A Non-negative Matrix Factorization Based Approach for Active Dual Supervision from Document and Word Labels

Chao Shen and **Tao Li**
School of Computing and Information Sciences
Florida International University
Miami, FL 33199 USA
{cshen001,taoli}@cs.fiu.edu

Abstract

In active dual supervision, not only informative examples but also features are selected for labeling to build a high quality classifier with low cost. However, how to measure the informativeness for both examples and feature on the same scale has not been well solved. In this paper, we propose a non-negative matrix factorization based approach to address this issue. We first extend the matrix factorization framework to explicitly model the corresponding relationships between feature classes and examples classes. Then by making use of the reconstruction error, we propose a unified scheme to determine which feature or example a classifier is most likely to benefit from having labeled. Empirical results demonstrate the effectiveness of our proposed methods.

1 Introduction

Active learning, as an effective paradigm to optimize the learning benefit from domain experts' feedback and to reduce the cost of acquiring labeled examples for supervised learning, has been intensively studied in recent years (McCallum and Nigam, 1998; Tong and Koller, 2002; Settles, 2009). Traditional approaches for active learning query the human experts to obtain the labels for intelligently chosen data samples. However, in text classification where the input data is generally represented as document-word matrices, human supervision can be obtained on both documents and words. For example, in sentiment analysis of product reviews, human labelers can label reviews as positive or negative, they can also label the words that elicit positive sentiment (such as "sensational" and "electrifying") as positive and words that evoke negative sentiment (such as "depressed" and "unfulfilling") as negative. Recent work has demonstrated that labeled words (or feature supervision) can greatly reduce the number of labeled samples for building high-quality classifiers (Druck et al., 2008; Zaidan and Eisner, 2008). In fact, different kinds of supervision generally have different acquisition costs, different degrees of utility and are not mutually redundant (Sindhwani et al., 2009). Ideally, effective active learning schemes should be able to utilize different forms of supervision.

To incorporate the supervision on words and documents at same time into the active learning scheme, recently an active dual supervision (or dual active learning) has been proposed (Melville and Sindhwani, 2009; Sindhwani et al., 2009). Comparing with traditional active learning which aims to select the most "informative" examples (e.g., documents) for domain experts to label, active dual supervision selects both the "informative" examples (e.g., documents) and features (e.g., words) for labeling. For active dual supervision to be effective, there are three important components: a) an underlying learning mechanism that is able to learn from both the labeled examples and features (i.e., incorporating supervision on both examples and features); b) methods for estimating the value of information for example and feature labels; and c) a scheme that should be able to trade-off the costs and benefits of the different forms of supervision since they have different labeling costs and different benefits.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 949–958,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

In Sindhwani et al.'s initial work on active dual supervision (Sindhwani et al., 2009), a transductive bipartite graph regularization approach is used for learning from both labeled examples and features. In addition, uncertainty sampling and experimental design are used for selecting informative examples and features for labeling. To trade-off between different types of supervision, a simple probabilistic interleaving scheme where the active learner probabilistically queries the example oracle and the feature oracle is used. One problem in their method is that *the values of acquiring the feature labels and the example labels are not on the same scale*.

Recently, Li et al. (Li et al., 2009) proposed a dual supervision method based on constrained non-negative tri-factorization of the document-term matrix where the labeled features and examples are naturally incorporated as sets of constraints. Having a framework for incorporating dual-supervision based on matrix factorization, gives rise to the natural question of *how to perform active dual supervision in this setting*. Since rows and columns are treated equally in estimating the errors of matrix factorization, another question is can we address *the scaling issue in comparing the value of feature labels and example labels*.

In this paper, we study the problem of active dual supervision using non-negative matrix tri-factorization. Our work is based on the dual supervision framework using constrained non-negative tri-factorization proposed in (Li et al., 2009). We first extend the framework to explicitly model the corresponding relationships between feature classes and example classes. Then by making use of the reconstruction error criterion in matrix factorization, we propose a unified scheme to evaluate the value of feature and example labels. Instead of comparing the estimated performance increase of new feature labels or example labels, our proposed scheme assumes that a better supervision (a feature label or a example label) should lead to a more accurate reconstruction of the original data matrix. In our proposed scheme, *the value of feature labels and example labels is computed on the same scale*. The experiments show that our proposed unified scheme to query selection (i.e., feature/example selection for labeling) outperforms the interleaving schemes and the scheme based on expected log gain.

The rest of this paper is organized as follows: the related work is discussed in Section 2, and the dual supervision framework based on non-negative matrix tri-factorization is introduced in Section 3. We extend non-negative matrix tri-factorization to active learning settings in Section 4, and propose a unified scheme for query selection in Section 5. Experiments are presented in Section 6, and finally Section 7 concludes the paper.

2 Related Work

We point the reader to a recent report (Settles, 2009) for an in-depth survey on active learning. In this section, we briefly cover related work to position our contributions appropriately.

