TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

Classification of Web based Recipes using Random Forests Algorithm

[1] Isura Nirmal, [2] H. A. Caldera
[1], [2] University of Colombo School of Computing, Colombo, Sri Lanka.

Article Info
Volume 82
Page Number: 12731 - 12737
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 24 February 2020*

Abstract:
Food plays a major role in human civilization as not only being a physiological need but being a major factor for defining the culture and society. Preparation of a dish is depending on country and have several unique features in it named as its cuisine. Recipe, being the instruction manual of a food preparation reveals the specific cuisine it belongs to. Web based recipe portals nowadays contain the thousands of recipes from all over the world. Even though there are many published and unpublished works, majority of them haven't gone deep in to tuning the classifier and looked for the best features when predicting the cuisine. Thus, our work aims to apply Random Forests algorithm using a large web based recipe data set in classifying the recipe to its originating cuisine and revealed that the cuisine of a given recipe can be predicted in ~79-80% accuracy.

Keywords*: Random Forest, Principle Component Analysis, Food Classification, Cuisine identification.*

I. INTRODUCTION

The term "Cuisine" tells the style of cooking, use of ingredients for food preparation which is appropriate for serving. Every region of the world, country and ethnic group has its own cuisine. Recipe can be called as an instance of a cuisine which contains ingredients as well as cooking instructions indigenous to that cuisine.

Cooking process involves with several decision making steps such as the correct amount to add and how much time to cook. This is non trivial as the some spices are used to enhance the flavor and eggs are often used to maintain the thickness.

This given recipe can be traced back to its origin country because of the unique features it yields. Therefore, it's worthy to study the ability to extend the classification technique in machine learning to classify a given recipe for its cuisine.

Nowadays, web based recipe portals contain the thousands of recipes from all over the world. Single food may have several recipes including different flavor choices of users. People often put their own version of the recipe on the web that has surprisingly become tasty. In scientific point of view, the combination of ingredient and cooking procedure must have resulted this tasty outcome. In the light of this, many researches have emerged recently from matrix factorization, classification, clustering and network based analysis fields. However, the laborious text book related data collection and data curation, the size of the sample and feature vector seemed to have limited the ability to understand the deeper bindings inside the recipe spaces [1] [2] [3]. Even though there are many published and unpublished works which are leveraged by the competitions from sites like Kaggle[1] [4], majority of them haven't gone deep in to tuning the classifier and look for best features when predicting the cuisine and analyses the performance of the chosen classifier.

Thus, our work aims to analyze the performance of tree based ensemble classifier Random Forests for a reasonably larger dataset in classifying a given recipe to its originating cuisine.

II. RELATED WORKS

According to the Neurology, flavor is mainly a

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

synergetic stimuli of both taste stimulus from the gustatory receptors (taste buds) in tongue and the olfactory receptors inside the nasal cavity [5] [6] [7]. However, [8] has given the chemical somatosensory stimulus a similar importance for flavor perception. Heat, crispness, color and the texture of the food also have a positive effect on human flavor perception levels [9].

Ahn et al. [10] tried to prove a food pairing hypothesis among the western cooks using the ingredient pairs and their flavor compounds which were extracted from web recipe portals. This study has gone a greater length in incorporating the flavor chemicals in to the study to prove the food pairing hypotheses ("Ingredients sharing flavor compounds are more likely to taste well together than ingredients that do not"). The hypothesis was proven true for the Western cuisine and false for Eastern Cuisines. This study was a milestone in the literature and after that several other related works supported this claim [2] [11]. This work supports our work by inferring the fact that the incorporation of cuisine related data leads to a better recommendation models.

The early appearances of Computational recipe planning can be found in CHEF model which was published in 1980s [12]. It replicated the human cognitive process in culinary decision making in Szechwan Cooking.

Recipe recommendation and user ratings prediction have been the other areas of interest in many published works. Freyene et al. [13] and Forbes et al [14] primarily worked on the user ratings prediction for a newly generated recipe.

