Estimating the Expected Effectiveness of Text Classification Solutions under Subclass Distribution Shifts

Nedim Lipka, Benno Stein
Bauhaus-Universität Weimar
99421 Weimar, Germany
Email: lipka.nedim@gmail.com, benno.stein@uni-weimar.de

James G. Shanahan
The Church and Duncan Group, Inc.
San Francisco, CA 94131, USA
Email: james.shanahan@gmail.com

***Abstract*—Automated text classification is one of the most important learning technologies to fight information overload. However, the information society is not only confronted with an information flood but also with an increase in "information volatility", by which we understand the fact that kind and distribution of a data source's emissions can significantly vary. In this paper we show how to estimate the *expected effectiveness* of a classification solution when the underlying data source undergoes a shift in the distribution of its subclasses (modes). Subclass distribution shifts are observed among others in online media such as tweets, blogs, or news articles, where document emissions follow topic popularity.**

To estimate the expected effectiveness of a classification solution we partition a test sample by means of clustering. Then, using repetitive resampling with different margin distributions over the clustering, the effectiveness characteristics is studied. We show that the effectiveness is normally distributed and introduce a probabilistic lower bound that is used for model selection. We analyze the relation between our notion of expected effectiveness and the mean effectiveness over the clustering both theoretically and on standard text corpora. An important result is a heuristic for expected effectiveness estimation that is solely based on the initial test sample and that can be computed without resampling.

***Keywords*-classification; concept drift; unknown distributions; model selection; clustering**

I. Introduction

The reliable evaluation of a classification solution is a basic necessity, be it for model selection purposes or for assessing the effectiveness with respect to a given task. Today's information filtering and retrieval tasks that deal with class label prediction in volatile environments render an evaluation more complicated. Large parts of statistical evaluation and machine learning research rely on the assumption that the provided as well as the future examples are independent and identically distributed (i.i.d.) with regard to the underlying probability distribution. It is known that this is often not the case in real-world scenarios, e.g., if examples from a time-varying stream are to be classified.

User-generated content on the Web such as news articles, blog posts, and tweets exhibit large variations of the underlying distribution characteristics; especially Twitter exemplifies the volatile nature of "trendy" topics, as illustrated by Liu et al. [1] with a new interactive visualization technique.

Forman [2] subdivides the phenomena of distribution changes over time, also known as concept drift, into three types:

(1) Class distribution shift: the sample of a class remains i.i.d., but the ratio between the classes varies.
(2) Subclass distribution shift: the sample of a subclass remains i.i.d., but the sample of the class and the classes overall does not.
(3) Fickle concept drift: the ground truth of the class labels changes.

The second type, a subclass distribution shift, models the dynamics in online media best and forms the basis for the paper in hand. This type is also known as covariate shift [3] if the shift occurs across subclass boundaries.

Our paper addresses the problem of *evaluating* classification solutions if subclass distribution shifts are likely to occur in the future and if one has no knowledge about how a shift will evolve. Research in semi-supervised learning, domain adaptation [4], and sampling bias correction are related to our problem, but the most approaches assume that knowledge of the target distribution (the distribution for which a classification solution is applied) is given. Other research, which disregards the target distribution, e.g. [2] and [3], focuses on machine learning within concept drifts and visualization.

We propose an evaluation framework that accounts for the nature of online media streams by estimating the expected effectiveness of a classification solution under subclass distribution shifts. In online media such as news streams, articles from the past remain retrievable while a new article substream emerges whenever a new topic becomes interesting. The evolution of the article distribution is hence not arbitrary, and one can expect density changes within local regions that lead to a subclass distribution shift.[1] Currently there is no means for a reliable evaluation nor for a model selection in this scenario, even if the changes of the target distribution are of a homogeneous nature.

[1] Also the growth rate of high density areas over the time is not random: Yang and Leskovec [5] studied the dynamics of attention of content within online media and identified six dominant temporal patterns.