Active Learning/Active Dual Supervision Most prior work in active learning has focused on pooled-based techniques, where examples from an unlabeled pool are selected for labeling (Cohn et al., 1994). With the study of learning from labeled features, many research efforts on active learning with feature supervision are also reported (Melville et al., 2005; Raghavan et al., 2006). (Godbole et al., 2004) proposed the notion of feature uncertainty and incorporated the acquired feature labels into learning by creating one-term mini-documents. (Druck et al., 2009) performed active learning via feature labeling using several uncertainty reduction heuristics using the learning model developed in (Druck et al., 2008). (Sindhwani et al., 2009) studied the problem of active dual supervision from examples and features using a graph-based dual supervision method with a simple probabilistic method for interleaving feature labels and example labels. In our work, we develop our active dual supervision framework using constrained non-negative tri-factorization and also propose a unified scheme to evaluate the value of feature and example labels. We note the very recent work of (Attenberg et al., 2010), which proposes a unified approach for the dual active learning problem using expected utility where the utility is defined as the log gain of the classification model with a new labeled document or word. Conceptually, our proposed unified scheme is a special case of the expected utility framework where the utility is computed using the matrix reconstruction error. The utility based on the log gain of the classification model may not be reliable as small model changes resulted from a single additional example label or feature label may not be reflected in the classification performance (Attenberg et al., 2010). The empirical comparisons show that our proposed unified scheme based on reconstruction error outperforms the expected log gain.

Dual Supervision Note that a learning method that is capable of performing dual supervision (i.e., learning from both labeled examples and features) is the basis for active dual supervision. Dual supervision is a relatively new area of research and few methods have been developed for dual supervision. In (Sindhwani and Melville, 2008; Sindhwani et al., 2008), a bipartite graph regularization model (GRADS) is used to diffuse label information along both sides of the document-term matrix and to perform dual supervision for semi-supervised sentiment analysis. Conceptually, their model implements a co-clustering assumption closely related to Singular Value Decomposition (see also (Dhillon, 2001; Zha et al., 2001) for more on this perspective). In (Sandler et al., 2008), standard regularization models are constrained using graphs of word co-occurrences. In (Melville et al., 2009), Naive Bayes classifier is extended, where the parameters, the conditional word distributions given the classes, are estimated by combining multiple sources, e.g. document labels and word labels. Our work is based on the dual supervision framework using constrained non-negative tri-factorization.

3 Learning with Dual Supervision via Tri-NMF

Our dual supervision model is based on non-negative matrix tri-factorization (Tri-NMF), where the non-negative input document-word matrix is approximated by 3 factor matrices as $X \approx GSF^T$, in which, X is an $n \times m$ document-term matrix, G is an $n \times k$ non-negative orthogonal matrix representing the probability of generating a document from a document cluster, F is an $m \times k$ non-negative orthogonal matrix representing the probability of generating a word from a word cluster, and S is a $k \times k$ non-negative matrix providing the relationship between document cluster space and word cluster space.

While Tri-NMF is first applied in co-clustering, Li et al. (Li et al., 2009) extended it to incorporate labeled words and documents as dual supervision via two loss terms in the objective function of Tri-NMF as following:

$$\begin{aligned}\min_{F,G,S} \quad & \|X - GSF^T\|^2 \\ & + \alpha \operatorname{trace}[(F - F_0)^T C_1 (F - F_0)] \\ & + \beta \operatorname{trace}[(G - G_0)^T C_2 (G - G_0)].\end{aligned} \tag{1}$$

Here, $\alpha > 0$ is a parameter which determines the extent to which we enforce $F \approx F_0$ to its labeled rows. C_1 is a $m \times m$ diagonal matrix whose entry $(C_1)_{ii} = 1$ if the row of F_0 is labeled, that is, the class of the i-th word is known and $(C_1)_{ii} = 0$ otherwise. $\beta > 0$ is a parameter which determines the extent to which we enforce $G \approx G_0$ to its labeled rows. C_2 is a $n \times n$ diagonal matrix whose entry $(C_2)_{ii} = 1$ if the row of G_0 is labeled, that is, the category of the i-th document is known and $(C_2)_{ii} = 0$ otherwise. The squared loss terms ensure that the solution for G, F in the otherwise unsupervised learning problem be close to the prior knowledge G_0, F_0. So the partial labels on documents and words can be described using G_0 and F_0, respectively.

