Chapter 6

Future Work

6.1 Generalization

- As discussed above, the accuracy of the leave-one-out cross-validation test was low because of the limited number of data points. The confusion matrix in Figure 6.1 shows that about one third of the predictions were incorrect.
- These results can typically be improved by increasing the amount of training data.
- Performed experiment to gauge how the amount of training data affects the accuracy of the LOOCV test. This is shown below in Figure 6.2.

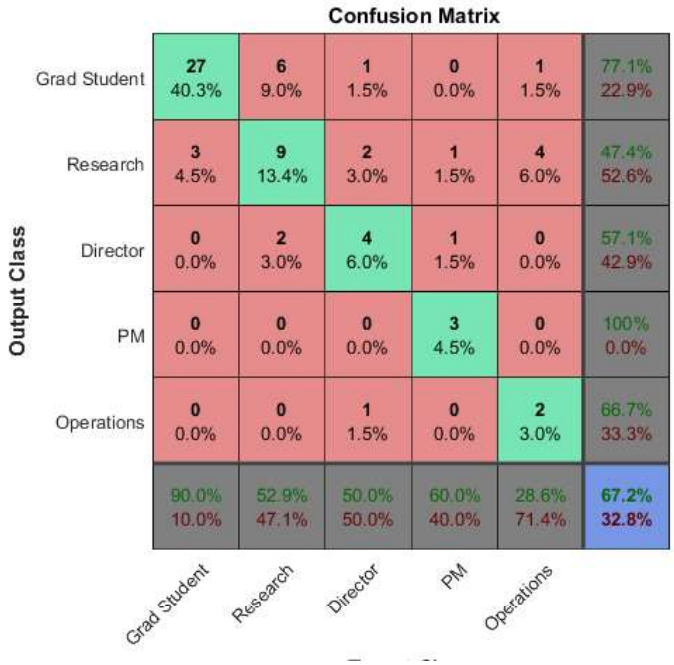

Figure 6.1: Prediction accuracy was higher for graduate students, who have more uniform behaviors and more training data. However, for each category, the majority of the people classified were true members of that group, save for research. Note that even for those classified as research, there were more true researchers than any other class.

6.2 Deep Learning

- Techniques such as random forests work well for small, low-dimensional data. However, imagine if the size of the dataset were greatly increased to billions of people without labels. In this case, a more sophisticated model would be necessary for successful classification. This is required to accurately learn the complicated structure inherent to such large real-world datasets. Suppose a representative sample of volunteers from the dataset self-identifies themselves. This would produce a set of labeled training data, transforming the problem space to semi-supervised learning.

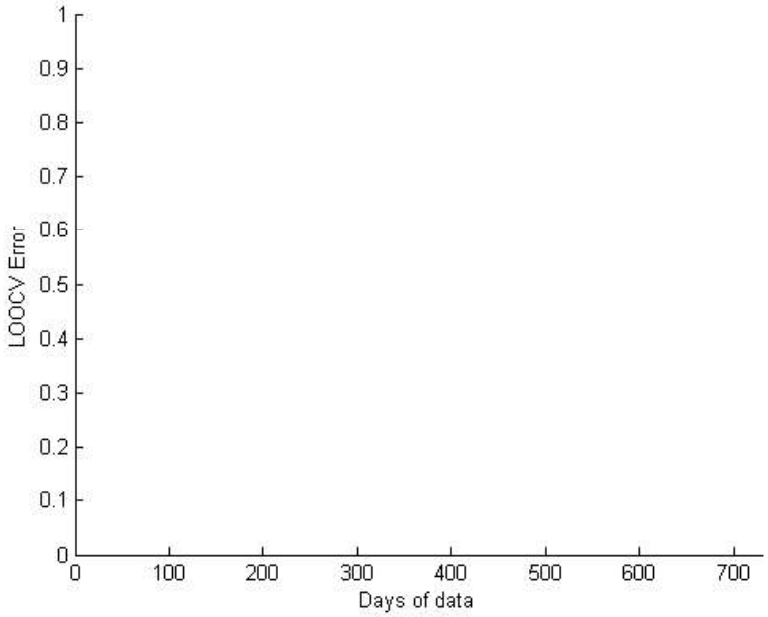

Figure 6.2: Effects of using more data for leave-one-out cross-validation. [Still working on generating this, but I expect it to exponentially decrease. From this data, I can fit a curve to the data that I do have to estimate how well the algorithm will perform with more data.]

- Describe deep belief networks: A deep belief network is a generative graph with nodes that represent stochastic variables. There are several layers of these nodes, some of which are visible and some are hidden. Typically, the top two layers of nodes have undirected connections. Therefore, these layers are actually Restricted Boltzman machines (RBMs). The rest of the layers use directed edges, forming a Bayesian network. For each node, there exists a probability of activation, $p(s_i = 1)$, which is represented by the nonlinear sigmoid function applied to a weighted input from the layer above. Specifically,

$$p(s_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_j s_j w_{ij}\right)} \tag{6.1}$$

where j represents the ancestors of node i, w_{ij} are the weights on the connections between i

and j, and b_i is the bias associated with node i...

- The challenge with very deep networks is that they are very difficult to train [32]. Unsupervised pre-training can improve the efficency of training deep belief networks, as first described in [33]. In that work, Hinton describes a fast, greedy method to train these networks one layer at a time.

- Describe the process of unsupervised pre-training.

- Testing on the MNIST dataset of handwritten digits, unsupervised pre-training is shown to improve classification results over only supervised training [34]. Pre-training identifies a set of initial weights for the network that can ultimately lead to better classification results. An advantage of unsupervised pre-training is that, with small enough layers, it acts as a regularizer by decreasing variance and increasing the bias [33].

