Semi Supervised Learning for Classification of Forest Coverage Type

Sara Misra, Wenyu Huang, Junyan Pu | Carnegie Mellon University

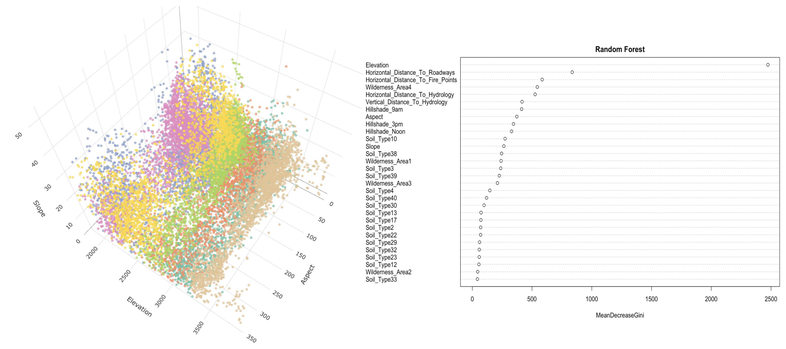
Introduction

The woodland covertype classification is an important problem in the efficient forest conservation. This is an excellent example of a semi-supervised learning problem as we have small samples of labelled data and significantly larger unlabelled samples with multiple features. In our data set we are classifying 7 types of forests using 10 different features.

Forest Types:

Spruce/Fir	Lodgepole Pine	Ponderosa Pine
Cottonwood/Willow	Aspen	Douglas-fir
Krummholz		

Feature Analysis



The figure on the left shows a scatter plot between the first three features in the dataset, colored by the forest cover type. We can get a rough idea on how Forest Cover Types are separated given the features. The figure on the right shows the rank of feature importance using Random Forest.

Method

Expectation-Maximization (EM)

We are using Bernoulli Naive Bayes as the classifier. Moreover, we use EM algorithm to find local optimum classifier. L is a set of labeled data, U is a set of unlabeled data. There are in total K=7 possible labels.

$$L = \{\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_n\}, Y = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}, U = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m\}$$
$$\text{where } \mathbf{x}_i, \mathbf{z}_i \in \mathbf{R}_i^p, \mathbf{y}_i \in \{1, 2, ..., K\}$$

E-step: for each $\mathbf{z}_i$, $\mathbf{w}_i = \mathtt{max}_{\mathbf{w}_i} P(\mathbf{w}_i \mid \mathbf{z}_i)$

M-step: $\Theta^* = \mathtt{argmax}_\Theta \sum_{i=1}^{m} logP(w_i, z_i \mid \Theta)$

S3VM

S3VM is short for Semi-Supervised Support Vector Machines, which aims to find a separator that maximizes the margin on both labeled and unlabeled data points. Specifically, we try to minimize:

$$\frac{1}{l}\sum_{i=1}^{l}\mathcal{L}^1(y_i', f(\mathbf{x}_i)) + \frac{\lambda'}{u}\sum_{i=l+1}^{l+u}\mathcal{L}^1(y_i, f(\mathbf{x}_{l+i})) + \lambda||f||_{\mathcal{H}}^2.$$

where L represents the loss function, f represents the decision function, l represents the number of labeled instance and u represents the number of unlabeled instance. As shown in Figure 2, we used differentiable surrogates for hinge loss function, and used gradient based Quasi-Newton framework to perform the optimization and find the optimal solution f.

