make predictions on a stream of future samples that may vary over time—for example, determining whether news articles belong to a specific topic, such as sports. Unfortunately, in the real world things change, and successful text classification methods fail when applied to time-varying samples. Despite the difficulty, this is nonetheless an economically important problem to tackle.

We subdivide the notion of concept drift into three common types:

1. Shifting Class Priors: the class distribution may shift over time, but the samples within a given class are iid stationary. For example, the proportion of Hepatitis A cases may increase with an epidemic, but the symptoms of the disease are invariant over time. For a robust method to track shifting class priors with limited training data, see [1].

2. Shifting Subclass Priors: each class is a union of (undiscovered) subclasses or themes, and the frequency distribution of these subclasses may shift over time. As above, the feature distribution given a subclass is stationary, but the feature distribution of the super-class will vary over time.

3. Fickle Concept Drift: individual cases may change their ground truth label over time. This setting is appropriate for recommender systems—the user may initially find some case relevant that is later not relevant. This appears to be the most difficult setting of class drift. If some assumption can be placed on how slowly or suddenly concepts shift, one may have some notion of how prior training labels may still be useful. In general, old training data is no better than unlabeled samples.

A variant on each of these types is when the domain has some recurrent or even periodic behavior. For example, in spam classification, there is a periodic theme of Christmas-related spam every December. We performed an analysis of the top 100 most predictive words for each week of the Reuters2000 benchmark (806,791 topic-labeled news articles over 365 days [2]) and determined that top words that stop being predictive after a few weeks often re-surface in the top 100 later on. We hypothesize that these represent *recurrent themes* of type 2 concept drift: recurrent shifting subclass priors, where we do not know the subclasses. It is this type of concept drift we attempt to address in this paper.

2 Daily Classification Task (DCT)

Concept drift is admittedly a difficult research area. To promote its study, we define a conducive problem formulation we call the *Daily Classification Task* (DCT). In it, time is discretized into periods, e.g. days, and of the many cases each day, a limited size random iid sample is provided as a labeled training set. Classification accuracy or F-measure is computed on each day of a benchmark dataset, and the average is reported over *all* days; this gives a natural incentive for methods that improve quickly with only a few past days available (rather than begin the average after day 200, for example). For research purposes, the size T of the daily training set should be selected so that the learning curve is still climbing. Later we also consider a variant task, *hindsight DCT*, which provides some or all class labels for past test cases—a reasonable assumption for certain real-world settings.

The strawman is simply to train a state-of-the-art classifier on each daily training set, and classify that day's cases. To surpass this baseline, we would like to leverage past training data somehow. An obvious idea is to use a sliding window that retains data from the most recent W days, including today's training data. We refute this popular idea with an empirical evaluation on Reuters2000 in Section 4.

3 Temporal Inductive Transfer Method (TIX)

Ideally we should be able to leverage the learned models from the past. This is the challenging domain of inductive transfer, here applied temporally to a single classification task to help cope with concept drift, specifically recurring themes.

We propose the following method: The most recent L previous daily classifiers are applied to all the new day's cases. Their predictions provide L additional features that may be potentially predictive in the daily learning task. If a news theme reoccurs that was popular <L days ago, a newly trained classifier may be able to leverage the predictions made by the old classifiers that were trained while the theme was previously popular.

Since each daily classifier depends on previous classifiers, the recurrence relation implies that all classifiers remain in use for all time. Intelligent pruning may someday be devised, but for this purpose of this paper, we preserve only the L most recent daily strawman classifiers, which operate independently of one another.

4 Empirical Evaluation

We conducted a series of experiments in the DCT formulation on Reuters2000. We focused on the binary classification task of predicting which of each day's news articles belong to the GCAT topic (government/social, ~30%). Each day we made available T=100 training cases. Rather than test on the many thousands remaining, we considered only the first 400 articles of each day, selecting T of these at random for training. We report the F-measure (macro-) averaged over all days.

The base classifier we used is a linear support vector machine (SVM) trained on binary bag-of-words features, as implemented by the WEKA library (title+body text lowercased, alpha only, max 50K words from each training set). In TIX, the L additional features are each binary outputs of past strawman classifiers.

For comparison, we also evaluated an 'oracle' variant, in which we let the L past classifiers train on the entire day's data, including the testing data. We also step them forward by one day. Thus, the first of the L additional features is a prediction based on the entire day's data as training. It does not provide a perfectly predictive feature, since the SVM does not memorize its training set, as kNN would. The final SVM trains on only the T=100 training cases, plus the L additional features.