Our evaluation framework adapts common statistical evaluation measures to estimate the expected effectiveness and to select the best classification solution in terms of the lowest worst case effectiveness. Given a classification solution m we partition the test sample and evaluate m on each partition. The partitioning is constructed by a clustering algorithm that identifies regions of similar examples. The examples in a cluster are likely to behave similarly, a fact which is known as cluster hypothesis in information retrieval: "closely associated documents tend to be relevant to the same requests". We create new test samples by varying the ratio between the clusters, and, based on these variations, we estimate the expected effectiveness. In addition, we consider the effectiveness of each cluster in isolation and study the effectiveness distribution over the clustering of the test sample. By assuming a constant effectiveness per cluster we derive a second statistics that approximates our idea of expected effectiveness under subclass distribution shifts. Our contributions comprise:

(1) Two statistics to assess the expected effectiveness of a classification solution under subclass distribution shifts.
(2) A probabilistic notion of the expected effectiveness for model selection.
(3) An empirical validation of the assertions of our model selection approach for different corpora.

II. Related Work

A concept drift is a change of the distribution of examples over time. Research in this field can be divided into concept drift detection and concept drift handling—with the objective to compile a machine learning algorithm that detects a drift and adapts to it. Concept drifts occur either gradually or abruptly and are empirically observable in labeled and unlabeled samples [6], e.g. by monitoring the prediction quality, the distribution, or clustering parameters such as densities, centers, or shapes. Vreeken et al. [7], for example, estimate the differences between two samples by techniques based on compression and covering characteristics, while Anderson et al. [8] use the distances between density estimates. Standard drift detection employs statistical hypothesis testing for the randomness of the samples, such as the Wald-Wolfowitz test or more advanced tests [9].

Concept drift handling has become an important research topic in recent years—the most popular machine learning methods are adapted to handle it, and theoretical results have also been extended to concept drift phenomena. Advanced window-based approaches are given in [10], [11], for example. The former paper proposes a window-based one-class ensemble, whilst the latter proposes a window-based ensemble for learning from positive and unlabeled data in order to accurately select and classify unlabeled examples for reuse. Huang [12] extends a sampling strategy for active learning by monitoring a possible concept drift in unlabeled data. Aggrawal et al. [13] present a classification method that adapts to changes of the underlying data stream by dynamically selecting an appropriate training sample, and Hulten et al. [14] focus on novel decision tree learning algorithms, where outdated subtrees are revisited and recreated.

III. Expected Effectiveness under Subclass Distribution Shifts

We define a *classification solution* m as a tuple of two functions: a model formation function α and a classifier h. A feature vector is denoted as $\mathbf{x}$; it represents a document d in text classification tasks. The model formation function $\alpha : d \rightarrow \mathbf{x}$ defines this representation and is an important factor within the process of building m. The statistical learning theory considers $\mathbf{x}$ as an instance of a real-valued multivariate random variable X, and the assigned class label y as an instance of a binary random variable $Y \in \{-1, 1\}$, governed by the joint probability distribution $P(X, Y)$. The classifier $h : \mathbf{x} \rightarrow y$, also called hypothesis, is provided by a machine learning algorithm or a domain expert; h predicts the class label of a given feature vector. We define m as tuple (α, h) in order to emphasize the fact that each classification solution has an underlying design process.

The performance of m's predictions is quantified by measures such as recall, precision, or accuracy. Without loss of generality we use the term *effectiveness* as a generic term for such measures and presume the range $[0; 1]$. A higher effectiveness $e(m, S)$ corresponds to a larger value of a measure when applied to a labeled test sample S, and hence to a better prediction quality of m.

It is assumed in this paper that the documents are emitted by stochastic processes. It is also assumed that each process is stationary and emits the documents of a subclass independently and identically distributed. Both assumptions qualify for many real-world classification problems. A subclass distribution shift occurs if the emission rates of the processes differ in the course of time. Note that under a subclass distribution shift, all measures that rely on the confusion matrix fail.

