Structured Multi-label Transductive Learning: a Case Study in Lexicon Acquisition

Kevin Duh[*]
Dept. of Electrical Engineering
University of Washington, Seattle
duh@ee.washington.edu

Katrin Kirchhoff
Dept. of Electrical Engineering
University of Washington, Seattle
katrin@ee.washington.edu

Abstract

This paper explores structured multi-label transductive clustering for learning lexical part-of-speech information in sparse-data scenarios. We propose three ways of extending existing transductive clustering schemes for binary label assignment to the multi-label case. Preliminary experimental results demonstrate that appropriately defined priors on label assignment hypotheses are crucial for obtaining good performance.

1 Introduction

Most previous work on structured multi-label learning has been carried out in an inductive, supervised learning framework [1, 2], where labeled data is available to learn a classification function, which is then used to classify a new, unseen test set. Recently, extensions to the semi-supervised setting have been proposed [3], where a classification function is learned from partially labeled data. *Transductive learning* [4] is another learning framework that typically makes use of unlabeled data; however, in contrast to inductive learning, the goal here is to only infer labels for the unlabeled data points in the training set rather than to learn a general classification function that can be applied to future data sets. This means that the test data is known a priori and can be used to construct better hypotheses.

Many problems in speech and natural language processing (NLP) can arguably be approached most fruitfully within a transductive learning framework. When developing NLP applications for new languages or domains, labeled data is typically sparse, but often the developer is interested in inferring labels for a particular data set only, rather than for all possible future data sets of the same type. As an example, we consider the task of acquiring a lexicon with part-of-speech (POS) annotations. POS tagging is often the first step for many other NLP applications (such as parsing, information extraction, etc.), and various methods for training taggers in an unsupervised way have been proposed. All of these, however, crucially depend on a reliable lexicon listing the possible POS tags for each word, which is typically constructed by hand or acquired from a previously tagged data set. For many resource-poor languages, these are usually not available; we therefore consider the problem of learning a POS lexicon from a small set of hand-annotated words and a larger set of raw text data representative of the domain we are interested in. Our goal is only to infer the correct assignment of possible POS labels to the unlabeled words in our (finite) word list; we do not expect to be able to classify any future word of the language in question. A transductive learner fulfilling this goal must be able to assign multiple labels to a single data point, and should be able to take advantage of dependencies between labels: for instance, the tags "proper noun" and "common noun" can often be applied to the same word type. In addition, the number of individual tags per word is knot known in advance.

[*] Work funded by NSF Grant IIS-0326276 and an NSF Graduate Fellowship for the 1st author.

```
1   Apply a clustering algorithm to $X_{m+u}$ to generate $C-1$ different partitions with $\tau$ clusters
    each ($2 \leq \tau \leq C$).
2   For each partition
    - generate a label hypothesis $h_\tau$ that labels each cluster with the most frequent tag
      among its labeled data points
    - calculate the bound for $h_\tau$ as defined in Eq. 1.
3   Output the classification of $X_u$ using the hypothesis with the smallest bound.
```

Figure 1: Pseudo-code for transductive clustering.

2 Transductive Clustering

Our learning problem is described formally as having a data sample $X_{m+u} = \{x_1, \ldots, x_{m+u}\}$, from which we draw randomly without replacement a training sample $X_m \subset X_{m+u}$ to receive the labels $y_1, ..., y_m$. The remaining set $X_u = X_{m+u} \backslash X_m$ is the unlabeled (test) sample – thus, training and test set are dependent. In our context, an obvious way to use both X_m and X_u to construct a possible labeling h for X_u is to cluster both sets (labeled and unlabeled words) jointly, and to use the labeled words in each cluster to label the unknown words. Recently, such transductive clustering schemes have been analyzed with respect to their expected performance on the unlabeled data (e.g. [6]). Let $R_h(X) = \frac{1}{M} \sum_{i=1}^{M} l(h(x_i), y_i)$ be the risk of some hypothesis h on data set X (of size M), with $l(h(x), y) \in [0, B]$ being a loss function. It has been shown (see [6] for the proof) that with probability $1 - \delta$ over choices of X_m ($\delta \in (0, 1)$), the risk on the unlabeled data, $R_h(X_u)$, can be bounded by

$$R_h(X_u) \leq \hat{R}_h(X_m) + B\sqrt{\left(\frac{m+u}{u}\right)\left(\frac{u+1}{u}\right)\left(\frac{\ln 1/p(h) + \ln 1/\delta}{2m}\right)} \qquad (1)$$