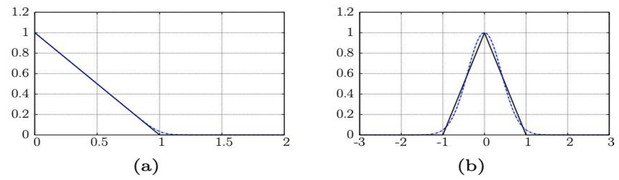
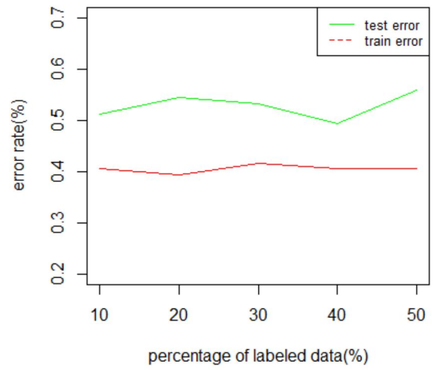
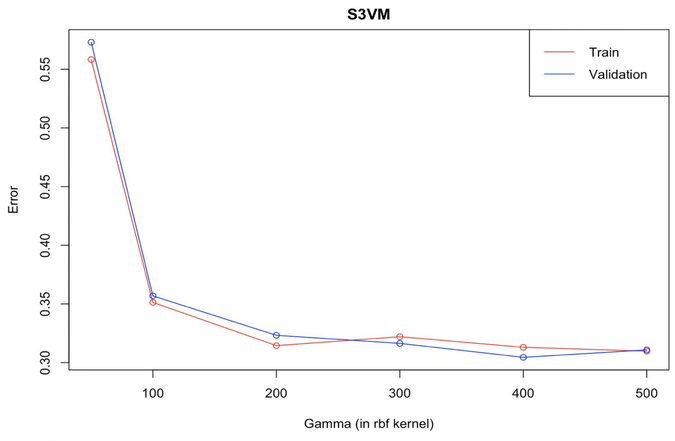


Figure 2: The hinge loss $\mathcal{L}(y,t) = \max(0, 1-yt)$ and its differentiable surrogate $\mathcal{L}(y,t) = \frac{1}{\gamma}\log(1+\exp(\gamma(1-yt)))$ with $y = +1$ and $\gamma = 20$ are shown in Figure (a). The effective hinge loss function $\mathcal{L}(t) = \max(0, 1-|t|)$ along with its differentiable surrogate $\mathcal{L}(t) = \exp(-st^2)$ with $s = 3$ are shown in Figure (b).

Results



The figure above shows the testing and training error in EM algorithm. It is tested on different portions of labeled data, ranging from 10% to 50%. We can see that overall there is no big difference across size of labeled data. In addition, there is a relatively constant gap between testing and training error.

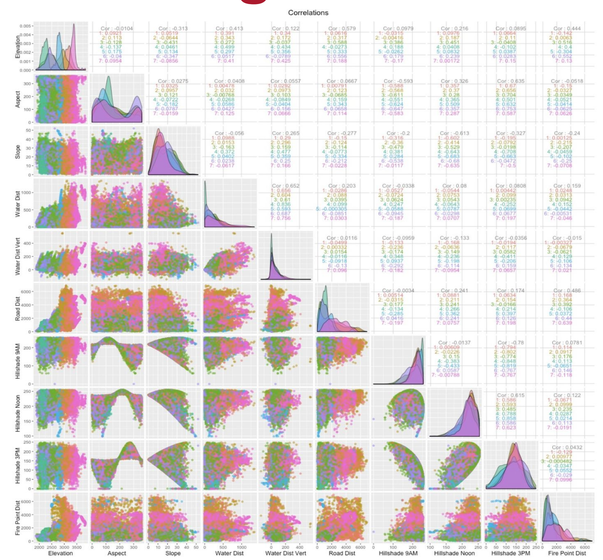


The figure above shows the curve of training and validation error versus different value of gamma (tuning parameter in rbf kernel).

Performance of S3VM:

Using S3VM with Linear Kernel, the accuracy on the test set is 59.31%. For S3VM with RBF Kernel, the highest accuracy on the test set is 69.49%.

Co-training



The figure above shows the correlation between ten continuous features color by the forest cover type. We can infer from the plot that some features are conditionally independent given their cover types. The conditional independence between features meets the basic assumption of co-training.

Conclusion

We have improved our classification using S3VM over our EM Baseline algorithm, going from 55% accuracy to 69% accuracy in terms of classification on a 20% labelled data training set. However, we will be looking into improving this accuracy as described below.

Future Work

For further work, we plan on:

1. Manual tuning of weight parameters for EM Baseline, as while we have importance based on Information Gain of each feature with respect to the label there is no calculation for the optimal weights for the algorithm.

2. Implement multi-class co-training algorithms using deep learning network. We will try to use insights from feature analysis to select views because we observed some conditional independence between features.

Acknowledgements

We would like to thank Prof. Leila Wehbe and Brynn Edmunds the TA staff, and our project mentor Gi Bum Kim, who have supported us in this project for their advice and guidance.