We introduce $\mathrm{E}_t[e]$, the *expected effectiveness* of a classification solution m under subclass distribution shifts, as the weighted average of the effectiveness that m achieves under all possible subclass distribution shifts. In situations where the development of the underlying distribution cannot be predicted, the expected effectiveness provides a sensible means for model selection. The exact computation of the expected effectiveness is not possible since the underlying emission processes cannot be controlled to produce all possible distribution shifts.

A. An Expected Effectiveness Estimate

We estimate the expected effectiveness $\mathrm{E}_t[e]$ of m under subclass distribution shifts by identifying subclasses of the underlying stochastic processes and by modeling different distribution shifts via resampling. We associate subclasses

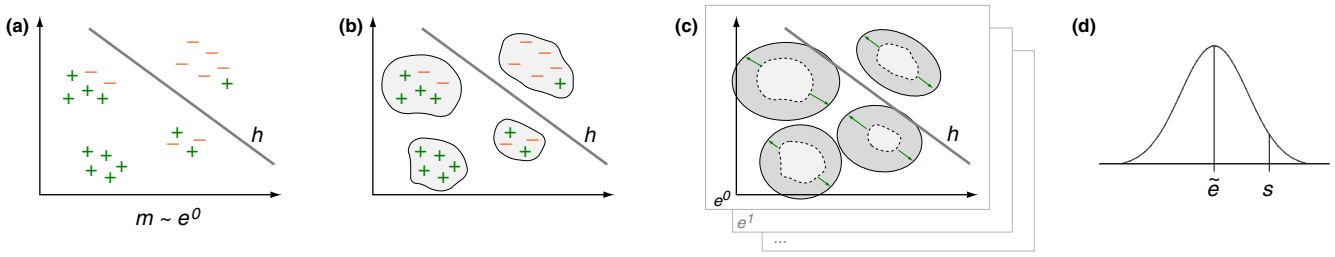

Figure 1. The estimation of the expected effectiveness of a classification solution m (with classifier h) happens in the following steps: (a) input for the estimation, which consists of m and a given test sample of positive and negative labeled examples, (b) cluster analysis of the test sample in order to identify the modes of the distribution, (c) variation of the cluster sizes and application of m, (d) sample statistics (mean $\widetilde{e}$ and standard deviation s) of the achieved effectiveness.

with the clusters of a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$, $C_i \subseteq S$, $i = 1, \ldots, k$, where $\mathcal{C}$ is an exclusive and complete partitioning of the feature vectors in the sample S. The difference between a clustering on the one hand and a categorization by humans on the other is that the latter is based on the interpretation of real-world objects, while a clustering analyzes densities (MajorClust), variances (k-means, Ward's method), or distributions (expectation-maximization clustering) of feature vectors. If the documents within a cluster are considered as realizations of a single stochastic process, it is likely that this process emits documents in a high-similarity region of the population.

Distribution shifts are modeled by resampling the documents within the clusters. An increase or decrease of the documents in a high-similarity region (as specified by a cluster) implies that the probability density function of the *global* probability distribution of documents and class labels will change. If, for example, the density values of the global probability density function increase inside a specific region, the density values outside will decrease due to normalization. In our considerations, we constrain the modeling of distribution shifts by preserving the local (cluster-specific) characteristics of the distribution. Stated another way, the probability distribution of a cluster C, $P_C(Y|X)$, remains i.i.d., and the distribution of clusters sizes varies.

Given a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$, let $\mathcal{S}$ be a set of samples where each $S \in \mathcal{S}$ is compiled by a unique weighting over $\mathcal{C}$:

$$S \;=\; sample(C_1) \cup \cdots \cup sample(C_k).$$

The estimate $\widetilde{e}$ for the expected effectiveness $\mathrm{E}_t[e]$ of m under subclass distribution shifts is defined as the sample mean over $\mathcal{S}$:

$$\widetilde{e} \;=\; 1/|\mathcal{S}| \sum_{S \in \mathcal{S}} e(m, S). \tag{1}$$

The sample variance s^2 of e can hence be written as:

$$s^2 \;=\; 1/|\mathcal{S}| \sum_{S \in \mathcal{S}} (e(m, S) - \widetilde{e})^2. \tag{2}$$

The symbol $\widetilde{e}$ is used instead of $\bar{e}$ to emphasize that the mean is computed from a specifically constructed set $\mathcal{S}$.

