Predicting Cuisine from Ingredients

Rishikesh Ghewari
rghewari@ucsd.edu

Sunil Raiyani
sraiyani@ucsd.edu

ABSTRACT

Over the years, people have tried to explore new ingredients and incorporate them into recipes or produce new recipes all together. One of the obvious relations that we would like to explore is the relation between ingredients and cuisines. We use the yummly data-set to study the problem of predicting cuisine of a recipe based on it's ingredients. On testing several classifiers we observed that SVM works best for this prediction task.

Keywords

cuisine prediction; classification

1. INTRODUCTION

Food is an indispensable part of our lives. The most basic element with which one can identify a food item are its ingredients. Ingredients are the atomic components of food. Over the years, people have tried to explore new ingredients and incorporate them into recipes or produce new recipes all together. However, the choice of ingredients is characterized by geographical locality. One of the factors responsible for this behaviour could be the similarity in availability of an ingredient in a particular geographic region. This has resulted in the set of recipes being divided into geographic classes known as cuisines.

One of the obvious relations that we would like to explore is the relation between ingredients and cuisines. It is quite apparent that availability and popularity are important factors influencing the choice of ingredients in a recipe. People in different regions have different taste preferences and hence tend to favor a particular set of ingredients in comparison to the other. Thus, there seems to be a strong co-relation between these two entities.

In this assignment, we will try to develop a model to classify a recipe based on the ingredients it uses.

2. DATASET

The Yummly[1] dataset used for the prediction task consists of 39,774 recipes. Each recipe is associated with a particular cuisine and a particular set of ingredients. Initial analysis of the data-set revealed a total of 20 different cuisines and 6714 different ingredients. Italian cuisine, with 7383 recipies dominates the dataset while brazilian cuisine, with 467 recipes is least dominating.

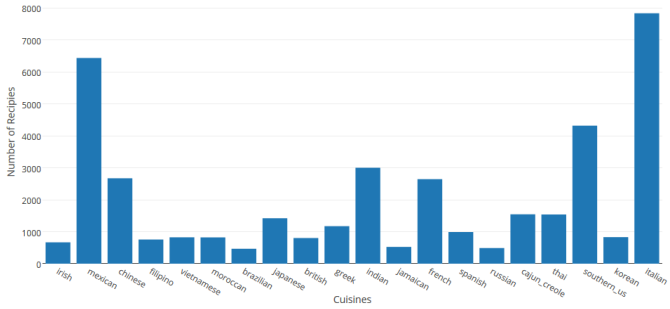

Figure 1: Cuisine distribution

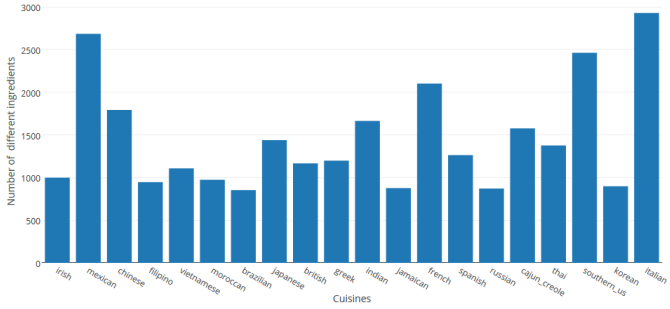

Figure 2: Ingredient distribution over cuisines

A cuisine can often be identified by its distinctive ingredients. The ingredients most associated with each cuisine (using normalized pointwise mutual information) in the dataset are:

- Brazilian: cachaca, acai
- British: stilton cheese, suet
- Cajun / Creole: Cajun seasoning, andouille sausage
- Chinese: Shaoxing wine, Chinese five-spice powder
- Filipino: lumpia wrappers, calamansi
- French: Gruyere cheese, Cognac
- Greek: feta cheese, Greek seasoning
- Indian: garam masala, ground turmeric

- Irish: Irish whisky, Guinness
- Italian: parmesan cheese, ricotta cheese
- Jamaican: scotch bonnet chiles, jerk seasoning
- Japanese: mirin, sake
- Korean: Gochujang, kimchi
- Mexican: corn tortillas, salsa
- Moroccan: couscous, preserved lemon
- Russian: beets, buckwheat flour
- Southern (US): buttermilk, grits
- Spanish: chorizo, serrano ham
- Thai: red curry paste, fish sauce
- Vietnamese: fish sauce, rice paper

The ingredients used by recipes from different cuisines are not very distinct. Often, recipes with different cuisines use very similar ingredients.

