Active Learning for Text Classification: Using the LSI Subspace Signature Model

Weizhong Zhu
Department of Information Sciences
City of Hope Medical Center
Los Angeles, CA, USA
weizz2004@yahoo.com

Robert B. Allen
Department of Library and Information Science
Yonsei University
Seoul, Korea
rballen@yonsei.ac.kr

***Abstract*—Supervised learning methods rely on large sets of labeled training examples. However, large training sets are rare and making them is expensive. In this research, Latent Semantic Indexing Subspace Signature Model (LSISSM) is applied to labeling for active learning of unstructured text. Based on Singular Value Decomposition (SVD), LSISSM represents terms and documents as semantic signatures by the distribution of their local statistical contribution across the top-ranking LSI latent dimensions after dimension reduction. When utilized to an unlabeled text corpus, LSISSM finds the most important samples and terms according to their global statistical contribution ranking in the corresponding LSI subspaces without prior knowledge of labels or dependency to model-loss functions of the classifiers. These sample subsets also effectively maintain the sampling distribution of the whole corpus. Furthermore, tests demonstrate that the sample subsets with the optimized term subsets substantially improve the learning accuracy across three standard classifiers.**

Keywords — active learning, classifiers, Latent Semantic Indexing Subspace Signature Model, text categorization

I. INTRODUCTION

Progress has been made in developing accurate classifiers, such as Naïve Bayes [17], K-Nearest Neighbor (KNN) [6], Rocchio [14, 19], SVM [20], and regression algorithms. However, these techniques do not solve the difficulty of labeling a large number of training examples. In practice, raw text is frequently unlabeled and to explicitly tagged them is very expensive. Alternatively, learners can start with a large pool of unlabeled examples and then strategically decide which label to probe, which example to label, and how many training examples and labels to select without reducing the accuracy of the classifiers. This latter approach has triggered interest in active learning. Potentially, active learning methods could substantially decrease the size of the set of training examples through bootstrapped sampling and iterative learning without degrading the performance of the classifiers.

The goal of our study is to develop an active learning method to improve the sampling selection process for a large unlabelled text corpus. We propose an active learning method based on LSISSM document ranking [26] to select training documents and possible conceptual labels from the approximated document subspace and the conceptual term subspace. This approach ranks the documents and the conceptual features regarding to their statistical contribution in the latent subspaces generated by LSISSM. Then, the top-ranked documents represented by the most important conceptual feature subsets are selected to generate the sample subsets. These subsets reflect the sample distribution of the whole corpus on the text categories. The performance of active learning methods for text categorization is mainly determined by three factors: (1) classification algorithms; (2) optimized features; and (3) optimized training samples. LSISSM introduces a straightforward unsupervised approach to target feature selection and sample selection. For an unlabelled text corpus it could determine the best labeling order of the documents as well as the best conceptual feature subsets to represent these labeled documents based on their LSISSM semantic signatures' global contribution/importance.

II. LITERATURE REVIEW

Active learning differs from "learning from examples" in that the learning algorithm assumes at least some control over the selection of training examples from the input domain [5]. The purpose of active learning is to produce good classifiers with fewer labels and training points than what would be required by traditional machine learning methods. Instead of the random selection of labels and training points, a better strategy is to choose the labels adaptively. For instance, one could start by querying some random data points to get a rough sense of where the decision boundary falls and then gradually refine the estimate of the boundary by querying points in its immediate vicinity. The algorithm could also attempt to find the labels and data points whose statistical properties make them especially informative and important. However, a common problem in these approaches is sampling bias. That is, the set of labeled points may not reflect the underlying data distribution. This makes it hard to show that the classifiers learned have good statistical properties, for instance, that they converge to an optimal classifier in the limit of infinitely many labels.