4 Dual Supervision with Explicit Class Alignment

4.1 Modeling the Relationships between Word Classes and Document Classes

In the solution to Equation 1, we have $S = G^T XF$, or

$$S_{lk} = g_l^T X f_k = \frac{1}{|R_l|^{1/2} |C_k|^{1/2}} \sum_{i \in R_l} \sum_{j \in C_k} X_{ij}, \tag{2}$$

where $|R_l|$ is the size of the l-th document class, and $|C_k|$ is the size of the k-th word class (Ding et al., 2006). Note that S_{lk} represents properly normalized within-class sum of weights ($l = k$) and between-class sum of weights ($l \neq k$). So, S represents the relationship between the classes over documents and the classes over words. Under the assumption that the i-th document class should correspond to the i-th word class, S should be an approximate diagonal matrix, since the documents of i-th class is more likely to contain the words of the i-th class. Note that S is not an exact diagonal matrix, since a document of one class apparently can use words from other classes (especially G and F are required to be approximately orthogonal, which means the classification is rigorous). However, in Equation 1, there are no explicit constraints on the relationship between word classes and document classes. Instead, the relationship is established and enforced implicitly using existing labeled documents and words.

In active learning, the set of starting labeled documents or words is small, and this may generate an ill-formed S, leading to an incorrect alignment of word classes and document classes. To explicitly model the relationships between word classes and document classes, we constrain the shape of S via an extra loss term in the objective function as follows:

$$\begin{aligned}\min_{F,G,S} \quad & \|X - GSF^T\|^2 \\ & +\alpha\,\mathrm{trace}[(F - F_0)^T C_1 (F - F_0)] \\ & +\beta\,\mathrm{trace}[(G - G_0)^T C_2 (G - G_0)] \\ & +\gamma\,\mathrm{trace}[(S - S_0)^T (S - S_0)]\end{aligned} \tag{3}$$

where S_0 is a diagonal matrix.

How to Choose S_0 If there is no orthogonal constraint on F, G and I-divergence is used as the objective function, it can been shown that the factors of Tri-NMF have probabilistic interpretation (Ding et al., 2008; Shen et al., 2011):

$$\begin{aligned} F_{il} &= P(w = w_i | z_w = l), \\ G_{jk} &= P(d = d_j | z_d = k), \\ S_{kl} &= P(z_d = k, z_w = l), \end{aligned} \tag{4}$$

where w is word variable, d is document variable, and z_w, z_d are random variables indicating word class and document class respectively. F and G represent posterior distributions for words and documents, and S represents the joint distribution of document class and word class. With such an interpretation, S_0 can be easily decided in balanced classification problems with each diagonal entry equals to one over the number of classes.

However, in our setting of Tri-NMF, orthogonal constraints are enforced on F, G and Euclidean distance is used as the objective function. To precompute S_0, one way is to first solve the optimization problem Equation 1 with another constraint that S should be diagonal. Alternatively, to keep it simple, we ignore the known label information and just assume there exists a diagonal matrix S_0 and two orthogonal matrices G, F, that

$$GS_0F^T \approx X.$$

Then

$$\begin{aligned} \mathrm{trace}[XX^T] &\approx \mathrm{trace}[GS_0F^TFS_0^TG^T], \\ &= \mathrm{trace}[S_0S_0^TF^TFG^TG], \\ &= \mathrm{trace}[S_0S_0^T], \\ &= \textstyle\sum_k (S_0)_{kk}^2. \end{aligned} \tag{5}$$

So if a classification problem is balanced with K classes, S_0 can be estimated as following:

$$(S_0)_{kl} = \begin{cases} \sqrt{\frac{\mathrm{trace}[XX^T]}{K}} & l = k, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

4.2 Computing Algorithm

This optimization problem can be solved using the following update rules

$$\begin{aligned} G_{jk} &\leftarrow G_{jk} \frac{XFS + \beta C_2 G_0}{(GG^TXFS + \beta GG^TC_2G)_{jk}}, \\ S_{jk} &\leftarrow S_{jk} \frac{F^TX^TG + \gamma S_0}{(F^TFSG^TG + \gamma S)_{jk}}, \\ F_{jk} &\leftarrow F_{jk} \frac{X^TGS^T + \alpha C_1 F_0}{(FF^TX^TGS^T + \alpha C_1 F)_{jk}}. \end{aligned} \tag{7}$$

The algorithm consists of an iterative procedure using the above three rules until convergence.

Theorem 4.1 *The above iterative algorithm converges.*

Theorem 4.2 *At convergence, the solution satisfies the Karuch-Kuhn-Tucker (KKT) optimality condition, i.e., the algorithm converges correctly to a local optima.*

Theorem 4.1 can be proved using the standard auxiliary function approach (Lee and Seung, 2001).