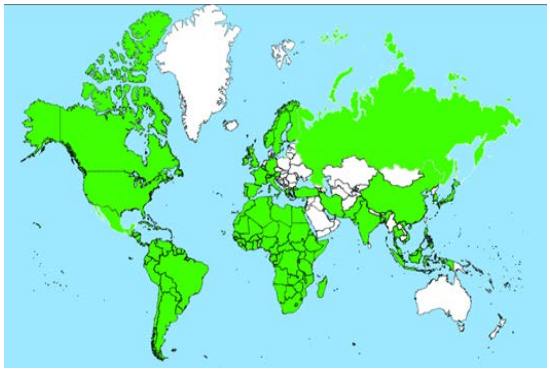

Figure 1: Cuisine Coverage Map

Table1: Cuisine-Ingredient Model

Cuisine	Ingredient 1	Ingredient 2	Ingredient 3	… …	Ingredient 381

Nonetheless, Teng et al. [15] used a web scraped recipe data set to implement an ingredient network, cooking procedures and substitution of the ingredients. The latter has given 79.2% accuracy for the user ratings classifier. These works ignored the fact that the cuisine specific features need to be understood in order to make a realistic dish. These studies all together didn't consider the feature richness of the data set, coverage of cuisines and performance of classifiers.

Observing the lack of identification of the cuisine at the beginning of recommendation models, Kim et al [3] highlighted that each country or ethnic group used unique ingredient set and those ingredient combinations are showing spatial relationship. The methods used were hierarchical clustering, ingredient network analysis and classification. Although the size of the data set is not large (5917 recipes), it has balanced cuisine coverage. The methodology used for data curation is not reliable as the volunteers' general culinary knowledge was used to recognize the cuisine of a given recipe.

We were able to uncover that considerable number of researches have studied machine learning and data mining applications in culinary domain but no work has been considered the rigorous analysis of data set and algorithms for classification task. In our study, we are trying to find the optimal and maximum settings to use for recipe classification task.

III. DATASET AND DATA MODELLING

A. Data Set

We have used the data set used by [10] and the Figure 1 depicts the coverage of the worldwide cuisine of the data set. The data set has total number of 54,498 recipes. Figure 2 depicts percentage of recipes from each cuisine.

Our data model is tabularized as follows. Table I portrays the occurrence matrix (binary matrix with

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

filling degree of 2.16%) which has a feature vector of 382 (381 ingredients and the cuisine).

Our preliminary analysis on the recipe dataset revealed that there are outliers in each cuisine but in tolerable levels (North American Cuisine had 14 and South Asian cuisine had no outliers. Other Cuisines also reported outliers in 5 to 7 range).

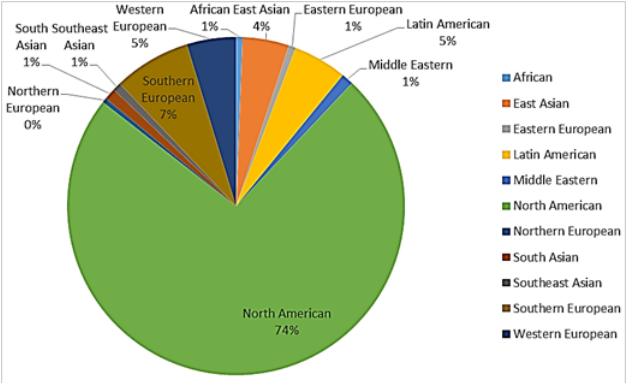

Figure 2: Number of Recipes

The main reason for outliers are as obvious as those recipes actually exits in the cuisine but rare. Thus, removing such outliers lead to an information loss.

The other major characteristic in the data was the large class imbalance. North American data set consisted of 41,524 recipes against total 56,498 total recipes which is 73.49 of the total set. (Figure 2)

B. Methodology

For predicting the Cuisine, we have utilized Random Forests algorithm to classify the given recipe and predict the cuisine.

Random Forests Classifier was selected due to the fact that the dimension of the feature vector is significantly high (381) and the high class imbalance problem discussed earlier in this work. Over fitting and the deviation of results due to the outliers are well handled by Random Forests Classifier because the classification model is generated with dense randomness with Bootstrapped samples. Random forest Classifier uses Bootstrapping; making random resamples of size N sample with replacements. (Set of recipes may appear over multiple times in each bootstrapped sample [16]. The size of the data set is high (54,498) and the sparsity of the data set turned out to be very low (2.16%).

Since we are reducing the effects of class imbalance, classifier related controls as Algorithmic Ensemble Techniques are applied. We tuned the algorithm to select Bootstrapping and Out-of-Bag. We considered the accuracy score as the main output from the classification model providing the ratio of how many recipes are correctly classified in to their correct cuisines and the total number of recipes in the cuisine. Our baseline accuracy score is 0.7349 (73.49%) (ZeroR algorithm). For the Random Forest implementation, we chose sklearn's implementation[2].

Principle Component Analysis (PCA) is fundamentally a dimensionality reduction algorithm for both classification and data visualization. We specifically used PCA to address the high dimensionality characteristic in our dataset when classification occurs. PCA would not only reduce the high dimensionality but identify if there are redundant features.