A range of active learning algorithms and techniques have been developed that seeks to address these challenges. Most of them employ an aggressive, adaptive sampling strategy, and many show promise in experimental studies. Lewis and Catlett [15] proposed a probabilistic sampling method to decrease the error rate of the C4.5 classifier. The Query-By-

Committee (QBC) learning algorithm [11] observes a stream of unlabeled data and makes spot decisions about whether or not to ask for a sample point's label. They show that if the data are drawn from the surface of the d-dimensional unit sphere and the hidden labels correspond to a homogeneous linear separator from the uniform distribution, then it is possible to achieve generalization error e after seeing O(d/e) points and requesting just O(d log 1/e) labels, which is an exponential improvement over the usual O(d/e) sample complexity of linear separators in a supervised setting. This remarkable result is tempered by the complexity of the QBC algorithm which involves computing volumes of intermediate version spaces. McCallum and Nigam [18] integrated a QBC learning approach with an EM classifier which reduces more than one third of the training samples without degrading the performance of the classifier. Gilad-Bachrach et al. [13] used kernels to extend the QBC model to a non-linear scenario by sampling an approximated low-dimension space with hit and run random walk.

Dasgupta [7] proposed a greedy active learning rule to identify the upper and lower bounds of labeling. His study suggests that in using linear separators which are non-homogeneous the sample complexity necessarily shoots up to 1/e. Dasgupta et al. [8] introduced a simple variant of the perceptron model which achieves the same sample complexity bounds as that of the QBC model. The membership query model [1] assumed that every point in the data will be perfectly separated according to certain hypotheses and the learner is allowed to query the label of any point in the input space and prove these perfect hypotheses. However the experiments suggest the model is of limited practical value.

Against the misspecification of the above models, recent studies on weighted least-square active learning which concerns "importance" are demonstrated to be more robust. Sugiyama [21] proposed ALICE, an importance-based least-square active learning method, which predicts the conditional expectation of the generalization error. Beygelzimer et al. [3] introduced an importance weighting scheme for active learning to correct sampling bias by controlling variance. Tong and Koller [22] developed an efficient margin approximation and SVM-based iterative learning approach that significantly decreases the size of the training examples. Yang et al. [25] introduced a SVM-based active learning approach which predicts multiple labels for a single document sample. Li et al. [16] built an active learning framework to label single data points with non-domain and domain specific features.

Thus, there are many types of active learning techniques which achieve excellent performance in different domains. However, they do not identify samples for outlier text categories. Nor do they propose how to order unlabelled samples without prior knowledge of the labels and the classifiers. We believe that LSISSM provides a novel sample-selection process to address these issues.

III. LSI SUBSPACE SIGNATURE MODEL (LSISSM)

LSISSM [26, 27] builds a semantic content representation of unstructured text to simulate the semantic associations among conceptual terms and documents. Each term/document is represented as a normalized semantic signature. The signatures are used to rank, match and group related terms/documents. The mathematical foundation of LSISSM is Singular Value Decomposition (SVD), which is also the basis of Latent Semantic Indexing (LSI) [9]. SVD decomposes the term-document matrix A into a term-concept matrix U, a diagonal concept-concept C matrix, and a document-concept matrix V: $A = UCV^T$.

However, LSISSM is fundamentally different from LSI from the perspectives of term/document signature representation, dimensionality reduction and similarity measures. Ding [10] developed a probabilistic model which suggested that the contribution of LSI dimensions follows the Zipf distribution [12]. Motivated by Ding's conclusion that the singular value squared is proportional to the statistical contribution of the latent concept dimensions, LSISSM defines the normalized contribution of the approximated subspace to the overall semantic space as the ratio of the sum of the singular value squared of the subspace, to the sum of the singular value squared of the entire space. Experiments [27] show that even with a subspace which makes major contribution (e.g., 80%) the number of the latent dimensions can be large after applying LSISSM. This indicates that the few hundreds of top latent concept dimensions which are typically used in traditional LSI may not cover all the important topics which dominate the performance of clustering or text categorization techniques. LSISSM controls the upper boundary of the number of latent concept dimensions selected and transfers the dimension reduction problem to a feature reduction challenge by the step-wise term-picking algorithm [28]. In this research, we apply the same ranking mechanism to pick the most important samples from the LSI document subspace regarding to the accumulated global contribution of the documents. This unique training-dataset-generation process relies on the importance and discrimination of the LSISSM semantic signatures and is fully independent of the learning processes of the classifiers.