Proof of Theorem 4.2: Proof for the update rules of G, F is the same as in (Li et al., 2009). Here we focus on the update rule of S. We want to minimize

$$\begin{aligned} L(S) = \quad & \|X - GSF^T\|^2 \\ & +\alpha\,\mathrm{trace}[(F - F_0)^T C_1 (F - F_0)] \\ & +\beta\,\mathrm{trace}[(G - G_0)^T C_2 (G - G_0)] \\ & +\gamma\,\mathrm{trace}[(S - S_0)^T (S - S_0)]. \end{aligned} \tag{8}$$

The gradient of L is

$$\frac{\partial L}{\partial S} = 2F^T FSG^T G - 2F^T X^T G + 2\gamma(S - S_0)$$

The KKT complementarity condition for the non-negativity of S_{jk} gives

$$[2F^T FSG^T G - 2F^T X^T G + 2\gamma(S - S_0)]_{jk} S_{jk} = 0.$$

This is the fixed point relation that local minima for S must satisfy, which is equivalent with the update rule of S in Equation 7. ∎

5 A Unified Scheme for Query Selection Using the Reconstruction Error

5.1 Introduction

An ideal active dual supervision scheme should be able to evaluate the value of acquiring labels for documents and words on the same scale. In the initial study of dual active supervision, different scores are used for documents and words (e.g. uncertainty for documents and certainty for words), and thus they are not on the same scale (Sindhwani et al., 2009). Recently, the framework of Expected Utility (Estimated Risk Minimization) is proposed in (Attenberg et al., 2010). At each step of the framework, the next word or document selected for labeling is the one that will result in the highest estimated improvement in classifier performance as defined as:

$$EU(q_j) = \sum_{k=1}^{K} P(q_j = c_k) U(q_j = c_k), \quad (9)$$

where K is the class number, $P(q_j = c_k)$ indicates the probability that q_j, j-th query (a word or document), belongs to the k-th class, and the $U(q_j = c_k)$ indicates the utility that q_j belongs to the k-th class. However, the choice of the utility measure is still a challenge.

5.2 Reconstruction Error

In our matrix factorization framework, rows and columns are treated equally in estimating the errors of matrix factorization, and the reconstruction error is thus a natural measure of utility. Let the current supervision knowledge be G_0, F_0. To select a new unlabeled document/word for labeling, we assume that a good supervision should lead to a good constrained factorization for the document-term matrix, $X \approx GSF^T$. If the new query q_j is a word and its label is k, then the new factorization is

$$\begin{aligned} & G^*_{j=k}, S^*_{j=k}, F^*_{j=k} \\ = \; & \arg\min_{G,S,F} \|X - GSF^T\|^2 \\ & \quad + \alpha \,\mathrm{trace}[(G - G_0)^T C_2 (G - G_0)] \\ & \quad + \beta \,\mathrm{trace}[(F - F_{0,j=k})^T C_1 (F - F_{0,j=k})] \\ & \quad + \gamma \,\mathrm{trace}[(S - S_0)^T (S - S_0)], \end{aligned} \quad (10)$$

where $F_{0,j=k}$ is same as F_0 except that $F_{0,j=k}(j,k) = 1$. In other words, we obtained a new factorization using the labeled words. Similarly, if the new query q_j is a document, then the new factorization is

$$\begin{aligned} & G^*_{j=k}, S^*_{j=k}, F^*_{j=k} \\ = \; & \arg\min_{G,S,F} \|X - GSF^T\|^2 \\ & \quad + \alpha \,\mathrm{trace}[(G - G_{0,j=k})^T C_2 (G - G_{0,j=k})] \\ & \quad + \beta \,\mathrm{trace}[(F - F_0)^T C_1 (F - F_0)] \\ & \quad + \gamma \,\mathrm{trace}[(S - S_0)^T (S - S_0)], \end{aligned} \quad (11)$$

where $G_{0,j=k}$ is same as G_0 except that $G_{0,j=k}(j,k) = 1$. In other words, we obtained a new factorization using the labeled documents. Then the new reconstruction error is

$$RE(q_j = k) = \|X - G^*_{j=k} S^*_{j=k} F^*_{j=k}\|^2. \quad (12)$$

So the expected utility of a document or word label query, q_j, can be computed as

$$EU(q_j) = \sum_{k=1}^{K} P(q_j = k) * (-RE(q_j = k)). \quad (13)$$

To calculate the $P(q_j = k)$, which is the posterior distribution for words or documents, probabilistic interpretation of Tri-NMF is abused. When a query q_j is a word, $P(q_j = k)$ is

$$\begin{aligned} & P(z_w = k | w = w_i) \\ \propto \; & P(w = w_i | z_w = k) \textstyle\sum_{j=1}^{K} P(z_w = k, z_d = j) \\ = \; & F_{ik} * \textstyle\sum_{j=1}^{K} S_{kj}, \end{aligned} \quad (14)$$

otherwise,

$$\begin{aligned} & P(z_d = k | d = d_i) \\ \propto \; & P(d = d_i | z_d = k) \textstyle\sum_{j=1}^{K} P(z_w = j, z_d = k) \\ = \; & G_{ik} * \textstyle\sum_{j=1}^{K} S_{jk}. \end{aligned} \quad (15)$$

5.3 Algorithm Description

Computational Improvement: It can be computationally intensive if the reconstruction error is computed for all unknown documents and words. Inspired by (Attenberg et al., 2010), we first select the top 100 unknown words that the current model is most certain about, and the top 100 unknown documents that the current model is most uncertain about. Then we identify the words or documents in this pool with the highest expected utility (reconstruction error). Equations 14 and 15 are used to perform the initial selection of top 100 unknown words and top 100 unknown documents.