We have utilized two major approaches for identifying the best classification results in the data set. First, we ran Random

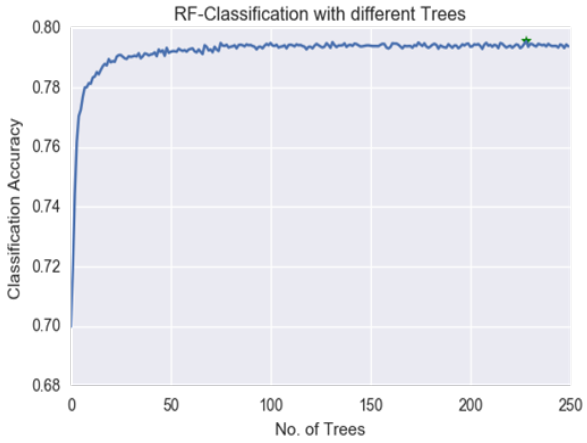

Figure 3: Classification with different trees (10-fold CV)

Table II: Classification Summary for different Trees

	Accuracy
Min	0.699780
Max	0.795460 (marked as asterisk)
Mean	0.791686

Forests classifier without any further data curation with different foldings (up to 100 folds) and number of trees (up to 250 trees) to analyses the behavior

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

and identify the most optimal and maximum points. Secondly, we applied PCA to reduce the high dimensionality of the recipes data set and project them into the manageable number of dimensions and analyzed the performance. Finally, the identified maximum and optimal settings in the first model are applied to Random Forests model with Principle Components to verify whether any improvements can be achieved in classification accuracy.

IV. ANALYSIS OF THE RESULTS

A. Optimal Settings

To observe the performance of the Random forests classifier in different settings to find the optimal setting, we have first applied 10-fold cross validation (CV) for varying numbers of trees. Figure 3 illustrates the accuracy of the classifier against the number of trees. First, we continuously incremented the number of trees up to 250 and then checked the accuracy by incrementing numbers of trees by 100 up to 1000 (Figure 5) This two processes (100 times up to 1000 trees) was separated due to the fact that growing a forest of this scale takes high computational power and time.

As we can observe, Classification Accuracy starts from 0.7 and saturate from 50 trees and remain ~0.79 throughout the plot (std= 0.008911). Figure 5 further proves that even though the number of trees reaches 1000, it doesn't have an impact on accuracy. Maximum accuracy hit was same with 700 trees grown.

As tabularized in Table II, the mean accuracy (0.791686) was first hit at 50 trees and the max accuracy (0.79546) was hit at 228 trees. Therefore, we conclude that the 50 trees for this data set is the optimal setting and 228 as the maximum setting.

To check whether the number of cross validation may have

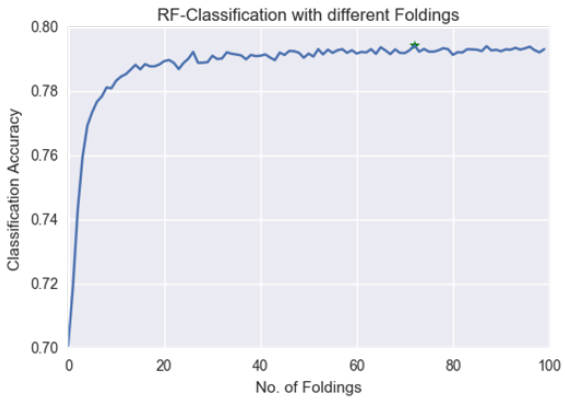

Figure 4: Classification with Different Foldings (100 trees)

Table III: Classification Summary for different Foldings

	Accuracy
Min	0.700750
Max	0.794440 (marked as asterisk)
Mean	0.788074

an effect on the accuracy, we ran the model on different folding starting from 1 to 100 (fixing the number of trees in the forest to 100 trees). Figure 4 visualizes that the trend stabilized around 50-fold cross validation and doesn't show significant improvements (std = 0.013290). The maximum accuracy hit was at 72-fold cross validation (0.79444) and mean accuracy was first hit at 15-fold cross validation (0.788074) as tabularized in Table III.

Analyzing the varying numbers of cross validations and number of trees while fixing one variable mainly gave the idea about the point where the accuracies are saturating. Although, computation power to identify those points are great, this will drastically reduce the computational power needed in the next steps as the wastage beyond these settings prone to have no improvement of the results.

