CIS 520 Project Final Report

Woodpecker (Xiang Deng, Yiren Lu, Dongni Wang)

Fall 2015

For the final project, we developed a system for gender prediction (male/female) from the language of their tweets and the image they post with their twitter profile. We were given a training set of 4998 labeled training samples and a testing set of 4997 testing samples. Each sample has 5000 words features, 7 pre-extracted image features and 30000 raw RGB image pixel features.

In our system, we used seven classifiers on different feature sets and combined them using the stacking method. The six classifier are: a logistic regression model on words features, an ensemble model consists of 300 decision stump trees using LogitBoost on selected words and image features, a SVM model with intersection kernel on selected words and image features, a SVM model with intersection kernel on selected and normalized words and image features, an ANN model with 2 hidden layer each with 100 and 50 nodes, a SVM model with RBF kernel on PCA-ed HOG features on face-detected images, and a SVM model with RBF kernel on PCA-ed LBP features on face-detected images. For the stacking method, we took the raw outputs (probabilities) of the six basic models mentioned above trained with 80% of training samples and trained a logistic regression model using the other 20% of training sample. Our final full model achieved an overall accuracy of 92.42%. In order to meet the time and space constraint for the competition, we dropped the SVM model on PCA-ed LBP features and replaced the SVM model on PCA-ed HOG features with one bagging of logistic regression classifiers on raw HOG features. The submitted model for final competition achieved an accuracy of 91.04%.

In the following sections, we present the cross-validation accuracies of each methods we tried and discuss the rationale of our final model. We also provide some interesting visualization such as the most predictive words and the visualization of auto-encoder.

1 Methods

In this section, we report the results of multiple methods we tried for feature extraction, dimension reduction, and classification.

1.1 Data preprocessing

1.2 Feature Selection

To extract features from the raw word and image features, we experimented with multiple feature selection methods, including Information Gain, BNS

1.3 Dimension Reduction

1.4 Classification

2 Experiment Analysis

In this section, we analyze the results of our experiments of multiple methods for feature extraction, dimension reduction, and classification.

2.1 Feature Selection/Extraction

2.2 Dimension Reduction

2.3 Classification

The table of approaches and their associated 5-fold cross-validation classification accuracies are shown in the Table 1

Approach			Accuracy (%)
Feature	Dimension Reduction	Classifier	
Words + Image features	PCA(500)	Ridge Regression + Sigmoid	≈ 70%
Words + Image features	PCA(320)	Ridge Regression + Sigmoid	≈ 79%
Words + Image features	PCA(2000)	Logistic Regression	≈ 85%
Words	None	Logistic Regression	85.96%
Words	IG(1000)*	Logistic Regression	85.79%
Normalized-Words	None	Naive Bayes	72.25%
Words	IG(100)	multinomial Naive Bayes	77.95%
Words	None	Bernoulli Naive Bayes	79.59%
Words	IG(350)	Bernoulli Naive Bayes + EM	82.49%
Words	PCA(2000)	Artificial Neural Network	≈ 86%
Words	IG(76)	K-Nearest Neighbor (L2)	72.89%
Words	IG(84)	K-Nearest Neighbor (Minkowski)	71.43%
Words	IG(95)	Random Forest	83.32%
Words	None	K-means	≈ 60%
Words	IG(1000)	Decision stumps + LogitBoost	89.11%
Face-detected Image RGB	PCA(100)	Random Forest	≈ 69%
Raw HOG features over Face-detected Image RGB	None	Logistic Regression	≈ 80%
Raw HOG features (Face/eyes/ nose-detected Image RGB)	Logistic Regression	None	≈ 81%
Raw HOG features (Face/eyes /nose-detected Image RGB) + Gaussian Pyramid	None	Logistic Regression	≈ 82%
Raw HOG features (Face/eyes /nose-detected Image RGB) + Gaussian Pyramid	None	SVM (RBF kernel)	≈ 84%
Dense LBP (Face-detected Image RGB)	None	SVM (RBF kernel)	≈ 85%
*IG represents Information Gain			

Table 1: Experimental results of single classifiers

	Approach	
Feature	Preprocessing	C
Words + Image features	IG(1000) for trees	Logistic + Neural Netwo
Words + Image features	IG(1000) for trees	Logistic + Neural Netwo
Words + Image features + Image RGB	IG(1000); Face detection + PCA(100)	Logistic + Neural Network + Ensemble
Words + Image features + Image RGB	IG(1000); face-HOG	Logistic + Neural Network + Ensem
Words + Image features + Image RGB	IG(1000); face/eyes/nose-HOG	Logistic + Neural Network + Ensem

Table 2: Experimental results of ensemble classifiers

3 Visualization

In this section, we include some interesting visualization obtained during the process of analyzing data, training, tuning, and testing our models.