Figure 1 illustrates the estimation procedure:

(a) Classification solution m. Feature space and document representation are defined by the model formation function α; h is a classifier built by a learning algorithm.
(b) Clustering of the documents.
(c) A set of samples, each one constructed by scaling the number of documents in a cluster by resampling.
(d) Distribution of e over the set $\mathcal{S}$ of samples.

B. An Expected Effectiveness Heuristic

The estimate $\widetilde{e}$ of the expected effectiveness is based on a sufficiently large set $\mathcal{S}$ of samples. We now devise a second statistics $\widehat{e}$ for $\mathrm{E}_t[e]$, which considers only the characteristics of the clustering $\mathcal{C}$ of the test set S.

In this regard we evaluate for each cluster $C \in \mathcal{C}$, $|\mathcal{C}| = k$, its cluster-specific effectiveness e_C of m. Since the effectiveness is likely to be the same on similar documents, it can be assumed that m's effectiveness for a cluster C remains stable under a subclass distribution shift. This assumption relates to the clustering assumption in the field of semi-supervised learning: "If points are in the same cluster, they are likely to be of the same class."

The (overall) effectiveness e of m given S is the weighted sum of the cluster-specific effectiveness values:

$$e \;=\; w_1 e_{C_1} + \cdots + w_k e_{C_k},$$

where w_i is the weight given by the relative size of the cluster $|C_i|/|S|$. Using vector notation, with a positive real-valued weight vector $\mathbf{w}, |\mathbf{w}| = k$, and effectiveness vector $\mathbf{e} = (e_{C_1}, \ldots, e_{C_k})$, the effectiveness is $e = \mathbf{w}^T\mathbf{e}$, where $\mathbf{w}^T$ denotes the transpose of $\mathbf{w}$.

We now assume that the effectiveness for a cluster shows no variation at different points in time:

$$\mathbf{e} \;\equiv\; \mathbf{e}^{(0)} \;=\; \mathbf{e}^{(1)} \;=\; \ldots \tag{3}$$

Since this assumption depends on the degree to which the clustering assumption is fulfilled, we call the statistic $\widehat{e}$ for the expected effectiveness $\mathrm{E}_t[e]$, introduced below, a heuristic. Note that Assumption (3) does not imply a constant (overall) effectiveness e.

Under Assumption (3), the effectiveness e varies only with the change of the weights $\mathbf{w}$. We model the weight vector $\mathbf{w}$

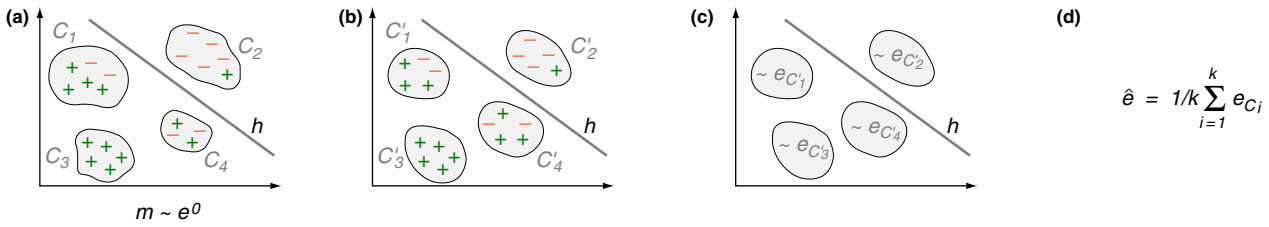

Figure 2. Heuristic estimation of the expected effectiveness of m: (a) application of cluster analysis to the test sample in order to identify the modes of the distribution (this corresponds to Step (b) in Figure 1), (b) adaptive sampling of the clusters in order to equalize different cluster sizes that may be given in the sample, (c) application of m for each cluster in isolation, (d) mean $\widehat{e}$ of the cluster-specific effectiveness values.