Algorithm 1 Active Dual Supervision Algorithm Based on Matrix Factorization

```
INPUT: $X$, document-word matrix; $F_0$, current labeled words; $G_0$, current labeled documents; $O$, the oracle
OUTPUT: $G$, classification result for all documents in $X$
  1. Get base factorization of $X$: $G, S, F$.
  2. Active dual supervision
  repeat
    $D$ is the set of top 100 unlabeled documents with most uncertainty;
    $W$ is the set of top 100 unlabeled words with most certainty;
    $Q = D \cup W$;
    for all $q \in Q$ do
      for $k = 1$ to $K$ do
        Get $G^*_{q=k}, F^*_{q=k}, S^*_{q=k}$ by Equation 10 or Equation 11 according to whether the query $q$ is a document or a word;
      Calculate $EU(q)$ by Equation 13;
    $q* = \arg\max_q EU(q)$;
    Acquire new label of $q^*$, $l$ from $O$;
    $G, F, S = G^*_{q^*=l}, F^*_{q^*=l}, S^*_{q^*=l}$;
  until stop criterion is met.
```

The overall algorithm procedure is described in Algorithm 1. First we iteratively use the updating rules of Equation 7 to obtain the factorization G, F, S based on initial labeled documents and words. Then to select a new query, for each unlabeled document or word in the pool and for each possible class, we compute the reconstruction error with new supervision (using the current factorization results as initialization values). It is efficient to compute a new factorization due to the sparsity of the matrices. The document-term matrix is typically very sparse with $z \ll nm$ non-zero entries while k is typically also much smaller than document number n, and word number m. By using sparse matrix multiplications and avoiding dense intermediate matrices, updating F, S, G each takes $O(k^2(m+n)+kz)$ time per iteration which scales linearly with the dimensions and density of the data matrix (Li et al., 2009). Empirically, the number of iterations that is needed to compute the new factorization is usually very small (less than 10).

6 Experiments

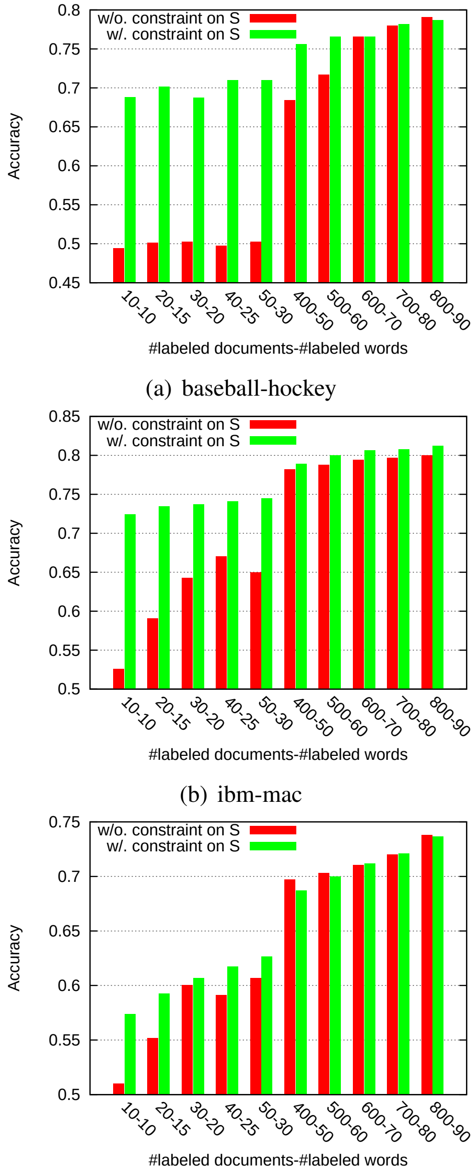
6.1 Experiments Settings

Three popular binary text classification datasets are used in the experiments: ibm-mac (1937 examples), baseball-hockey (1988 examples) and med-space (1972 examples) datasets. All of them are drawn from the 20-newsgroups text collection[1] where the task is to assign messages into the newsgroup in which they appeared. Top 1500 frequent words in each dataset are used as features in the binary vector representation. These datasets have labels for all the documents. For a document query, the oracle returns its label. We construct the word oracle in the same manner as in (Sindhwani et al., 2009): first compute the information gain of words with respect to the known true class labels in the training splits of a dataset, and then the top 100 words as ranked by information gain are assigned the label which is the class in which the word appears more frequently. To those words with labels, the word oracle returns its label; otherwise, the oracle returns a "don't know" response (no word label is obtained for learning, but the word is excluded from the following query selection).

