

## Research Summary

Paramveer S. Dhillon (pasingh@seas.upenn.edu)  
Computer and Information Science, University of Pennsylvania

My graduate research has, to date, concentrated on developing Machine Learning methods for domains with structured data, namely Natural Language Processing and Genomics. The main commonality that these domains have is that there is inherent structure in their data. E.g., in language tasks, features naturally fall into classes based on syntax and semantics, and in genomics the features are expression levels of genes that can be divided into gene families. Most state-of-the-art learning models do not take advantage of such structure and they assume, simplistically, that all features belong to a single equivalence class. Furthermore, since problems in these domains have large numbers of potential features, building computationally efficient learning methods is important. Statistical efficiency is even more important, as there is often paucity of labeled data. For example, there are only  $\sim 10$  labeled examples per word in SENSEVAL-2 (Word Sense Disambiguation Data) (Florian and Yarowsky, 2002).

In my Master's thesis we proposed three related models to tackle the above problems. Firstly, we introduced an  $\ell_0 - \ell_0$  sparsity penalty (Dhillon et al., 2008) which allows us to introduce sparsity at two levels: at the level of feature classes and also at the level of individual features within that feature class. Secondly, we extended this penalty to the case of feature selection for multiple related tasks (Dhillon et al., 2009b). The basic idea behind the above methods is similar to the  $\ell_1/\ell_\infty$  penalty for joint regularization (Quattoni et al., 2008; Turlach et al., 2005) and the  $\ell_1/\ell_2$  penalty for enforcing "group sparsity" (Yuan and Lin, 2006; Obozinski et al., 2009), but our methods ( $\ell_0 - \ell_0$  penalty) give much sparser models, which is highly desirable in genomics and language. Since solving the  $\ell_0$  penalty is NP-Hard our methods select features greedily using a stepwise search in the feature space. Thirdly, we proposed a way of doing transfer learning by transferring a feature relevance prior from similar tasks to improve the accuracy of supervised learning algorithms on tasks which have less labeled data available (Dhillon and Ungar, 2009).

As part of a separate project I have also worked on incorporating a general set of constraints in Named Entity Recognition (NER) Systems. Generally, the state-of-the-art NER systems use standard sequence models like Conditional Random Fields, which can only handle local features. But there are some non-local constraints that are very useful and can give us improved F-measures like consistency constraints (e.g. every occurrence of word "Einstein" in a document should be tagged as person (PER)), list constraints (e.g. if there is a list of seminar speakers they should have the same entity type i.e PER.), conjunction and disjunction constraints (e.g. If we have *John and/or James* then both of them should have same entity type). There has been some previous work in this direction; (Sutton and McCallum, 2004) use Skip-chain CRFs for this task and (Finkel et al., 2005) used Gibbs sampling. The main problem with these approaches is that they are custom-tailored for a particular type of constraint (i.e. consistency constraints) and they do not generalize well to other set of constraints. We proposed an approach in which we use these consistency constraints to constrain the model posteriors in generalized EM algorithm (Graca et al., 2008), thereby allowing us to learn these valuable non-local constraints. We have got some encouraging results by using our approach on CoNLL '02 and '03 NER shared tasks and this project is still underway.

Besides this I have also done research in Computer Vision namely in Human Activity Classification in videos. Generally, the activity classification algorithms either use only appearance information or only motion information for classifying human actions. We proposed a model which uses both appearance and motion cues to build robust and highly discriminative histograms to classify human actions in videos and gives state-of-the-art accuracy on KTH Human Action Dataset (Dhillon et al., 2009a).

## References

- P. S. Dhillon and L. H. Ungar. Transfer learning, feature selection and word sense disambiguation. In *Annual Meeting of the Association of Computational Linguistics, (ACL)*, August 2009.
- P. S. Dhillon, D. Foster, and L. Ungar. Efficient feature selection in the presence of multiple feature classes. In *International Conference on Data Mining (ICDM)*, pages 779–784, 2008.
- P. S. Dhillon, Sebastian Nowozin, and Christoph Lampert. Combining appearance and motion for human action classification in videos. In *ViSU'09 (International Workshop on Visual Scene Understanding), CVPR 2009*, June 2009a.
- P. S. Dhillon, B. Tomasik, D. Foster, and L. Ungar. Multi-task feature selection using the multiple inclusion criterion (mic). In *European Conference on Machine Learning (ECML)-PKDD*, Lecture Notes in Computer Science. Springer, September 2009b.
- J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.
- R. Florian and D. Yarowsky. Modeling consensus: classifier combination for word sense disambiguation. In *EMNLP '02*, pages 25–32, 2002.
- J. Graca, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press, Cambridge, MA, 2008.
- G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 2009.
- A. Quattoni, M. Collins, and T.J. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR 2008*, pages 1–8, 2008.
- C. Sutton and A. McCallum. Collective segmentation and labeling of distant entities in information extraction. In *In ICML Workshop on Statistical Relational Learning and Its connections to Other Fields.*, 2004.
- B.A. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 47(3): 349–363, 2005.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Stat. Society: Series B (Stat. Methodology)*, 68(1):49–67, February 2006. ISSN 1369-7412.