as a k-dimensional random variable $W = (W_1, \ldots, W_k)$. Without knowledge about future topic distributions, our risk minimization strategy is to consider all possible vectors $\mathbf{w}$ as equally likely, whereas the L^1-norm of $\mathbf{w}$ is always one: $\sum_{i=1}^{k} w_i = 1$. As a consequence, the vectors $\mathbf{w}$ lie on a simplex and W is Dirichlet distributed, $W \sim Dir(\alpha)$, with concentration hyperparameter $\alpha = (1, \ldots, 1)^T$, $|\alpha| = k$. The resulting mean and variance are:

$$\begin{aligned} \mathrm{E}[W_i] &= \frac{\alpha_i}{\alpha_0} = 1/k, \quad \text{where } \alpha_0 = \sum_{j=1}^{k} \alpha_j, \\ \mathrm{Var}[W_i] &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} = \frac{k-1}{k^2(k+1)}. \end{aligned}$$

Let e' denote the random variable $W_1 e_{C_1} + \cdots + W_k e_{C_k}$. The heuristic $\widehat{e}$ of the expected effectiveness $\mathrm{E}[e]$ of m under subclass distribution shifts is defined as the mean of e':

$$\begin{aligned} \widehat{e} &= \mathrm{E}[W_1]\, e_{C_1} + \cdots + \mathrm{E}[W_k]\, e_{C_k} \\ &= 1/k\ e_{C_1} + \cdots + 1/k\ e_{C_k} = 1/k \sum_{i=1}^{k} e_{C_i}. \end{aligned} \quad (4)$$

Due to the central limit theorem, e' is normally distributed and its variance is:

$$\mathrm{Var}[e'] = \frac{k-1}{k^2(k+1)}.$$

$\mathrm{Var}[e']$ is independent of the cluster-specific effectiveness values and hence constant for all classification solutions. Therefore, this variance cannot be exploited for model selection purposes. We conclude that the difference between $\mathrm{Var}[e']$ and the sample variance s^2 of e (Equation 2) reflects to what extent Assumption (3) is violated, say, to what extent the effectiveness within the clusters is not constant.

Assumption (3) is strict, but not unrealistic, and we can show in the experimentation section that our heuristic has in practice only a small approximation error (about 3%) for estimating the expected effectiveness under a subclass distribution shift.

Figure 2 illustrates the computation of the heuristic:

(a) *Input.* A classification solution m along with a clustering $\mathcal{C}$ of the test sample with k clusters.

(b) *Adaptive Sampling.* The k clusters are resampled to the same size, $|C'_1| = \cdots = |C'_k|$, to get better effectiveness estimates for each cluster.

(c) *Effectiveness Estimation.* For each cluster the effectiveness $e_{C'}$ of m on C' is estimated.

(d) *Output.* $\widehat{e}$, the mean of the $e_{C'}$, which represents a heuristic estimate of the expected effectiveness under distribution shifts in the clustering (Equation 4).

C. Model Selection via Expected Effectiveness

Given a clustering $\mathcal{C}$ of a test set S, model selection means to choose a classification solution m from a set of solution candidates M. If the expected effectiveness $\mathrm{E}_t[e]$ is approximated under Assumption (3) as $\widehat{e}$, the model selection problem can be tackled by choosing the model with the highest $\widehat{e}$. If the expected effectiveness is approximated via the estimation procedure (Figure 1) as $\widetilde{e}$, additional model selection information in form of a probabilistic lower effectiveness bound Θ can be provided.