- This labeled data can be used for backpropagation to fine-tune the network. Describe the process backpropagation with the small set of labeled data.

- In conjunction with greedy pre-training, backpropagation has been shown to improve both the optimization and generalization of a DBN. Hinton et al. showed that unsupervised pre-training followed by backpropagation of a DBN outperformed traditional feed forward neural networks on the MNIST dataset [35].

- Analogous paper for comparison:

  Much of the research in deep learning today is focused on image classification and analysis,

such as the MNIST dataset. However, DBNs have been applied to text analysis as well. In [36], a sophisticated version of a DBN was used to construct a sentiment classifier. The purpose of this research is to label online reviews as positive, negative, or neutral. The difference between this method and other DBN approaches is that this work replaced some of the hidden layers of the network with a feature selection step. Overall, this improves the training learning efficiency of the algorithm. This method can classify sentiments more accurately and train more quickly than previous semi-supervised algorithms.

Bibliography

[1] S. Radicati and Levenstein, *Email statistics report, 2015-2019.* Technical report, 2015.

[2] D. Fisher, A. J. Brush, E. Gleave, and M. A. Smith, "Revisiting Whittaker & Sidner's email overload ten years later," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pp. 309–312, ACM, 2006.

[3] B. Klimt and Y. Yang, "Introducing the Enron Corpus.," in *CEAS*, 2004.

[4] E. M. Bahgat, S. Rady, and W. Gad, "An E-mail Filtering Approach Using Classification Techniques," in *AISI2015, November 28-30, 2015, Beni Suef, Egypt*, vol. 407, pp. 321–331, Cham: Springer International Publishing, 2016.

[5] R. Shams and R. Mercer, "Classifying Spam Emails Using Text and Readability Features," in *ICDM 2013*, pp. 657–666, Dec. 2013.

[6] B. He, Z. Li, and N. Yang, "A Novel Approach for Email Clustering Based on Semantics," in *WISA, 2014*, pp. 269–272, Sept. 2014.

[7] P. S. Keila and D. B. Skillicorn, "Structure in the Enron email dataset," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 183–199, 2005.

[8] Z. Sofershtein and S. Cohen, "Predicting Email Recipients," pp. 761–764, ACM Press, 2015.

[9] Q. Hu, S. Bao, J. Xu, W. Zhou, M. Li, and H. Huang, "Towards building effective email recipient recommendation service," in *SOLI 2012*, pp. 398–403, July 2012.

[10] A. Nordb, "Data Visualization for Discovery of Digital Evidence in Email," 2014.

[11] K. K. Waterman and P. J. Bruening, "Big Data analytics: risks and responsibilities," *International Data Privacy Law*, vol. 4, pp. 89–95, May 2014.

[12] G. M. Namata, L. Getoor, and C. Diehl, "Inferring formal titles in organizational email archives," in *Proc. of the ICML Workshop on Statistical Network Analysis*, 2006.

[13] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report," *Information sciences institute technical report, University of Southern California*, vol. 4, 2004.

[14] G. Tang, J. Pei, and W.-S. Luk, "Email mining: tasks, common techniques, and tools," *Knowledge and Information Systems*, vol. 41, pp. 1–31, June 2013.

[15] K. Yelupula and S. Ramaswamy, "Social network analysis for email classification," in *Proceedings of the 46th Annual Southeast Regional Conference on XX*, pp. 469–474, ACM, 2008.

[16] S. Martin, B. Nelson, A. Sewani, K. Chen, and A. D. Joseph, "Analyzing Behavioral Features for Email Classification.," in *CEAS*, 2005.

[17] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, Nov. 1994.

[18] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[19] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations," *arXiv:cond-mat/0303264*, Mar. 2003. arXiv: cond-mat/0303264.

[20] G. Wilson and W. Banzhaf, "Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis," in *CEC'09*, pp. 3256–3263, IEEE, 2009.

[21] R. Rowe, G. Creamer, S. Hershkop, and S. J. Stolfo, "Automated social hierarchy detection through email network analysis," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 109–117, ACM, 2007.

[22] J. Shetty and J. Adibi, "Ex employee status report.," 2004. [http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls]. _Internet Archive_.[https://web.archive.org/web/20131126121206/http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls], Accessed 1/30/2016.

[23] J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Computing Surveys (CSUR)*, vol. 31, no. 4es, p. 5, 1999.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the Web.," 1999.

[25] P. G. Lind, M. C. Gonzlez, and H. J. Herrmann, "Cycles and clustering in bipartite networks," *Physical Review E*, vol. 72, Nov. 2005.

[26] J. Saramki, M. Kivel, J.-P. Onnela, K. Kaski, and J. Kertsz, "Generalizations of the clustering coefficient to weighted complex networks," *Physical Review E*, vol. 75, Feb. 2007.

[27] S. P. Borgatti and D. S. Halgin, "Analyzing affiliation networks," *The Sage handbook of social network analysis*, pp. 417–433, 2011.

[28] U. Brandes and D. Fleischer, *Centrality measures based on current flow*. Springer, 2005.

[29] E. Estrada and N. Hatano, "Communicability in complex networks," *Physical Review E*, vol. 77, no. 3, p. 036111, 2008.

[30] M. E. Newman, "Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality," *Physical review E*, vol. 64, no. 1, p. 016132, 2001.

[31] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[33] G. E. Hinton, "To recognize shapes, first learn to generate images," *Progress in brain research*, vol. 165, pp. 535–547, 2007.

[34] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[35] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[36] P. Ruangkanokmas, T. Achalakul, and K. Akkarajitsakul, "Deep Belief Networks with Feature Selection for Sentiment Classification,"