B. PCA analysis for dimensionality reduction

In PCA, we first analyzed the cumulative explained variance increasing number of components (1 to 381). Explained variance is a major output from the PCA model which gives the idea of how much information is reflected from the

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

newly generated each principle component. Cumulative explained variance provided the sum of explained variances which gives an insight about total coverage of the information. However, cumulative explained variance hasn't reached to 100% before the number of principles components equals with the actual feature vector (Figure 7). By the light of this, we can conclude there are no redundant dimensions in our data set. However, 90% of information can be retained with 113 principle components which is more than three times dimensionality reduction. Therefore, there are dimensions in this dataset which are less important. Figure 7 further depicts the key points identified in the cumulative explained variance plot.

The accuracies generated by the key points identified in the PCA is tabularized in Table IV.

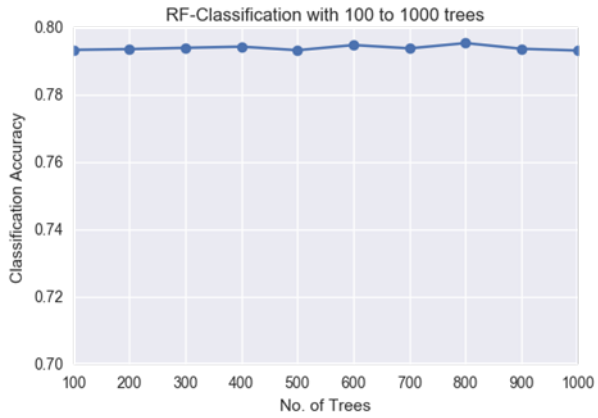

Figure 5: Classification with multiples of 100 trees (10-fold CV)

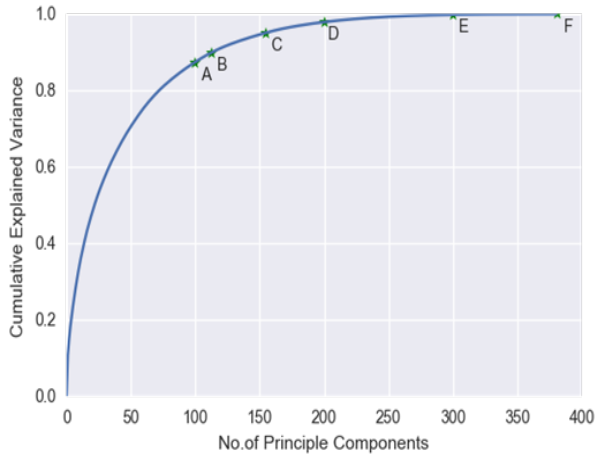

Figure 7: PCA Analysis

Table IV: Cumulative Explained Variances and Classification Accuracies

Principle Components	Cumulative Explained Variance	Accuracy (10-fold/100 trees)	Accuracy (72-fold/228 trees)
A(100)	0.8728	0.78130	0.78231
B(113)	08987	0.78162	0.78132
C(155)	0.9505	0.78091	0.78171
D(200)	0.9789	0.78105	0.78314
E(300)	0.9984	0.78286	0.78374
F(381)	1.0000	0.78274	0.78357

The main observation was that the classification accuracy didn't significantly (std= 0.000852) vary depending on the number of principle components. This infers the fact that a substantial level of accuracy can still be reached with less principle components. Using 1/3 of the components (113 principle components retained 90% of information) gave 0.78162 accuracy where all the components gave 0.78274 accuracy which is only 0.00112 increment. For this experiment we used only 10-fold cross validation with 100 trees in Random Forest as the fixed variables to measure the effect only from increment of Principle Components. But our preliminary results suggested that 72-fold cross validation and 228 trees gave the maximum results. As the final test on our data set, we ran the model with that setting.

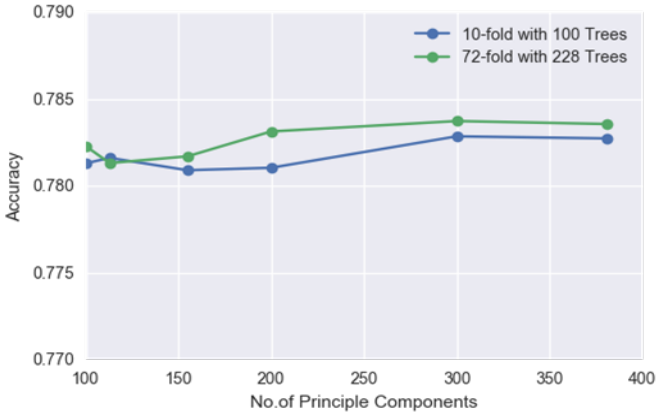

Figure 6: PCA Components-Classification Accuracy

The results were promising. It was performing better than the 10-fold cross validation in 100 trees settings even in Principal Components. The last column of Table IV tabularizes the findings and

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

Figure 6 visualizes the results against the 10-fold cross validation with 100 tress test.

