What is Cooking?

Authors: Alexander Bobowski & Tyler Paulsen

b. Who has done this before.

The project is on Kaggle.com, so many other data scientists have attempted the classification. For the project, our team will only be classifying one or two types of cuisine to simplify the initial problem. If the project seems to be going well, the algorithm may be extended to classify all cuisines. The leading classifier is currently at a .82090 correct classification rate (Kaggle). The difficulty in creating an accurate classifier, evidenced by the low correct classification rate, shows that there are many potential issues with the data that need to be addressed.

c. What challenges did they face

There have been a number of different challenges faced by the competitors. Most of the proposed solutions for this project have used the bag of words approach, while others models suffering from overfitting (Ashar). Given that most of the solutions to this data mining exercise have implemented the Bag of Words algorithm, most of the discussion has focussed on preprocessing of ingredient strings for efficient implementation.

Observing posts on the Kaggle forums relating to this challenge, most of the competitors struggled with interpreting ingredient strings. Some of the more interesting techniques used statistics and dictionary libraries to determine the stem of a word (Chiu). For example, the stem of a word for “thighs” would be “thigh”. The stem can be taken further by guessing words that are similar, but just have added adjectives; for example, “ground beef” vs. “lean ground beef”. (Orfano)

Some competitors experimented with what data to remove from the set. Some have removed ingredients in almost every recipe -- such as salt -- and others removed ingredients that only appeared in a handful of cuisines.

d. Are there any ethical considerations.

Yummly.com has set this up with no reward other than knowledge. The company could be using the top algorithms to generate revenue with a new feature added to the site, or improving an existing one. This does not raise issues with the creation of a classifier, but rather the concept of the competition itself.

e. Is there a business case for this work?

Kaggle has supplied a business case for the users of the website to follow (Yummly). The business case says who provided the dataset, and the rules for the competition. Unlike other competitions on kaggle, this one does not have any prizes for the top solutions. Kaggle also provides a forum for the competitors to show off their code and to discuss about the different problems with their current implementation.

f. What issues are there with the data

The dataset provided by Kaggle has many issues that may not appear at first glance. The data is all text based, and this presents major problems if not dealt with. The data is provided as a json list by Yummly. The list is broken up into two major categories: what our team is classifying (cuisine), and the ingredients. The data seems to be pulled directly from their site without any prior cleaning.

Since the data is all natural language based, there are going to be major issues with the text not being spelled correctly, similar words, and special characters (&, $, #), etc. The forums present an elegant solution to these problems: using a function to measure the similarity between words and grouping similar ingredient strings.

The algorithms presented in class have dealt with perfectly formatted data with a limited number of possible attribute values: weather=(rain,hot,cold) or power=(on,off). The provided data set has one attribute that will have to be classified on, and attributes can take on any combination of ingredients: cuisine=italian, ingredients=”salt,sauce,pepper,...” and cuisine=”indian”,ingredients=”salt,curry,...”.

The first step in homework assignments has included data visualization, usually in the form of a scatter plot. Given that the data provided for this challenge is natural language strings, visualization is more difficult than with numerical data. Additional preprocessing will have to be performed in order to create a meaningful visualization. Data visualization in this case may prove to be more time consuming that it is worth.

The data set provided by Yummly is very large, so the initial classification scheme should be limited in scope. By limiting the initial scope of the project and just classifying one or two cuisines, the problem can be simplified and a the accuracy of an experimental scheme can be tested without extending it to all possible data where more overlap between ingredients exists. The provided data set has over 39,000 entries, and as such it is out of the scope of the project to classify them all. Through visualization of the data, a meaningful decision about what cuisines should be included in our initial model can be made.

g. How clean is the data?

The data provided from Yummly is very dirty; cleaning it presents an interesting challenge. Any true solution will require both manual and automated data cleaning. Creating an automated solution that successfully avoids false positives is necessary, as improperly renaming an ingredient will throw off the model and create an incorrect classifier.