Results are averaged over 10 random training-test splits. For each split, 30% examples are used for testing. All methods are initialized by a random choice of 10 document labels and 10 word labels. For simplicity, we follow the widely used cost model (Raghavan and Allan, 2007; Druck et al.,

[1] http://www.ai.mit.edu/people/jrennie/20_newsgroups/

2008; Sindhwani et al., 2009) where features are roughly 5 times cheaper to label than examples, so we assume the cost is 1 for a word query and is 5 for a document query. We set $\alpha = \beta = 5$, $\gamma = 1$ for all the following experiments[2].



(a) baseball-hockey

(b) ibm-mac

(c) med-space

Figure 1: Comparing the performance of dual supervision via Tri-NMF w/ and w/o the constraint on S.

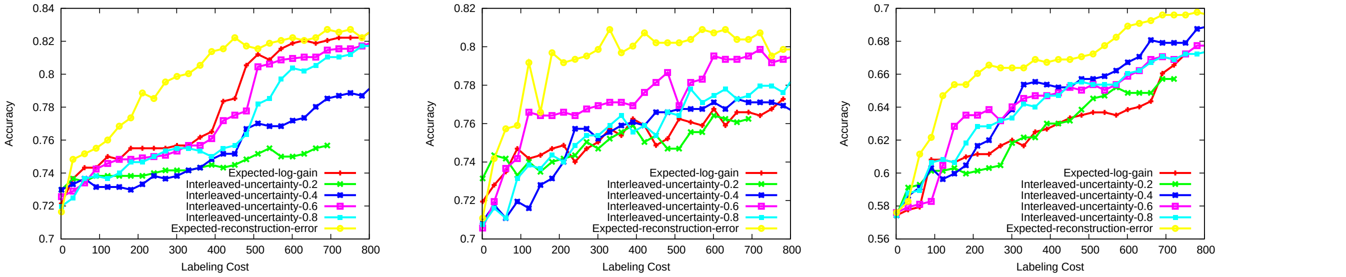
6.2 Experimental Results

Effect of Constraints on S in Constrained Tri-NMF Figure 1 demonstrates the effectiveness of dual supervision with explicit class alignment via Tri-NMF as described in Section 4. When there are enough labeled documents and words, the constraints on S have a relative small impact on the performance of dual supervision. However, in the beginning phase of active learning, the labeled dataset can be small (such as 10 labeled documents and 10 labeled words). In this case, without the constraint of S, the matrix factorization may generate incorrect class alignment, thus lead to almost random classification results (around 50% accuracy), as shown in Figure 1, and further make unreasonable the following evaluation of queries.

Comparing Query Selection Approaches Figure 2 compares our proposed unified scheme (denoted as *Expected-reconstruction-error*) with the following baselines using Tri-NMF as the classifier for dual supervision: (1). *Interleaved-uncertainty* which first selects feature query by certainty and sample query by uncertainty and then combines the two types of queries using an interleaving scheme. The interleaving probability (probability to select the query as a document) is set as 0.2, 0.4, 0.6 and 0.8. (2). *Expected-log-gain* which selects feature and sample query by maximizing the expected log gain. *Expected-reconstruction-error* outperforms interleaving schemes with all the different interleaving probability values with which we experimented. It also has a better performance than *Expected-log-gain*. Although log gain is a finer-grained utility measure of classifier performance than accuracy and has a good performance in the setting with a large set of starting labeled documents (e.g., 100 documents), it is not reliable especially in the setting with a small set of labeled data. Different from the *Expected-log-gain*, *Expected-reconstruction-error* estimates the utility using the matrix reconstruction error, making use of information of all documents and words, including those unlabeled.

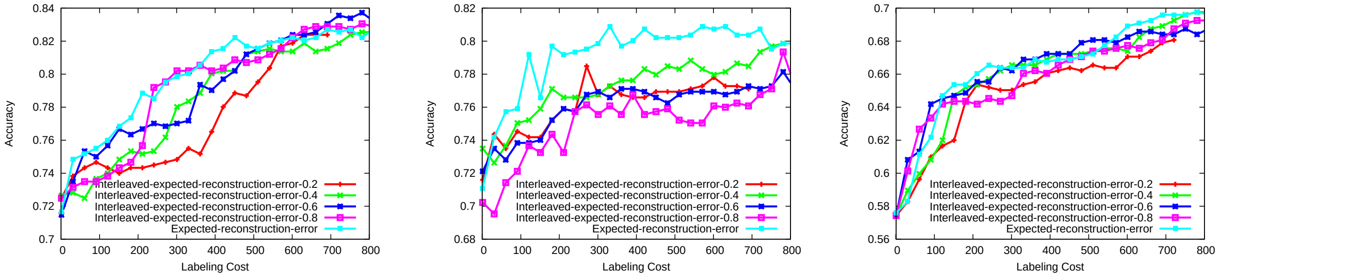
Comparing Interleaving Scheme vs. the Unified Scheme To further demonstrate the benefit of the proposed unified scheme , we compare it with its interleaved version: *Interleaved-expected-*

[2] We do not perform fine tuning on the parameters since the main objective of the paper is to demonstrate the effectiveness of matrix factorization based methods for dual active supervision. A vigorous investigation on the parameter choices is our further work.