We present such a lower bound Θ to show that the effectiveness of classification solution m is with a probability of $1 - \delta$ larger than Θ, if the subclass distribution varies. The effectiveness e is normally distributed, see Section III-B, with approximated mean $\widetilde{e}$ and variance s^2 (Equations (1) and (2)). We can estimate the parameters of the normal distribution for each solution in M and infer Θ with the inverse of its cumulative distribution function, also known as quantile function:

$$\begin{aligned} F^{-1}(\delta;\, \mu, \sigma^2) &= \mu + \sigma\sqrt{2}\ \mathrm{erf}^{-1}(1 - 2\delta) \\ \Theta &= F^{-1}(\delta;\, \widetilde{e}, s^2), \end{aligned} \quad (5)$$

where F^{-1} denotes the quantile function of the normal distribution for $1 - \delta$, and erf^{-1} denotes the inverse error function. If δ is chosen to be 0.0228, the value of Θ is $\widetilde{e} + 2s$. Figure 3 shows an example for the accuracy Acc.

While the expected effectiveness estimate is useful to select the classification solution with the best expected effectiveness, the probabilistic lower bound is useful to select the solution that minimizes the risk of an *effectiveness drop* in the wild.

IV. Experimentation

This section validates our theoretical findings and demonstrates the use of the expected effectiveness concept. We

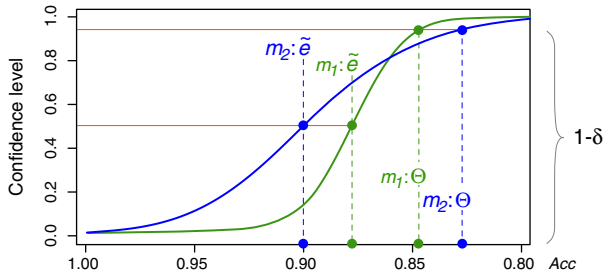

Figure 3. Comparison of classification solution m_1 ($\widetilde{e} = 0.88, s = 0.02$, green line) to classification solution m_2 ($\widetilde{e} = 0.90, s = 0.04$, blue line) with respect to their accuracies. At confidence level 0.5 solution m_2 has a higher expected accuracy than m_1. However, at confidence level $\delta = 0.05$, the lower bound (worst case) Θ under m_1 is better than Θ under m_2.

show for different corpora, classifiers, and measures, that

(1) the expected effectiveness under subclass distribution shifts, $\mathrm{E}_t[e]$, is normally distributed,
(2) the expected effectiveness can be applied as a probabilistic lower bound for model selection,
(3) the heuristic $\widehat{e}$ for assessing the expected effectiveness is a tight estimate.

In order to demonstrate the evaluation of classification solutions under subclass distribution shifts we focus on standard text classification corpora and standard machine learning algorithms. As text classification task the topic categorization problem is studied: given a set of topics, assign an unseen document to one of these topics.

A. Classification Solutions and Corpora

As mentioned at the outset, classification solutions combine a model formation function α and a classifier h. We vary the range of solutions by employing several machine learning algorithms, while α remains unchanged. The model formation function α is the following: The documents of each considered corpus are represented under a vector space model with term frequencies, whereas the dimensions correspond to the stemmed alphabetic words that occur at least 10 times. The Lovins stemmer is employed as stemming technology, the vectors are normalized. In order to run a vast amount of experiments within a reasonable time, 2 500 words with the highest information gain scores remain after further processing, while the scores are evaluated only on the training examples with discretized word frequencies. The following machine learning algorithms are employed in the experiments to learn h: linear support vector machine, naïve Bayes, C4.5 decision tree, and k-Nearest Neighbor.

The experiments are conducted on standard corpora: Reuters (RVC1), the Open Directory Project (ODP), and 20 Newsgroups (20NG), which are the most frequently used data sets in the field of topic categorization. A preprocessing of the corpora restricts them to documents that are uniquely classified and that have a minimum size of 1 kB. We construct binary classification tasks by selecting two categories instead of pursuing a one-versus-all strategy. From the large number of tasks that have been considered in our experiments we will report results for the task *Science versus Sports*, since these categories occur across all corpora.

