Active Learning for Recommender Systems with Multiple Localized Models

Meghana Deodhar, Joydeep Ghosh and Maytal Saar-Tsechansky
University of Texas at Austin, Austin, TX, USA.

For effective predictive modeling in large scale recommender systems, it is essential to have many customers rate a large number of products, i.e., obtain a large number of labeled data. However, most consumers often do not provide their preferences without proper incentives. Given a budget to reward consumers for their feedback, it would be beneficial to have a policy to suggest the ratings of which customers and for what products would be most cost-effective to acquire, so as to improve modeling the most. E-commerce businesses can use such a policy to cost-effectively acquire consumers' ratings or other forms of feedbacks, incrementally. This challenge can be mapped to the problem of active learning [2] in which a learner aims to intelligently select the labels of particularly informative examples from a pool of prospective acquisitions, so as to improve generalization accuracy the most for a given number of acquisitions.

While there are published results for active learning in a regression setting [4, 5], there has been little study of approaches that are applicable to important, practical scenarios. In particular, most proposed approaches aim at optimizing the training input density (non pool based) rather than evaluate prospective acquisitions from an available pool, and focus on linear least squares models. In addition, most of the literature on active learning, especially in the machine learning community involves learning in the context of classification problems [2]. Also, these methods consider predictions obtained by either a single "global" probabilistic classification model or an ensemble of global classifiers (e.g., bagging or boosting). As we discuss below, large scale data sets that are characteristic of recommender systems, suggest other types of modeling that are significantly more appropriate.

The data encountered in large scale recommender systems typically exhibit inherent heterogeneity among the customers and products. For instance, Amazon.com being a large retailer has customers who exhibit varied purchase patterns. As we demonstrate, in this setting it is beneficial to model the behavior of different groups of customers separately. Similarly, because the retailer's products span a very wide range of categories, it is also advantageous to model ratings for homogeneous product groups separately. Thus, rather than induce a single predictive model, one can represent such data by a set of multiple local models, such that each local model captures consumers' ratings in a certain region of the input (consumer/product) space. We propose a radically different active learning scheme that (a) leverages a collection of localized predictive models, and (b) generalizes to both classification and regression prediction problems. Our approach is also applicable to a wide range of model types, including non linear models such as MLPs, or regularized linear models (ridge/lasso).

Before developing our active learning strategy, we first discuss briefly how to learn a set of local predictive models to accurately represent such heterogeneous data. We recently proposed Simultaneous CO-clustering And Learning (SCOAL) [3], a versatile and effective framework for predictive modeling of large scale, heterogeneous, dyadic data. SCOAL interleaves simultaneous partitioning along both "customer" and "product" modes (co-clustering) and the construction of prediction models to iteratively improve both the assignment of a consumer's rating to a given cluster, as well as improve the fit of the model induced within each data cluster. SCOAL exploits both neighborhood information and the available customer/product attributes, thereby combining the benefits of collaborative filtering and of content based approaches. The framework can be viewed as simultaneous co-segmentation and classification (or regression), and we show is substantially better than independently clustering the data *a priori* followed by model induction. In SCOAL, each model

represents only a region of the input space, identified by the corresponding co-cluster, and is hence a local model. SCOAL improves interpretability as well as accuracy as compared to other modeling techniques such as a global regression/classification model, Bregman co-clustering [1], sequential clustering and modeling, and a non-linear MLP model.

We propose and empirically evaluate several active learning strategies, which exploit the local models learnt via the SCOAL meta-algorithm to intelligently select the data points in a large heterogeneous data setting. The key statistical challenge of active learning policies and the ones we explore here, is how to evaluate the benefit from acquiring an example's label, such as a customer's rating, before the rating is known. Active learning for classification models, offered a set of utility measures that capture the benefit from such acquisitions. In this paper, we propose several approaches to gauge the benefit from acquiring an example's *numerical* dependent variable. The first strategy we consider estimates a utility from acquiring *any* label from a given co-cluster. The rationale is to acquire examples in regions of the input space that are not well represented by the current set of models. Besides identifying informative acquisitions that improve the predictive accuracy, an important practical advantage of this approach is that at each phase the acquisitions are selected from a restricted and closely related subsets of customers and products. Thus, it increases the likelihood that any given customer is requested to provide feedback on closely related sets of products. We later show that estimating the prediction variance with respect to each individual prospective acquisition in a giver cluster may help differentiate between the benefits from different acquisitions within a cluster and can improve the selection of informative acquisitions. Lastly, we propose an acquisition strategy that caters to the relationship between model complexity and the increasing availability of data in the active learning setting. Specifically, capturing more complex structure in the data using multiple local models is likely to yield higher predictive accuracy as more data is acquired, while a single (hence less complex) model may yield a lower generalization error than multiple models when the data set is sparse (i.e., when many customers' ratings have not yet been acquired). We develop a hierarchical SCOAL active learning approach that begins with a single global model and adaptively increases the number of local models and thus the model complexity as more training data becomes available. Since the choice of model complexity and the selection of training points are not independent, improvements in data modeling also improves data acquisition and vice versa.

Our proposed approaches show promising results on several real data sets from a number of recommender system applications. We address the problem of predicting user-movie ratings on the MovieLens dataset, which consists of 100,000 ratings (1-5) from 943 users on 1682 movies. We also explore active learning on two different data sets created from the ERIM household panel data, (i) a data set of the purchases of products from a number of product categories by a set of customers and, (ii) a time ordered data set that captures changing customer purchase patterns. Both of the latter data sets capture purchasing behaviors of a large, heterogeneous set of customers for a variety of different product types. The SCOAL active learning techniques do significantly better than existing alternative approaches on these data sets.

References

[1] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *JMLR*, 8:1919–1986, 2007.

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

[3] M. Deodhar and J. Ghosh. A framework for simultaneous co-clustering and learning from complex data. In *KDD '07*, pages 250–259, 2007.

[4] V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

[5] D. Wiens. Robust weights and designs for biased regression models: Least squares and generalized m-estimation. *Journal of Statistical Planning and Inference*, 83(2):395 – 412, 2000.