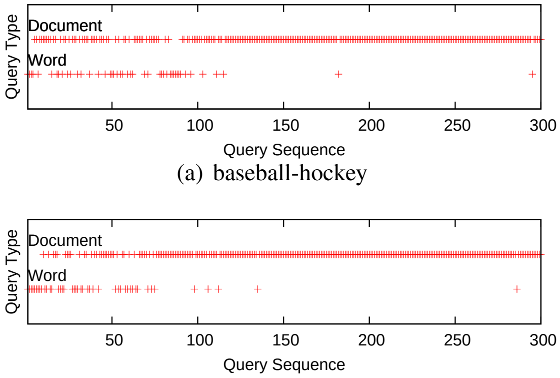
(a) baseball-hockey (b) ibm-mac (c) med-space

Figure 2: Comparing the different query selection approaches in active learning via Tri-NMF with dual supervision.



(a) baseball-hockey (b) ibm-mac (c) med-space

Figure 3: Comparing the unified and interleaving scheme based on reconstruction error.



(a) baseball-hockey

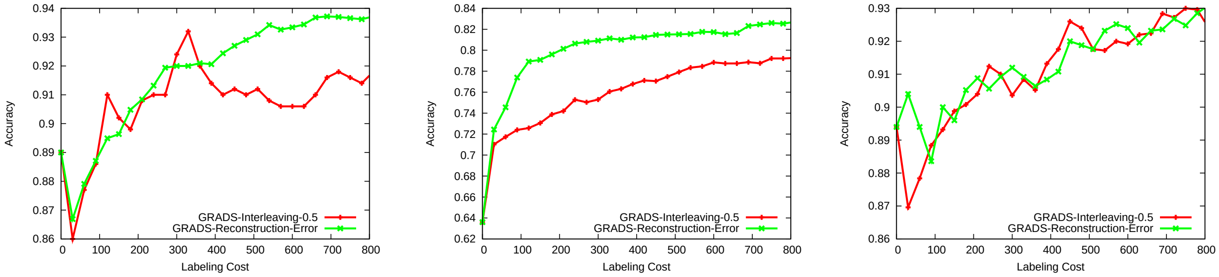
(b) ibm-mac

Figure 5: Example of query sequence.

construction-error which computes the utility of a query using the reconstruction error, but uses interleaving scheme to decide which type of query to select. We experiment with different interleaving probability values ranging from 0.2 to 0.8, which lead to quite different performance results. From Figure 3, the optimal interleaving probability value varies on different datasets. For example, the probability value of 0.8 is among the optimal interleaving probability values on baseball-hockey dataset but performs poorly on ibm-mac dataset. This observation also illustrates the need for a unified scheme, because of the difficulty in choosing the optimal interleaving probability value. Although the proposed unified scheme is not significantly better than its interleaving counterparts for all interleaving probability values on all datasets, it avoids the bad choices.

Figure 5 presents the sequence of different query types selected by our unified scheme and it clearly demonstrates the distribution patterns of different query types. At the beginning phase of active learning, word queries have much higher probabilities to be selected, which is consistent with the result of previous work: feature labels can be more effective than examples in text classification (Druck et al., 2008). And in the later learning phase, documents are more likely to be selected, since the number of words that can benefit the classification is much smaller than the effective documents.

Reconstruction Error vs. Interleaving uncertainty using GRADS It should be pointed out that *our unified scheme for query selection based on reconstruction error does not rely on the estimation of model performance on training data and can be easily integrated with other dual supervision mod-*



(a) baseball-hockey (b) ibm-mac (c) med-space

Figure 4: GRADS with reconstruction error and interleaving uncertainty.

els such as GRADS (Sindhwani et al., 2008). Figure 4 shows the comparison of GRADS using the interleaved scheme with an interleaving probability of 0.5, and using our unified scheme based on reconstruction error. Among the 3 datasets we used, the reconstruction error based approach outperforms the interleaving scheme on baseball-hockey and ibm-mac, and has similar performance with the interleaving scheme on med-space.

