CS412 Project Proposal

March 2017

1 Group Member

He Huang Ye Liu Zhu Wang Congying Xia Lichao Sun Fan Zhu

2 Dataset

Our project is about a competition named "What's Cooking?" on Kaggle[1]. The dataset is provided by Yummy [2]. Food is an important part of most cultures. Some of our strongest geographic and cultural associations are tied to a region's local foods. In the dataset, we have a train file and a test file. The train file includes the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format. In the test file, the format of a recipe is the same as the train file, only the cuisine type is removed, as it is the target variable we are going to predict. In the train file, there are 39774 recipes and 20 kinds of cuisines in total. In the test file, there are 9944 recipes that we need to predict. You can find the competition from the website [3]. And you can download the dataset from the website [4].

3 Machine Learning Task

Our goal with this project is to use recipe ingredients to predict the cuisine.

3.1 Problem definition: Multiclass Classification

- **Features**: all V ingredients, and each recipe is represented as a vector of length V.
- **Classes**: all 20 kinds of cuisines.

[1] www.kaggle.com
[2] www.yummly.com
[3] www.kaggle.com/c/whats-cooking
[4] www.kaggle.com/c/whats-cooking/data

3.2 Machine Learning Techniques:

1. **SVM (multiple classes version)**
2. **Random Forest**
3. **Naive Bayes**
4. **Class-based Association Rule Mining**
5. **Neural Network**
6. **Latent Dirichlet Allocation**

3.3 Visualization:

To reduce V-dimensional vectors to 2- or 3- dimensional ones
Visualization Methods:

- **PCA**
- **t-SNE**

4 Contributions:

1. Using many traditional ML techniques.
2. Compare the performance of each used technique.
3. Visualizing the recipes in 2-D/3-D space to show the closeness of recipes.