B. Cluster Analysis

The comparison of different clustering algorithms is beyond the scope of this paper. The appropriate choice depends on the application domain and the concrete classification task, and it cannot be expected that a single algorithm is consistently the best choice (no free lunch). Within our experiments the k-means algorithm is applied, which yields an non-overlapping (exclusive) clustering, i.e., $\forall_{i,j,i\neq j} C_i \cap C_j = \emptyset$. We set k to $\sqrt{|S|/2}$ and use Lloyd's algorithm for approximation.

For the RCV1 we explicitly show that the setting is able to identify appropriate subclasses: opposed to ODP and 20NG, the RCV1-corpus provides time stamps. Moreover, by visualizing the change of the top 100 most predictive words over time, the authors of [2] gave evidence for a subclass distribution shift in RCV1. With statistical randomness tests, we empirically validate the two main properties of subclass distribution shifts:

(1) the overall samples at different time stamps *are not* i.i.d. according to the same distribution, but
(2) the samples of single subclasses *are* i.i.d.

The most practical randomness tests operate on binary sequences, which are often constructed by dichotomizing a sequences of continuous values. For a sample S we consider the sequences of the Euclidean distances $d(\mathbf{x}^{(i)}, \mathbf{x}^{(i-1)})$ for $i = 2 \ldots |S|$, where $\mathbf{x}^{(i-1)}$ is the chronological predecessor of $\mathbf{x}^{(i)}$. A sequence of distances that is i.i.d. according to an unknown distribution indicates a process of data generation that can produce i.i.d. samples. As an illustrative example consider a random process that first produces fairly similar news articles on politics and after a while articles on sports. The corresponding (non-random) sequence of distances is a series of small distances followed by a large distance when the emission of sports articles begins. As a consequence, the samples drawn at different points in time are not representative for the same distribution.

In the following, we test the randomness of the process of emitting Reuters articles with the Wald-Wolfowitz runs test and the Bartels test on the chronological sequence of distances. We also test the isolated emission of articles within the subclasses defined by the clustering. Table I shows the analysis results: The p-values of the respective tests indicate that the null hypothesis ("The articles are i.i.d.") is rejected if the data is analyzed as a whole, but that it is accepted if each cluster is analyzed in isolation, i.e., the articles are possibly i.i.d.

C. A Model Selection Example

To apply model selection as described in Section III-C, the effectiveness has to be normally distributed under subclass

Table I
RANDOMNESS OF THE DATA GENERATION FOR THREE RUBRICS OF THE REUTERS CORPUS, QUANTIFIED BY BARTELS AND WALD-WOLFOWITZ TESTS. p-VALUES ARE COMPUTED FOR THE ENTIRE SETS (OVERALL) AS WELL AS FOR THE CLUSTERINGS (AVG. CLUSTER).

	Science		**Sports**		**Politics**	
	avg. cluster p-values	overall p-value	avg. cluster p-values	overall p-value	avg. cluster p-values	overall p-value
Bartels	0.169	≈ 0	0.139	≈ 0	0.141	≈ 0
Wald-Wolfowitz	0.274	≈ 0	0.202	≈ 0	0.105	≈ 0

Table II
SHAPIRO-WILK TEST TO ANALYZE IF THE EXPECTED EFFECTIVENESS IS NORMALLY DISTRIBUTED. THE W- AND p-VALUES ARE AVERAGED OVER THE CLASSIFIERS IN SECTION IV-A AND OVER ALL STANDARD MEASURES, WHICH ARE BASED ON THE CONFUSION MATRIX.