This further confirms that the best tuning parameters for Random Forests model for this data set is 72-fold cross validation with 228 trees. Even with size of 100 principal components as the feature vector, the results haven't drifted significantly.

CONCLUSION

During the initial analysis, we uncovered that there are some variable noise and a large class imbalance in the data set. We specifically used the Random Forest classifier because of the algorithmic level control it provides for identified characteristics of the dataset. The baseline for the dataset set is at the accuracy of 0.7349(ZeroR algorithm).

The main research finding is that the 72-fold cross validation with 228 trees proven to be the best settings for this data set. . We achieved maximum 0.79~0.8 Accuracy score in original data set with 381 dimensions.

PCA analysis was very effective when dimensionality reduction yet maintaining tolerable accuracy changes. Using 113 principle components, 90% of information is retained and gave 0.78274 accuracy (10-fold cross validation with 100 trees). This further proves that even though there are no redundant information in the features in original dataset, some features are important than others. Applying the previously identified algorithmic tuning parameters (72-fold cross validation with 228 trees) further improved the classification results obtained by the PCA components.

We choose Random Forest model as only classifier for this classification as the problems identified in this data set was reasonably controlled by the controls provided with the algorithm. Other classification models may further improve the results.

As a future work, we intend on using this classification model for better analyzing the ingredient alteration patterns in a given recipe.

REFERENCES

1. O. Kinouchi, R. W. Diez-Garcia, A. J. Holanda, P. Zambianchi and A. C. Roque, "The non-equilibrium nature of culinary evolution," New Journal of Physics, vol. 10, pp. 1-15, 2008.
2. K. R. Varshney, L. R. Varshney and D. M. Jun Wang, "Flavor pairing in Medieval European cuisine: A study in cooking with dirty data," arXiv preprint arXiv:1307.7982, pp. 1-10, 2013.
3. K. J. Kim and C. H. Chung, "Tell Me What You Eat, and I Will Tell You Where You Come From: A Data Science Approach for Global Recipe Data on the Web," IEEE Access, vol. 4, pp. 8199-8211, 2016.
4. R. M. R. V. Kumar, M. A. Kumar and K. P. Soman, "Cuisine Prediction based on Ingredients using Tree Boosting Algorithms," Indian Journal of Science and Technology, vol. 9, no. 45, pp. 1-5, 2016.
5. D. M. BARRETT, J. C. BEAULIEU and R. SHEWFELT, "Color, flavor, texture, and nutritional quality of fresh-cut fruits and vegetables: desirable levels, instrumental and sensory measurement, and the effects of processing.," Critical reviews in food science and nutrition, vol. 50, no. 917969571, pp. 369-389, 2010.
6. B. Smith, "Perspective: Complexities of flavour," Nature, vol. 486, no. 7403, pp. S6-S6, 2012.
7. J. F. Delwiche, "You eat with your eyes first," Physiology and Behavior, vol. 107, no. 4, pp. 502-504, 2012.
8. G. K. Beauchamp and J. A. Mennella, "Flavor perception in human infants: Development and functional significance," Digestion, vol. 83, no. SUPPL, pp. 1-6, 2011.
9. B. C. Smith, "The Nature of Sensory Experience : the Case of Taste and Tasting," Phenomenology and Mind Online Journal, pp. 292-313 , 2013.
10. Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow and A.-L. Barabasi, "Flavor network and the principles of food pairing," Scientific Reports, vol. 1, p. 7, 2011.
11. A. Jaina, R. N. Kb and G. Baglerb, "Spices form the basis of food pairing in Indian cuisine," {arXiv preprint arXiv:1502.03815, 2015.
12. K. J. Hammond, "CHEF: A model of case-based

January-February 2020
ISSN: 0193-4120 Page No. 12731 - 12737

planning," Proceedings of the Fifth National Conference on Artificial Intelligence, vol. 1, p. 267–271, 1986.

13. J. Freyne and S. Berkovsky, "Intelligent Food Planning : Personalized Recipe Recommendation," Proceedings of the 15th international conference on Intelligent user interfaces, pp. 321-324, 2010.
14. P. Forbes and M. Zhu, "Content-boosted Matrix Factorization for Recommender Systems: Experiments with Recipe Recommendation," Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), p. 261, 2011.
15. 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, New York, NY, USA, 2012.
16. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.