4.1 Results

Figure 1a shows the average F-measure for strawman: simply train on the random T samples each day and test on the rest of the day's cases. This establishes a baseline performance of 76.1% F-measure for T=100, used hereafter. This graph confirms that the choice of T=100 is sufficient for some learning to occur, but not so much that additional data or features could provide no benefit.

Figure 1b shows the results for the sliding window technique, as we vary the window size W days. Thus, the training set contains W*T recent cases. Strawman corresponds to W=1. For larger W, we see that increasingly stale training data misleads the classifier badly. However, the decline with increasing W for this method can be used as a way to characterize the pace of concept drift in a dataset.

Figure 2a shows the results of the TIX model as we vary the number L of past classifiers retained. It nearly matches the performance of the strawman. Unfortunately, the predictions by the L past classifiers were apparently not valuable; we confirmed this by observing small weights given to these additional features. The figure also shows the oracle model, which rises with L. This shows that past classifiers, when trained with more data, begin to provide useful predictive features for the daily learning task. This validates the concept of temporal inductive transfer, although in this setting the inaccuracy of the past classifiers trained on only T=100 cases gives no better information than just examining the current text training data.

This finding inspired the *hindsight DCT* setting, in which all past labels become available. Figure 2b shows the performance of the TIX model for this task. Its performance improves with increasing L and approaches that of the oracle. Note that inductive transfer from L>=16 past days outperforms the oracle for L=1, which uses *all* of the current day's labels but ignores the past.

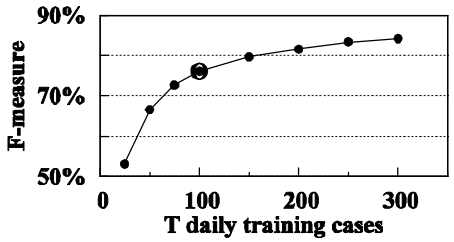

Figure 1a: Strawman

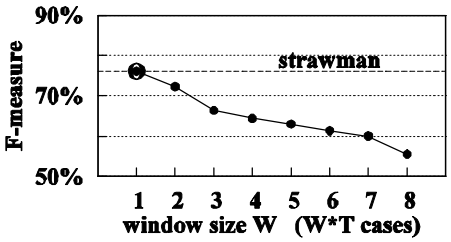

Figure 1b: Sliding window

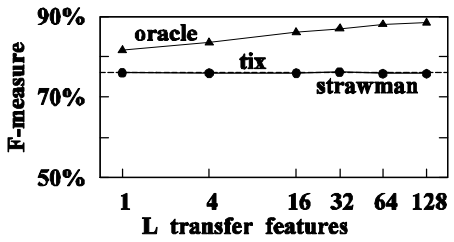

Figure 2a: Inductive transfer

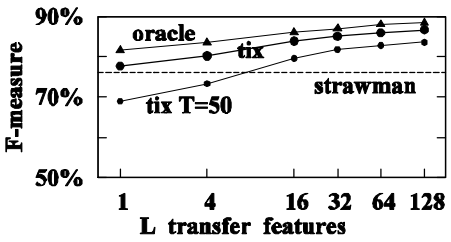

Figure 2b: Same, with hindsight

5 Discussion & Conclusion

We have shown the success of temporal inductive transfer for the hindsight DCT setting, which exposes the answer key for all past data. While useful in some real-world situations, in many others the past labels are not available without great expense. Thus, further research is called for in the pure DCT setting.

It is disappointing that TIX provided no benefit in the pure DCT setting. Recall that for this study, each past classifier did not build on previous classifiers, in order to break the recurrence. If they had instead each leveraged previous classifiers, one possibility is that they would then perform better, providing more predictive features. We are skeptical of this: since the outputs of strawman performed equal to those of the L=1 TIX classifier, there would presumably be no difference in performance if we were to replace the past strawman classifiers in TIX with the daily TIX classifiers that are trained with the additional L=1 feature. Even for L=1, the recurrence would extend to all previously learned classifiers.

A more promising avenue for future work includes hybridizing the temporal inductive transfer idea with related work in semi-supervised learning. The past labeled data provides for a richer setting than traditional semi-supervised learning. Other related work includes methods such as Stacking, and sequential prediction methods, which assume some predictive value in the specific sequence of cases.

Finally, although we showed that increasing the sliding window size W hurt performance, we acknowledge that in settings without DCT time discretization, some sliding window scheme may be needed to select a sample of recent data. Our results suggest strong pressure to minimize W in this case.

References

[1] George Forman. (2005) Counting Positives Accurately Despite Inaccurate Classification. *Proc. of the European Conf. on Machine learning* (ECML), pp. 564—575.

[2] D. Lewis, Y. Yang, T. Rose, F. Li. (2004) RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5(Apr):361—397.