Science versus Sports					
ODP		**RCV1**		**20NG**	
W	p-value	W	p-value	W	p-value
0.93	0.29	0.92	0.17	0.93	0.37

distribution shifts. We revisit the theoretical result that the effectiveness is normally distributed by employing the Shapiro-Wilk test, which has been shown to be one of the most powerful tests of normality [15]. The value W of the test is the ratio between two variance estimators for a random sample $e_1 < e_2 < \cdots < e_n$. The first variance estimator is the expected variance of an assumed normal distribution, while the second variance estimator is the bias-corrected variance of the given random sample, cf. [15]. A W close to one indicates a normal distribution. The high p-value of the Shapiro-Wilk test indicates that the null hypothesis ("The data is normally distributed.") cannot be rejected. For the results reported in Table II, we removed the 5 highest and lowest values in the evaluation since the test is very sensitive to outliers. With respect to all measures and classifiers the estimated effectiveness passed the test under the subclass distribution shift.

In addition, we conduct classification experiments with all mentioned classifiers and corpora and average the results over 10 different testing and training samples. The expected effectiveness is estimated on 1 000 different test samples based on the initial clustering. Considering the most commonly used measures, namely accuracy Acc, precision $Prec$, and recall Rec, and a probabilistic lower bound Θ, which results from $\delta = 0.0228$, the average approximation error of $\widetilde{e}$ by the heuristics $\widehat{e}$ is $3\,\%$.

V. CONCLUSION

We presented the notion of expected effectiveness and its probabilistic lower bound as a basis for preferring one classification solution over another when the underlying data source undergoes a shift in the distribution of its subclasses. Subclass distribution shifts occur in many real-world classification applications, and quite often one has no knowledge about how such a shift will evolve. Our idea is to prefer the solution that has the best probabilistic lower bound of its effectiveness. This bound is based on the expected effectiveness if all shifts are considered equally likely.

Our estimate of the expected effectiveness relies on a repetitive resampling of the clustered test sample for different margin distributions. Clustering is an appropriate method, as exemplified in the experimentation section for news articles, for the identification of those subclasses that are not subject to distribution shifts. The effectiveness within these subclasses is nearly constant. This observation suggests a heuristic for computing another expected effectiveness estimate, namely, to use the mean effectiveness over the clustering. In an empirical evaluation we applied the outlined considerations to standard text corpora, and we showed that the heuristic for the expected effectiveness has a low approximation error.

REFERENCES

[1] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proc. of CIKM*, 2009.

[2] G. Forman, "Tackling concept drift by temporal inductive transfer," in *Proc. of SIGIR*, 2006.

[3] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning for differing training and test distributions," in *Proc. of ICML*, 2007.

[4] P. Prettenhofer and B. Stein, "Cross-lingual adaptation using structural correspondence learning," *Transactions on Intelligent Systems and Technology*, vol. 3, pp. 13:1–13:22, 2011.

[5] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proc. of WDSM*, 2011.

[6] P. Lindstrom, B. Mac Namee, and S. J. Delany, "Drift detection using uncertainty distribution divergence," in *Proc. of ICDM Workshops*, 2011.

[7] J. Vreeken, M. van Leeuwen, and A. Siebes, "Characterising the difference," in *Proc. of SIGKDD*, 2007.

[8] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *Journal of Multivariate Analysis*, vol. 50, pp. 41–54, 1994.

[9] A. Dries and U. Rückert, "Adaptive concept drift detection," *Statistical Analysis and Data Mining*, vol. 2, pp. 311–327, 2009.

[10] Y. Zhang, X. Li, and M. E. Orlowska, "One-class classification of text streams with concept drift," in *Proc. of ICDM Workshops*, 2008.

[11] M. N. Nguyen, X. Li, and S.-K. Ng, "Ensemble based positive unlabeled learning for time series classification," in *Proc. of DASFAA*, 2012.

[12] S. Huang, "An active learning method for mining time-changing data streams," in *Proc. of IITA*, 2008.

[13] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "On demand classification of data streams," in *Proc. of SIGKDD*, 2004.

[14] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proc. of SIGKDD*, 2001.

[15] N. M. Razali and Y. B. Wah, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, pp. 21–33, 2011.