A. LSISSM Term Signatures and Document Signatures

LSISSM defines term and document signatures according to their local statistical contribution to the LSI latent concept dimensions. The contribution is calculated by the squared product of the singular value of the latent concept dimension and the least-square distance from the dimension to the original term or document. The signatures are generated in three steps:

Step 1: Use Singular Value Decomposition (SVD) to decompose the term-by-document matrix A into a term-concept matrix U, a diagonal matrix C and a document-concept matrix V as described above.

Step 2: For the U or V matrices, the top K dimensions are selected as the proximity of the overall subspace. In this model, the value of K is determined from:

$$T_d = \frac{\sum_{j=1}^{K} D_j}{\sum_{n=1}^{M} D_n}$$

Where D is the square of the singular value S and M is the total number of the latent concept dimensions with a non-zero contribution. T_d defines the contribution portion of the top K dimensions that contribute to the overall latent subspaces. With T_d set at 0.95, the K dimensions that contribute 95% to the overall subspaces are taken as useful information and the remainder is considered as noise.

Step 3: To a dimension n, the contribution value w_{in} of a term or a document X_i is calculated as $w_{in} = D_n X_{in}^2$ where X_{in} is the n^{th} dimension projection score.

A term or a document X_i is represented by a vector w_{in} as a signature, where n belongs to $\{1,.., K\}$ and K is determined by T_d. The global contribution of a term/document to the selected LSI subspaces is based on $T_g = \frac{\sum_{n=1}^{k} w_{in}}{\sum_{i=1}^{t}\sum_{n=1}^{K} w_{in}}$

where t is the total number of the terms/documents in the corpus.

B. LSISSM Signature Ranking

Zhu and Allen [26, 27] found an effective signature rating algorithm in LSISSM by using Global and Local Contribution Ranking (GLCR). GLCR iteratively picks terms/documents based on their rankings until these terms and documents collectively reach a predefined threshold of T_g. GLCR has two steps: (1) the global contribution of one term or one document is measured by T_g, and (2) the local contribution of that term/document is estimated by the absolute value of its projection score to one latent concept dimension. GLCR first selects a threshold T_l for the local contribution. T_l is determined empirically and is generally initialized by a portion to the mean of the absolute values of all the projection scores across the top K dimensions. GLCR scans the signature of every term or document. The selection starts from the dimension with the highest singular value and steps through the dimensions in descending order. As long as the absolute value of the projection score for a candidate is larger than T_l, it will be selected. Otherwise, the searching process goes to next dimension. Finally, each term or each document in the subset is ranked by the value of T_g.

IV. Text Corpus and Pre-processing Procedures

Three standard news corpora with labeled categories were selected:[1] Reuters 21587, TDT1, and TDT2. These corpora contain many outlier categories (i.e., with a sample size less than five). From the Reuters21578–Apte-90Cat corpus 2527 training news documents were selected that belong to 10 categories (see Table I). The sample size for each category varied from 1 to 1646. The TDT1 corpus includes three outlier categories; we generate a subset which includes all 25 categories and contains 1131 documents which are evaluated as "YES" rather than "BRIEF" in the evaluation sheet of TDT1. The sample size for each category ranged from 2 to 273 articles.

The full TDT2 corpus has 100 categories (news story topics). From that, we selected a subset consisting of 30 categories (topics), including six outlier categories, and containing 3349 documents which are evaluated as "YES" rather than "BRIEF" in the evaluation sheet of TDT2. The sample size for the categories ranged from 1 to 1132. Additionally, we assembled a different sample set in each corpus for independent testing of the classifiers. Overall, the test sets included 1031, 274 and 1790 articles from the Reuters, TDT1 and TDT2 collections respectively.

Pre-processing influences the quality of the sampling. We applied the Stanford part-of-speech (POS) tagging [23, 24], stop-word filtering to each corpus using the Google stop word list and Porter stemming. To be considered for a category, each document has to include at least one noun as identified by the POS tagger. Only nouns were included because our study focuses on the concept representations of the documents. The association between a noun and a document in the initial term-document matrix was weighted by traditional TFIDF and each column of the matrix is normalized to 1.0.