For the purpose of visualization, the data features were reduced to 2 dimensions using t-distributed stochastic neighbor embedding (t-SNE). A scatter plot was drawn using these 2-D features. Each point on the plot represents a recipe.

Observing the plot, we notice that Asian recipes appear together in the upper left part of the plot, and there are clear Indian, Mexican, and Cajun clusters, among others. Many other cuisines, however, are highly overlapping, which makes classification more challenging. For example, the center contains a mixture of European (French, Italian, British, Irish, Spanish) and Southern US cuisine.

3. PREDICTION TASK AND FEATURES

For the purpose of this assignment, we sought to predict the cuisine to which a recipe belongs based on its ingredients. We randomly shuffle the data-set consisting of 39774 recipes to remove any existing bias. This shuffled data-set is then divided into training(80%) and validation sets(20%). We use another data-set consisting of 9994 recipes as the test set.

The performance of the model is measured in terms of its prediction accuracy.

$$\texttt{Accuracy} = \frac{\texttt{Number of correctly predicted cuisines}}{\texttt{Total number of predictions}}$$

One of the most naive predictions for this task would be to find out the most frequently occuring cuisine in the training set and always predict this cuisine. In our data-set, it turns out that Italian is the most frequently occuring cuisine. We use an 'All Italian Baseline' to evaluate the performance of our model.

Validation set accuracy: 0.19229
Test set accuracy: 0.19268

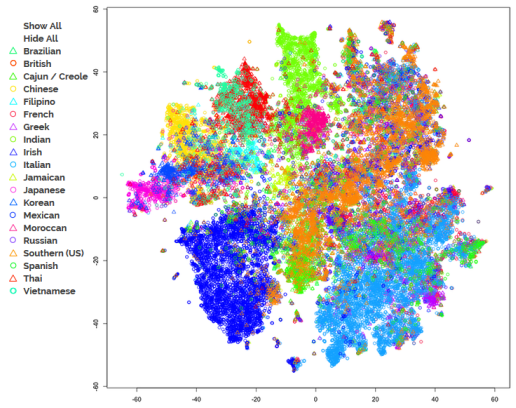

Figure 3: Scatter plot for all cuisines [1]

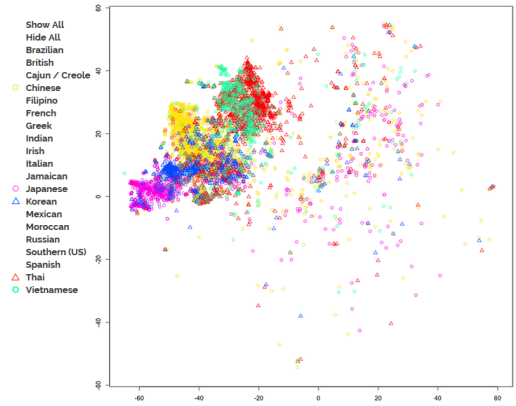

Figure 4: Scatter plot for all Asian cuisines [1]

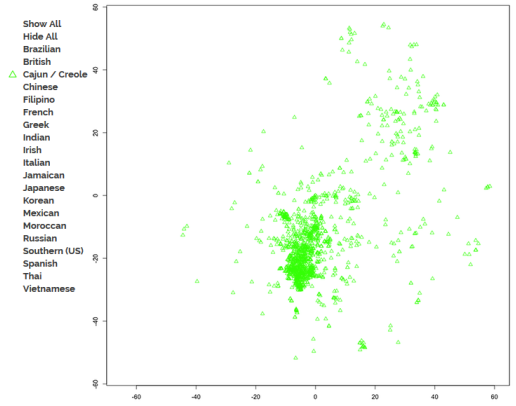

Figure 5: Scatter plot for Cajun cuisine [1]