Others who participated in this challenge attempted automated data cleaning of ingredients. Michael Goettsche, a competitor in the initial challenge, settled on a scheme which combined Levenshtein distance and cosine distance to create associations between ingredients present in the data set that are likely the same. By tweaking the thresholds involved in this solution, accuracy can be fine tuned. This list can then be manually inspected to remove false positives (Goettsche).

As Michael states, this solution successfully combines some similar ingredients but is not sufficient for all classification. In the example he gives of mozzarella cheeses, the names of ingredients vary so much that automated combination would be very difficult. In these cases, manual combination and substitution of ingredients would be required. Simple search-and-replace on ingredient names after manually viewing the results and filtering improper matches would supplement automatic combination.

Depending on the classification approach attempted, different types of data cleaning will need to be performed. Removing ingredients common across all cuisines, such as salt, is useful in some classification strategies to prevent overfitting. Other times, it may not be necessary. As such, how the data is cleaned and presented may need to be altered depending on the classification method used.

h. What methods and algorithms were used in the past?

Though the competitor's preferred models have not been covered in this class yet, they seem to be based on patterns we have studied. The model of choice has been the bag of ingredients approach; this is a variation of the bag of words (Kappa).

The bag of ingredients approach works off of grouping each classifier with a bag. The bag is then filled with attributes that can help classify the bag. The most important aspect of this model is its ability to group ingredients while disregarding grammar and word order (Bag-of-words Model).

i. Other issues related to the project.

On-the-fly preprocessing of ingredients for recipes presents an interesting challenge. For training data, automatic and manual data cleaning can be used to produce the best classifier. In the final implementation, all combination of ingredients and data processing must be done automatically. For decision tree implementation, this means that we must create a scheme that is able to reliably interpret ingredients and match them to decision points. Using the bag of words algorithm, creating relationships between ingredient list entries does not depend on exact matches.

j. What distance metrics are used for comparing data records?

A distance metric that can measure strings must be used to compare the data records. Some of the common string distance metrics are: Levenshtein distance, Jaccard coefficient, and cosine distance. For the problem on which our team working, it may be beneficial to find a library that can combine different types of string distances to provide a better result. The programming language R has a function, strdist, that uses several metrics for this including the well-known Levenshtein distance, cosine similarity, and even the phonetic Soundex algorithm (Goettsche).

k. How do you plan to solve the problem?

The bag of ingredients approach will be the easiest if the team wants to solve the entire dataset. Once one “bag” is created, it should be easy to extend the algorithm to other cuisines. This model’s approach is to create a bag for each type of cuisine. For each cuisine, create a bag containing the most used ingredients. When given an unclassified datapoint, look at each bag and determine what bag it is most like. The key to this model is the preprocessing of data. The team will need to determine which ingredients should be ignored during classification as well as reduce equivalent ingredient strings to a single entry (for example, Mozzarella vs Low Fat Mozzarella, Mushrooms vs. Mushroom, etc.).

l. How will you validate your results?

Since the data set is large, we can reduce it to the two or three types of cuisines we want to classify. The resulting dataset will then be split up further into a training, testing , and validation set. For the training set, we will filter noise datapoints and preprocess the data to produce the most accurate classifier possible. Once the team is confident that our model is correct, they can run it against the validation dataset to determine the accuracy of their classifier.

If we are able to classify all of the cuisines, we can run our model vs the given unclassified data set given by Yummly. The test dataset will be graded by the correct classification on Kaggle's server to determine the accuracy of the model.

m. What three algorithms from our course do you plan to use?

Three classification algorithms that have been taught in the Principles of Data Mining course that can be implemented. Since the problem involves classification, our team plans to try building a decision tree, implementing the naive bayes algorithm, and using k-nn. The Bag of Words algorithm may be used in conjunction with one or more of these classification schemes to represent our datapoints. If one of these methods proves successful for our reduced dataset, our team would could extend the scope of the project to include all cuisines present in the initial dataset.