V. Experimental Results

We selected the top ranked samples from the signature ranking (Section III.B) for the initial training set. To determine the generality of our approach, we applied three standard classification algorithms: Naïve Bayes, Rocchio, and KNN.

A. Strategy for Evaluation of LSISSM for Active Learning

We evaluated LSISSM by exploring whether it develops effective training sets for the classifiers. First, we inspected the training sets to determine whether they represent the range of categories found in the full corpus. Then, we examined classification accuracy compared to the full corpus. The classification procedure follows three steps:

Step1: Select the training sets with either the LSI Subspace Signature ranking algorithm (Section III.B) or by training point random sampling.

Step2: The learning curves of the classifiers are estimated through stratified n-fold cross validation [4] to determine the optimal size of the training set for improving the categorization. Thirteen training partition points are chosen to perform the n-fold cross validation: {0.00, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}.

Step3: For each training point, the independent testing sets are evaluated within the n-fold cross validation process.

The accuracy of each training point is measured by the average of the classification accuracy in each fold where the

[1]The Reuters collection was obtained from www.daviddlewis.com/resources/testcollections/reuters21578/
The TDT1 and TDT2 collections were obtained from LDC.

classification accuracy is defined as the number of documents classified as "YES" divided by the size of the testing corpora. The overall classification accuracy is the average accuracy across the 13 training partition points.

B. Term Signature Ranking

GLCR selects the term features which make the major contribution to the corpus and it controls the level of feature reduction. The model parameters, T_d, T_l and T_g, are varied within a broad range to maximize the information included while minimizing the noise. The values of T_d, T_l and T_g are between 0.0 and 1.0. T_d determines how many top dimensions are selected. A larger T_l value means less noise. T_g sets the upper boundary of the overall contribution of the term subspace if T_l and T_d are predefined. If T_d or T_g increases or T_l decreases, the size of term-feature sets increases.

C. Experimental Results for Active Learning

1) *Sample Selection Including Outlier Categories*

LSISSM document signature ranking is applied to generate training samples across the three corpora and the resulting variation of T_d, T_g and T_l is explored. After ranking by GLCR, the sampling distributions of the documents in the subset of the Reuters news articles with ten text categories are listed in Table I. R100 denotes the sample distribution in the full training set. RL25 denotes the sample subset selected by GLCR with T_g set as 0.25, T_d set as 0.95 and T_l set as 0.5. RL5080 denotes the sample subset selected by GLCR with T_g set = 0.50, T_d = 0.8 and T_l = 0.6.

If a subset is selected randomly, that subset probably will not do a good job of representing the different categories. For instance, we randomly selected five document subsets each with 608 documents, which is the same size as RL25. Those subsets always excluded the category "sun-meal" that contains only one sample (see Table I). By comparison, GLCR includes it consistently if T_g is not less than 0.25 with T_d set at 0.95 and T_l set at 0.5.

To replicate the observation about the inclusion of small categories found in the training samples in the Reuters collection, we examined the TDT1 and TDT2 corpora which have many more categories. For TDT1, T1100 in Table II summarizes the full training. T1395 denotes the sample subset selected by GLCR with T_g set as 0.395 (378 out of 1131), T_d set as 0.65 and T_l set as 0.3. T150 denotes the sample subset selected by GLCR with T_g set as 0.50 (488 out of 1131), T_d set as 0.65 and T_l set as 0.3. T190 denotes the sample subset selected by GLCR with T_g set as 0.90 (968 out of 1131), T_d set as 0.65 and T_l set as 0.3. We found that the two categories with the smallest sample size, "Kerrigan/Harding" and "Cuban riot in Panama", are included in the training set with our technique but are consistently missed by random selection.