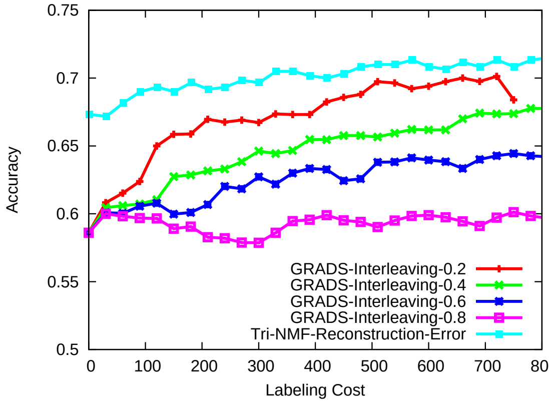


Figure 6: Comparing active dual supervision using matrix factorization with GRADS on sentiment analysis.

Comparing Active Dual Supervision Using Matrix Factorization with GRADS on Sentiment Analysis The sentiment analysis experiment is conducted on the movies review dataset (Pang et al., 2002), containing 1000 positive and 1000 negative movie reviews. The results are shown in Figure 6. The experimental results clearly demonstrate the effectiveness of our approach, denoted as *Tri-NMF-Reconstruction-Error*.

7 Conclusions

In this paper, we study the problem of active dual supervision, and propose a matrix tri-factorization based approach to address the issue, how to evaluate labeling benifit of different types of queries (examples or features) in the same scale. Following extending the nonnegative matrix tri-factorization to the active dual supervision setting, we use the reconstruction error to evaluate the value of feature and example labels. Experimental results show that our proposed approach outperforms existing methods.

Acknowledgement

The work is partially supported by NSF grants DMS-0915110, CCF-0830659, and HRD-0833093. We would like to thank Dr. Vikas Sindhwani for his insightful discussions and for sharing us with his GRADS code.

References

J. Attenberg, P. Melville, and F. Provost. 2010. A Unified Approach to Active Dual Supervision for Labeling Features and Examples. *Machine Learning and Knowledge Discovery in Databases*, pages 40–55.

D. Cohn, L. Atlas, and R. Ladner. 1994. Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

I.S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM.

C. Ding, T. Li, W. Peng, and H. Park. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international* *conference on Knowledge discovery and data mining*, pages 126–135. ACM.

C. Ding, T. Li, and W. Peng. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927.

G. Druck, G. Mann, and A. McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602. ACM.

G. Druck, B. Settles, and A. McCallum. 2009. Active learning by labeling features. In *Proceedings of the 2009 conference on Empirical methods in natural language processing*, pages 81–90. Association for Computational Linguistics.

S. Godbole, A. Harpale, S. Sarawagi, and S. Chakrabarti. 2004. Document classification through interactive supervision of document and term labels. *Knowledge Discovery in Databases: PKDD 2004*, pages 185–196.

D.D. Lee and H.S. Seung. 2001. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

T. Li, Y. Zhang, and V. Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, pages 244–252. Association for Computational Linguistics.

A.K. McCallum and K. Nigam. 1998. Employing EM and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Citeseer.

P. Melville and V. Sindhwani. 2009. Active dual supervision: Reducing the cost of annotating examples and features. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 49–57. Association for Computational Linguistics.

P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. 2005. An expected utility approach to active feature-value acquisition. In *Proceedings of Fifth IEEE International Conference on Data Mining*. IEEE.

P. Melville, W. Gryc, and R.D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the 2002 conference on Empirical methods in natural language processing*, pages 79–86. Association for Computational Linguistics.

H. Raghavan and J. Allan. 2007. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 79–86. ACM.

H. Raghavan, O. Madani, and R. Jones. 2006. Active learning with feedback on features and instances. *The Journal of Machine Learning Research*, 7:1655–1686.

T. Sandler, P.P. Talukdar, L.H. Ungar, and J. Blitzer. 2008. Regularized learning with networks of features. *Advances in Neural Information Processing Systems*, pages 1401–1408.

B. Settles. 2009. Active Learning Literature Survey. *Technical Report 1648*.

C. Shen, T. Li, and C. Ding. 2011. Integrating Clustering and Multi-Document Summarization by Bi-mixture Probabilistic Latent Semantic Analysis (PLSA) with Sentence Bases. In *Proceedings of the national conference on Artificial intelligence*. AAAI Press.

V. Sindhwani and P. Melville. 2008. Document-word co-regularization for semi-supervised sentiment analysis. In *Data Mining, Eighth IEEE International Conference on*, pages 1025–1030. IEEE.

V. Sindhwani, J. Hu, and A. Mojsilovic. 2008. Regularized co-clustering with dual supervision. *Advances in Neural Information Processing Systems*, 21.

V. Sindhwani, P. Melville, and R.D. Lawrence. 2009. Uncertainty sampling and transductive experimental design for active dual supervision. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 953–960. ACM.

S. Tong and D. Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

Omar F. Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*, pages 31–40. Association for Computational Linguistics, October.

H. Zha, X. He, C. Ding, H. Simon, and M. Gu. 2001. Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 25–32. ACM.