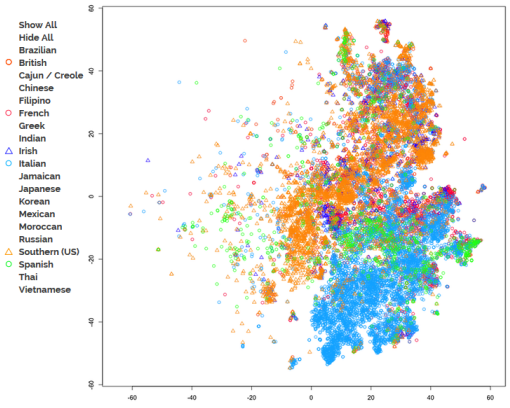

Figure 6: Scatter plot for all European and Southern US cuisines [1]

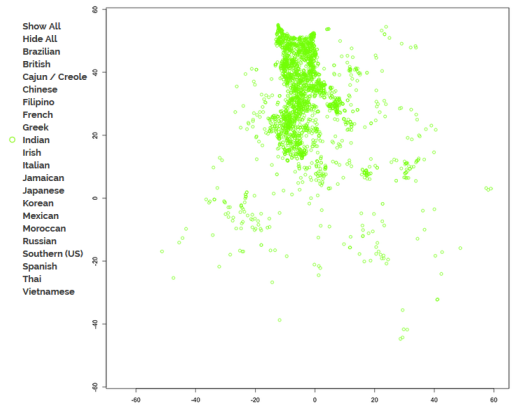

Figure 7: Scatter plot for Indian cuisine [1]

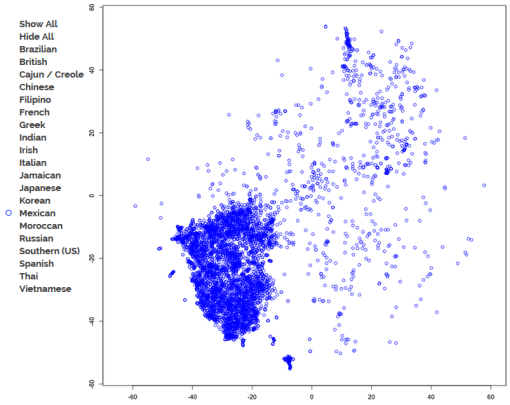

Figure 8: Scatter plot for Mexican cuisine [1]

We used a bag of words model for the ingredients. Firstly, we split the ingredients which had multiple words into separate words and considered each word as a ingredient. In this process we eliminated words which had lengths less than 3 to ignore words like 'of', 'or' etc. We represented each ingredient as a feature in the feature vector. If an ingredient is present in the recipe, the value of the corresponding feature is 1, otherwise it is 0. We removed capitalization to account for duplicate features. Finally, we were left with 2986 ingredients. Now this gives rise to a 2986 dimensional feature.

Furthermore, we used PCA to reduce the number of features to 1150. However, it doesn't help in improving the accuracy. So, we reduced the number of features by removing ingredients that are very rare from the feature vector. We removed all the ingredients with frequency less than 20. This left us with 1150 features. The accuracy obtained after this reduction was better than that obtained after using PCA.

We extended the feature vector by adding 20 more features wherein each feature corresponds to one cuisine. Firstly, we find the set ingredients corresponding to every cuisine.

Let I be the set of ingredients in a recipe. Let I_c be the set of all ingredients corresponding to cuisine c. Now the feature corresponding to cuisine c is $|I \cap I_c|$.

The feature vector was extended further by adding 20 more features similar to the previous 20 features. In this case, the feature corresponding to a particular cuisine represented the number of signature ingredients of that cuisine present in the recipe.

Finally, we used this 1190 dimensional feature vector to perform our prediction task.

There were certain feature representations which didn't work. When we tried stemming the ingredients to reduce the number of features e.g representing 'eggs' and 'egg' as the same feature, the accuracy of the resulting model was adversely affected. Similarly, we also tried representing each feature by its TF-IDF value rather than just 0 or 1. Here, the TF is one for all the ingredients. So we just divided by the document frequency of each ingredient. Surprisingly, this also resulted in reduced accuracy of the model.