For TDT2, T2100 in Table III denotes the full training set. T230 denotes the sample subset selected by GLCR with T_g set as 0.307, T_d set as 0.73 and T_l set as 0.6. T238 denotes the sample subset selected by GLCR with T_g set as 0.384, T_d set as 0.95 and T_l set as 0.6. The T_g values of T230 to T238 shift from 0.307 to 0.384. Those values reflect the lower bounds

TABLE I. SAMPLING DISTRIBUTION ACROSS 10 CATEGORIES OF THE REUTERS CORPUS USING GLCR.

Category/Subset	RL25	RL5080	R100	Category/Subset	RL25	RL5080	R100
Acq	518	991	1646	Palmkernel	1	2	2
Coffee	9	53	110	Sugar	10	29	125
Interest	41	58	339	Sun-meal	1	1	1
Iron-Steel	7	14	40	Veg-oil	7	29	87
Oat	1	1	8	Wheat	13	28	205

TABLE II. SAMPLING DISTRIBUTION ACROSS 25 CATEGORIES OF THE TDT1 CORPUS SELECTED BY GLCR.

Category/Subset	T1395	T150	T190	T1100	Category/Subset	T1395	T150	T190	T1100
Aldrich Ames	5	6	8	8	Kerrigan/Harding	1	1	1	2
Carlos the Jackal	4	4	8	10	Kobe Japan quake	31	39	65	84
Carter in Bosnia	8	15	32	34	Lost in Iraq	15	22	43	44
Cessna on White	6	9	13	14	NYC Subway	9	10	17	24
Clinic Murders (Salvi)	18	23	37	41	OK-City bombing	87	108	222	273
Comet into Jupiter	15	21	43	45	Pentium chip flaw	3	3	4	4
Cuban riot in Panama	1	1	2	2	Quayle lung clot	6	7	11	12
Death Kim Jong-il	11	15	46	58	Serbs down F-16	15	23	59	65
DNA in OJ trial	56	72	105	114	Serbs violate	18	25	77	91
Haiti ousts observers	1	3	9	12	Shannon Faulkner	5	5	7	7
Hall's copter (N.Korea)	36	44	80	97	USAir 427 crash	7	10	33	39
Humble, TX, flooding	8	8	18	22	WTC Bombing	8	10	21	22
Justice-to-be Breyer	4	4	7	8					

TABLE III. SAMPLING DISTRIBUTION ACROSS 30 CATEGORIES OF THE TDT2 CORPUS SELECTED BY GLCR.

Category/Subset	T230	T238	T2100
20001	194	322	1132
20005	25	29	41
20010	2	3	7
20012	9	26	151
20013	206	281	540
20014	1	1	2
20017	16	17	20
20019	52	55	110
20022	30	30	30
20023	34	46	125
20025	1	1	1
20026	66	67	70
20027	1	1	1
20030	2	2	2
20036	5	5	5

Category/Subset	T230	T238	T2100
20040	6	6	6
20044	29	47	280
20047	63	71	93
20050	11	11	11
20060	8	8	8
20065	55	55	60
20068	8	8	8
20071	36	40	203
20074	22	28	50
20076	19	44	324
20082	3	3	4
20084	1	1	5
20087	58	61	79
20092	3	3	3
20098	9	9	9

for GLCR to pick a subset which includes every category in the corpora if T_d varies from 0.73 to 0.95 and T_l is set as 0.6. By comparison, when randomly selecting 975 and 1226 documents which have the same size as T230 and T238, 9 and 4 categories are missed out of 30, respectively.

As expected the thresholds of T_d, T_l and T_g affect the scope of the sampling selection. GLCR provides a larger scope of parameters to select sampling sets without losing any categories. For instance, GLCR finds all the categories when T_g is set between 0.3 and 0.384 (T238), T_d set as 0.95 and T_l set as 0.6.

Thus, replicated across the three data sets we show the striking result that outlier categories are included in the training samples generated by GLCR but are frequently excluded by randomly generated sample sub-sets. Presumably this is because the outlier categories have high projection scores on one or more LSI latent concept dimensions and are semantically sufficiently distinctive to be identified by GLCR.