i.e. it is bounded by the empirical risk on the labeled data, $\hat{R}_h(X_m)$, plus a term that varies with the prior $p(h)$ of the hypothesis – this is a so-called PAC-Bayesian bound [5]. Out of a set H of multiple hypotheses we can now select the one with the lowest error bound. The application of this model selection scheme to transductive clustering was described in [7]; the pseudocode is given in Figure 1. In [7], only problems involving binary label assignments and small numbers of clusters were investigated. A 0/1 loss function (0 if $h(x_i) = y_i$, 1 otherwise) and a uniform prior over all hypotheses was used when calculating the bound. Generally, the slackness of the bound depends on the prior – for problems with a large number of hypotheses (e.g. due to a large number of labels or clusters or both) the uniform prior becomes to small and the bound is not reliable enough. Finding a good way of assigning priors is therefore critical.

3 Extension to Structured Multi-Label Assignment

In the situation where each data point can have $1, ..., k$ possible labels, there are three obvious ways of extending the above algorithm (for illustrative purposes, suppose $k = 3$ and the labels are A,B,C):

1. k **binary learners:** Apply k independent transductive learners, each indicating whether a label is present or absent (ie. A vs. ¬A, B vs. ¬B, etc.). The final labeling is the combined output of the k learners. This has the obvious disadvantage of not exploiting dependencies between different labels; neither does it ensure that each data point receives at least one and fewer than k positive labels.

2. **one n-ary learner:** use all n label combinations observed on X_m as individual labels (e.g. ”A-B”, ”A-B-C”, ”B-C”). Apply one transductive learner as described previously. The potential drawback of this method is that not all possible combinations of labels may

have been observed, particularly when the X_m is small.
3. **Multi-label learner:** Apply one transductive learner. When labeling a cluster, do not only use the majority vote among labels in a given cluster but choose all labels above some frequency threshold. Suppose there are 3 A's, 5 B's, and 2 C's in total and the threshold is the top 80%. We then label all data points ”A-B”; if it is top 50% or less, we use ”B” only.
In all three cases, the bound defined in Eq. 1 applies. In the first two cases, the 0/1 loss is used to compute the risk while in the third case the loss function is bounded by $[0, k]$. If uniform priors were used, $p(h)$ would be $1/(C-1)2^{\tau}$, $1/(C-1)n^{\tau}$, and $1/(C-1)2^{k\tau}$, respectively. As mentioned above, when n or k are large (as they are in practice), the prior probabilities will be too low and the bound will be too loose for effective model (hypothesis) selection. We therefore need to develop better ways of computing priors. Note that priors are also a way of integrating domain knowledge and dependencies between labels – this makes the PAC-Bayesian scheme more attractive than other model selection techniques such as BIC. Among the possible techniques we note:
- for case 1: using conditional priors to incorporate dependencies between learners (e.g. $p(h_B)|p(h_A)$). However, this assumes that labels are assigned in a fixed order (i.e. first the learner for A vs.¬A is applied, then the learner for B vs. ¬B, etc.), and the best ordering may be impossible to find.
- for cases 2 and 3: estimating priors for the joint occurrence of labels from the labeled data (though X_m may be small and priors therefore unreliable), or estimating the priors from some taxonomically similar language with more labeled data.

4 Experiments & Discussion

We present preliminary experimental results on the Wall Street Journal Penn Treebank task. This is not a resource-poor task per se but it is widely available, and we artificially limit the amount of labeled data for simulation purposes. The word list consists of 44k words and the number of possible tags (k) is 46. We ran experiments with 5k, 10k, 20k, 30k, and 40k randomly selected words as labeled data, leaving the remaining as unlabeled (test) data. The full sample is partitioned into sets of 10, 20, 30, ..., 2000 clusters using Brown clustering [8]. Our main goals are to determine (a) whether the PAC-Bayesian model selection criterion is more effective than choosing a hypothesis based on $\hat{R}_h(X_m)$ alone, and (b) which of the multi-label extensions described above work best. We experimentally investigate joint priors estimated from the labeled data vs. uniform priors. Results are reported as the accuracy on concatenated tags and F1-score on the individual tags (e.g. If reference=”A-B” & decision=”A”, then accuracy=0; precision=1, recall=0.5, F1-score=0.67).

Table 1 compares the performance of (1) n-ary with uniform prior [n-uni], (2) n-ary with estimated joint label priors [n-joint], and (3) multi-label with joint priors [multi]. The column [oracle] shows the best possible accuracy that can be achieved among all proposed hypotheses; the column [ER] reports the accuracy for the case where the hypothesis selection is based on the minimum empirical risk on the labeled data. (The former is an upper

Table 1: **Accuracy/F1-score comparison:** [m]=# labeled data. [oracle]= oracle performance when selecting the hypothesis with the lowest risk on the test data . [ER]= selection based on the empirical risk on the training data. [n-uni]= n-ary learner with uniform prior. [n-joint]=n-ary learner with estimated joint prior. [multi]=multi-label with estimated joint prior. Best results are in bold.