4. MODEL

We used a multi-class Support Vector Classifier to classify the recipes. To optimize the model, we used tuned the slack variable C on the validation set. The performance of the model was best observed for a value of C = 0.05

We have represented each data point in the data-set using ingredients as features. After visualizing this data-set we realized that the data is characterized by soft boundaries that separate different cuisines. In such a case, intuitively, one would choose a SVM classifier for the classification task.

We tried various multi-class classifiers like Naive Bayes, Logistic regression, Decision Tree, Random Forest, SVM and k-Nearest Neighbors. The results that we obtained confirmed our intuition about the characteristics of the data-set

and choice of the classifier.

Logistic Regression performed almost as well as SVM classifier, however, SVM gave a slightly higher performance.

5. LITERATURE REVIEW

Han Su et. al.[2] have worked on investigating if the recipe cuisines can be identified by exploiting the ingredients of recipes. In their paper they treat ingredients as features. Their study provided insights on which cuisines are most similar to each other. Also finding common ingredients for each cuisine. Most of the work in cooking related research has been on recipe recommendation and retrieval. Ueda et al. [5][4] proposed a personalized recipe recommendation method based on user's food preferences . A user's food preference in terms of ingredients is derived from his/her recipe browsing activities and menu planning history. Most models use ingredients to model recipes Wang et al. [6] model cooking procedures of Chinese recipes as directed graphs and proposed a substructure similarity measurement based on the frequent graph mining. Yang et al.'s[7] first identified the ingredients, gave each ingredient a probability label, and then used pairwise local features among the ingredients to determine the food category, by calculating the distance, orientation, and other properties between each pair of ingredients. Teng et al[3] have studied substitutable ingredients using recipe reviews by creating substitute ingredient graphs and forming clusters of such ingredients. There are various successful recipe recommendation systems there is very little work on analyzing the correlation between recipe cuisine and ingredients.

6. RESULTS

The accuracy reported on validation and test sets for different models are as follows:

Model	Accuracy (Validation Set)	Accuracy (Test Set)
SVM	0.81315	0.78228
Random Forest	0.74131	0.73803
Naive Bayes	0.73251	0.72335
Logistic Regression	0.81253	0.78158
KNN	0.60746	0.60358
Decision Tree	0.62589	0.62691

7. FUTURE WORK

The problem of finding cuisine based ingredients is very much like topic modeling. We could try various techniques that we use in topic modeling like LDA to model cuisines. Also instead of using individual ingredients as features we could try an n-gram model. This would increase the number of features but we can use PCA to get a lower number of features. There could be a low dimensional structure in cuisines with respect to ingredients. We could do SVD on these to find the low dimensional structure.

8. CONCLUSION

We observed that both Logistic Regression as well as SVM classifiers perform equally well in the prediction task. This better performance of these classifiers over others can be attributed to the soft boundary characteristic of the dataset. We saw that the bag of words model on ingredients works well in this task. This leads us to realization that the problem of predicting cuisines from ingredients is similar to the traditional topic modeling task. Future work would be exploring this similarity.

9. REFERENCES

[1] Understanding cuisines using a new dataset from yummly. http://www.yummly.com/insights/understanding-cuisines, Dec. 2015.

[2] H. Su, M.-K. Shan, T.-W. Lin, J. Chang, and C.-T. Li. Automatic recipe cuisine classification by ingredients. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 565–570. ACM, 2014.

[3] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 298–307. ACM, 2012.

[4] M. Ueda, M. Takahata, and S. Nakajima. Recipe recommendation method based on userâĂŹs food preferences. In *IADIS International Conference on e-Society*, 2011.

[5] M. Ueda, M. Takahata, and S. Nakajima. UserâĂŹs food preference extraction for personalized cooking recipe recommendation. *Semantic Personalized Information Management: Retrieval and Recommendation SPIM 2011*, page 98, 2011.

[6] L. Wang, Q. Li, N. Li, G. Dong, and Y. Yang. Substructure similarity measurement in chinese recipes. In *Proceedings of the 17th international conference on World Wide Web*, pages 979–988. ACM, 2008.

[7] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2249–2256. IEEE, 2010.