2) *Effect of the Model-based Feature Reduction on the Text Classifiers*

Feature reduction should improve the performance of the classifiers effectively and efficiently. To explore the value of LSISSM for feature reduction, GLCR was applied to rank the terms. T_d is set as 0.95 and T_l is set at 0.5. For the Reuters dataset, GLCR selects a feature subset of 2673 conceptual terms out of the entire set (9070 terms) with a maximum T_g (83.3%). For the TDT1 set, GLCR selects a feature subset of 1614 conceptual terms out of the entire set (8338 terms) with a maximum T_g (82.3%). For the TDT2 set, GLCR selects a subset of 3007 conceptual terms out of the whole set (17083 terms) with a maximum T_g (85.1%). Using these feature subsets, we examined learning curves for the three classifiers on the three full training sets (Table IV) and the independent testing sets (Table V).

In Tables IV and V, "ALL" denotes using all features and the baseline runs. The first columns of these tables denote the size of the feature subsets used in the training documents. The values in columns 3, 4 and 5 denote the average accuracy of the classifier across the 13 training points. The results indicate that the feature subsets consistently improve the performance of the Naïve Bayes classifier. Using the feature subsets, Rocchio and KNN achieve equal or slightly better evaluation scores. Thus, GLCR is effective when selecting small feature subsets to represent the overall concept space.

TABLE IV. AVERAGE LEARNING ACCURACY OF THE THREE CLASSIFIERS, NAÏVE BAYES, KNN AND ROCCHIO, ON THE FULL TRAINING SETS FOR REUTERS, TDT1 AND TDT2 USING FEATURE REDUCTION.

Top Terms	Dataset	Naïve Bayes	KNN	Rocchio
ALL	R100	.873	.832	**.914**
2673	R100	**.874**	**.846**	.913
1847	R100	**.875**	**.848**	.913
1000	R100	**.874**	**.851**	.912
ALL	T1100	.866	**.810**	**.753**
1614	T1100	**.874**	.808	.751
1579	T1100	**.875**	.808	.751
900	T1100	**.877**	.808	.750
ALL	T2100	.870	**.858**	**.896**
3007	T2100	**.877**	.857	.894
2528	T2100	**.877**	.856	.894
1600	T2100	**.879**	.856	.892

TABLE V. AVERAGE LEARNING ACCURACY OF THE THREE CLASSIFIERS, NAÏVE BAYES, KNN AND ROCCHIO, ON THE INDEPENDENT TESTING SETS FOR REUTERS, TDT1 AND TDT2 USING FEATURE REDUCTION.

Top Terms	Dataset	Naïve Bayes	KNN	Rocchio
ALL	R100	.875	.822	.910
2673	R100	**.877**	**.842**	**.911**
1847	R100	**.877**	**.843**	.909
1000	R100	.874	**.851**	.905
ALL	T1100	.768	**.728**	**.710**
1614	T1100	**.794**	.727	.708
1579	T1100	**.793**	.726	.708
900	T1100	**.786**	.723	.700
ALL	T2100	.787	.762	**.820**
3007	T2100	**.789**	**.764**	.816
2528	T2100	.787	**.764**	.816
1600	T2100	.787	**.764**	.815

Feature reduction also improves the efficiency of the classifier. For instance, with the KNN classifier and dataset R100, the total training time decreases from 5.71s to 2.72s and the training time per sample decreases from 0.71ms to 0.66ms.

3) *Effect of Model-based Sample Selection on the Text Classifiers*

GLCR ranks the documents and produces the training sample subsets. We found that those sample subsets produced better learning curves than those for the full training sets across corpora and classifiers. For example, Table VI shows the average accuracy of the three classifiers with the four training sample sets, RL25, Rand608, RL5080 and R100, across the 13 training points. R100 denotes the full training set and we treat its learning accuracy as the baseline. Rand608 denotes the training subset which is randomly picked with the same sample size as that of RL25. RL25 outperforms the learning accuracy of Rand608 by 2.72% using Naive Bayes, by 8.04% using KNN and by 6.94% using Rocchio. In addition, the average learning accuracy of the three classifiers using RL5080 is consistently better than those with the full training set (R100).