	Accuracy					F1-score				
m	oracle	ER	n-uni	n-joint	multi	oracle	ER	n-uni	n-joint	multi
5k	49.54	39.59	44.16	44.87	**47.36**	56.04	48.29	50.17	51.27	**54.68**
10k	50.70	43.91	44.47	**48.05**	47.22	57.69	51.91	50.72	54.56	**54.76**
20k	51.75	48.06	44.37	**48.37**	47.25	58.74	54.28	50.88	**55.03**	54.80
30k	52.35	**49.61**	44.30	48.37	47.26	59.01	**56.61**	50.67	54.88	55.01
40k	53.57	**51.81**	45.16	45.16	48.24	60.38	**58.21**	51.43	51.43	55.75

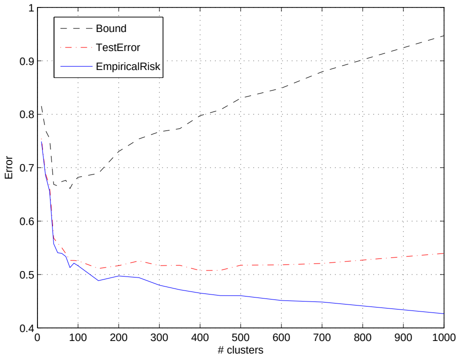

Figure 2: Plot of PAC-Bayesian bound, empirical risk on X_m, and actual test error on X_u for different number of clusters (different hypotheses). (Multi-label, m=5k)

limit to the performance of the transductive clustering scheme while the latter indicates the usefulness of the prior-based penalty term in the PAC-Bayesian bound).

Not surprisingly, the results highlight the importance of using good priors: Both methods that utilize priors over combinations of labels estimated from X_m ([n-joint],[multi]) outperform [n-uni]. We also note that [multi] tends to have a better F1-score but lower accuracy when compared to [n-joint], which is expected due to their different label-assignment schemes. Since no method is the clear winner, it is important to choose the proper evaluation criteria as one that best matches the goals of the downstream application/user of the lexicon. Finally, it is interesting to note that all three PAC-Bayesian learners outperform [ER] when the number of labeled data is small ($m << u$), but the usefulness of the bound decreases as the m increases, indicating that the empirical risk is actually a good indicator of test risk in those cases.

We also experimented with the k-binary case. Due to the unbalanced data problem created by the binary decomposition (i.e. the number of ”¬A” labels far exceeds ”A” labels), 40% of words received no positive labels from any of the k binary learners. This demonstrates the problem that arises when decomposing a multi-label problem without regard to its structure. The 60% of test words that were classified achieved an accuracy of 60.0% to 61.7% and F1-score of 69.7% to 72.5%.

Future work will concentrate on additional ways of defining hypothesis priors, comparisons with other model selection criteria, alternative clustering algorithms, and other applications beyond learning POS sets. It will also be interesting to investigate to what extent the transductive framework is still applicable when the assumption of unbiased training set selection does not hold: in many practical situations, the labeled data is not drawn uniformly at random but in a biased manner. For instance, only the most frequent words in the corpus may initially be labeled, or words that are cognates in some related language or dialect.

References

[1] Collins, M., & Koo, T. (2005) Discriminative reranking for natural language parsing. *Computational Linguistics 31(1):25-69*

[2] Crammer, K., & Singer, Y. (2001) On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*.

[3] Altun, Y., McAllester, D., & Belkin, M. (2005) Maximum margin semi-supervised learning for structured variables. In *Neural Information Processing Systems (NIPS)*

[4] Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley Interscience, New York.

[5] McAllester, D. (1999). Some PAC-Bayesian theorems. *Machine Learning, 37(3):255-363.*

[6] Derbeko, P., El-Yaniv, R., & Meir, R. (2004) Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of AI Research, 22:117-142.*

[7] El-Yaniv, R., & Gerzon, L. (2005) Effective Transductive Learning via Objective Model Selection. *Pattern Recognition Letters 26(13):2104-2115.*

[8] Brown, P.F., deSouza, P., Mercer, R., Watson, T.J, Della Pietra, V.J., Lai, J. (1992) Class-based n-gram models of natural language. *Computational Linguistics 18(4):467-479.*