For TDT1 and TDT2, the average learning accuracy of the selected subsets, T150, T190, T230 and T238, is a little lower than those of the full training sets, T1100 and T2100 (see Table VI). Because the random selection missed two or more categories in TDT1 and TDT2, we do not compare random sampling with LSISSM.

4) *Combined Effect of Model-based Feature Reduction and Sample Selection on the Text Classifiers*

The results in Table VII indicate that the performance of the sample subsets RL25, RL8050, T150, T230 and T238 on the independent tests are not as quite good as those for the full training sets, R100, T1100 and T2100 (see Tables V), but they are comparable. Moreover, feature reduction shrinks the difference because in most cases the feature subsets improve the performance of the three classifiers on the sample subsets. For T190, Naïve Bayes and Rocchio classifiers outperform T1100. This reflects the fact that a sample subset (968 out of 1131) with a small feature set (1614 out of 8338) trained by a classifier achieves better performance compared to that of the full training set with a full feature set. This sample selection suggests that an optimized sample candidate subset with human labeling could dramatically decrease the time and the cost to make a training corpus with class labels.

VI. DISCUSSION AND CONCLUSION

LSISSM document signature ranking selects crucial training samples and maintains the sampling distribution of the full set of text categories including outlier categories. Most modern active learning approaches do not consider how to retain the distribution of the categories in the sample subsets; rather, they are designed for single categories, which assume that the distribution is known or predefined [2]. Compared to most popular pool-based QBC methods which generally take text categories one at time and report the learning accuracy, LSISSM-based sample-selection approach considers all the categories at once and does not require either prior knowledge of the class labels or any model loss prediction on these classifiers to decide which training point to pick.

TABLE VI. COMPARISON OF AVERAGE LEARNING ACCURACY BETWEEN THE FULL TRAINING SET AND THE TRAINING SUBSETS OF THE REUTER CORPUS, TDT1 AND TDT2 USING NAÏVE BAYES, KNN AND ROCCHIO.

Dataset	Naïve Bayes	KNN	Rocchio
Rand608	.846	.784	.864
RL25	**.869**	**.847**	**.924**
T1100	**.866**	**.810**	**.753**
T150	.832	.758	-------
T190	.862	.806	.752

Dataset	Naïve Bayes	KNN	Rocchio
R100	.873	.832	.914
RL5080	**.893**	**.872**	**.945**
T2100	**.870**	**.858**	**.896**
T230	.848	.810	.872
T238	.858	.828	.887

TABLE VII. COMPARISON OF THE AVERAGE LEARNING ACCURACY OF NAÏVE BAYES, KNN AND ROCCHIO, TRAINED BY THE TRAINING SUBSETS OF THE REUTER CORPUS, TDT1 AND TDT2 WITH FEATURE REDUCTION AND TESTED BY THE INDEPENDENT TESTING SETS.

Top Terms	Dataset	Naïve Bayes	KNN	Rocchio
ALL	RL25	.806	.772	.839
2673	RL25	**.826**	**.789**	**.861**
ALL	RL8050	.836	.802	.879
2673	RL8050	**.846**	**.813**	**.885**
ALL	T150	.719	**.695**	-------
1614	T150	**.735**	.692	-------
ALL	T190	.769	**.727**	**.712**
1614	T190	**.800**	.724	.709
ALL	T230	.726	.702	.775
3007	T230	**.738**	**.704**	**.778**
ALL	T238	.770	.740	**.801**
3007	T238	**.774**	**.743**	.798

Controlled by a fairly large scope of parameters, the signature ranking algorithm of LSISSM identifies the representative samples and the label candidates for all the categories. LSISSM is particularly useful for finding optimized sample subsets and potential conceptual labels for a large unlabeled text corpus. The conceptual feature subsets generated by GLCR significantly enhance the performance of the clustering algorithms [27]. The effects of LSI subspace signature ranking on both unsupervised and supervised classification are consistent and identical. In future work, LSISSM-based clustering will be applied to automatically grouping the top-ranking sample documents and identifying the group representatives as the candidate samples.

REFERENCES

[1] D. Angluin, "Queries revisited," TCS, vol. 313(2), pp. 175-194, 2004.

[2] M. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," J. Comp. Sys. Sci. vol. 75(1), pp.78-89, 2009.

[3] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," ICML, pp. 49-56, 2009.

[4] L. Breiman and P. Spector, "Submodel selection and evaluation in regression: The X-random case," Int. Stat. Rev., vol. 60, pp. 291-293, 1992.

[5] D. Cohn, L. Atlas, and R.. Ladner, "Improved generalization with active learning," Mach. Learn., vol. 15, pp. 201-221, 1994.

[6] T.Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. IT. vol. IT-13, pp. 21–27, 1967.

[7] S. Dasgupta, "Analysis of greedy active learning strategy," NIPS, vol. 18, 2005.

[8] S. Dasgupta, A. Kalai, and C. Monteleoni, "Analysis of perceptron-based active learning," JMLR, vol. 10, pp. 281-299, 2009.

[9] S. Deerwester, S. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," JASIST, vol. 41(6), pp. 391-407, 1990.

[10] C.H. Ding, "A probabilistic model for Latent Semantic Indexing," JASIST, vol. 56(6), pp. 597-608, 2005.

[11] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the Query by Committee Algorithm," Mach. Learn. vol. 28(2-3), pp. 133-168, 1997.

[12] A. Gelbukh and G. Sidorov, "Zipf and Heaps Laws' coefficients depend on language," Intel. Text Proc. Comp. Ling., 2001, pp. 332–335.

[13] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Query by committee made real," NIPS, vol. 18, 2005.

[14] D.J. Ittner, D.D. Lewis, and D.D. Ahn, "Text categorization of low quality images," SDAIR, pp. 301–315, 2005.

[15] D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," ICML, pp. 148-156, 1994.

[16] L.M. Li, X.M. Jin, J.L. Pan, S.J.L., and J.T.Sun, "Multi-domain active learning for text classification," ACM SIGKDD, pp.1086-1094, 2012.

[17] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," AAAI/ICML Workshop Learning for Text Categorization, pp. 41-48, 1998.

[18] A. McCallum and K. Nigam, "Employing EM and Pool-Based active learning for text classification," ICML, pp. 350-358, 1998.

[19] J.J. Rocchio, "Relevance feedback in information retrieval", in G. Salton (Ed.) SMART Retrieval System - Experiments in Automatic Document Processing, New York, Prentice Hall. 1971.

[20] B. Schölkopf, A. Smola, R. Williamson, and P.L. Bartlett, "New support vector algorithms," Neur. Comp., vol. 12, pp. 1207-1245, 2000.

[21] M. Sugiyama, "Active learning in approximately linear regression based on conditional expectation of generalization error," JMLR vol. 7, pp. 141-166, 2006.

[22] S. Tong and D. Koller, "Support Vector Machine active learning with applications to text classification," JMLR, vol. 2, pp. 45-66, 2001.

[23] K. Toutanova, D. Klein, D.C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," HLT-NAACL, pp. 252-259, 2003.

[24] K. Toutanova and D.C. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," SIGDAT Conf. on Emp. Methods in NLP and Very Large Corpora, 2000, pp. 63-70.

[25] B. Yang, J.T.Sun, T.J. Wang, and Z.Chen, "Effective multi-label active learning for text classification," ACM SIGKDD, pp. 917-926, 2009.

[26] W. Zhu, "Text Clustering and Active Learning Using a Latent Semantic Indexing (LSI) Subspace Signature Model and Query Expansion," Doctoral Dissertation, College of Information Science and Technology, Drexel University, 2009.

[27] W. Zhu and R.B. Allen, "Document clustering using the LSI Subspace Signature Model," JASIST, vol. 64(4), pp. 844-860, 2013.

[28] W. Zhu, and C. Chen, "Storylines: Visual exploration and analysis in latent semantic spaces," Int. J. Comp. Graphics (Special Issue on Visual Analytics), vol. 31(3), pp. 338-349, 2007.