Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

Influence of Food Habits in Gender Classification: A Support Vector based Approach

Akansha
Dept. of CSE, B.C.E.T, Durgapur, West Bengal, India

Prasenjit Kumar Patra
Dept of CSE, B.C.E.T, Durgapur, West Bengal, India

Nilanjan Dey
Dept. of CSE, Techno India, Kolkata, West Bengal, India

Abstract – People of different regions, age group as well as gender have different tastes and food habits. The objective of this project was to determine the differences between the food habits of male and female of a certain age group. The main goal was then to investigate patterns between male and female menu related choices and test whether the machine can classify the gender using the choices of menu or not. An online survey was conducted among the people, mainly in between the age group of 18-25 years in the region of West Bengal, mainly of Durgapur, by the use of Google Forms. This survey involved the entries from 544 males and 465 females. The results then obtained were classified using SVM Classifier. The final result was obtained with an accuracy of 89.99%, and the correlation between food habits between males and females were found, thus confirming that there is a vast difference in the food habits and menu choices of male and female.

Index Terms - Classification, SVM, Google Form.

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th& 26th September (2015) conducted by B. S. Anangpuria Instituteof Technology & Management, Village-Alampur, Ballabgarh-Sohna Road,Faridabad.

1. INTRODUCTION

Food indeed is basic necessity of life. Food is of paramount importance around the world- not just because of the fundamental need to consume it to sustain life but because of its cultural and social connotation. Food habit can be defined as the ways in which people select, cook, serve and eat foods that are available to them. Food habits differ from one ethnic group to another, from one country to another, one state to another state and also from person to person. Food habits are shaped by environment, available ingredients, climate, and even factors like class and income. The food habits of people vary greatly as we move from one place to another. Many cultures have their own recognizable cuisine which in simple terms is a specific set of cooking tradition that uses various kinds of spices or combination of flavours that are unique to that particular culture. Many cultures have diversified their foods by means of preparation, cooking methods, and manufacturing also.

The food habits of people across the globe vary widely. It is accepted that nutrition is most significant part of the environment that is introduced into the human body and the eating patterns are relevant component of cultural reference models [1]. For instance, Sushi is synonymous with Japan, fish and chips conjure up the images of Britain and tacos are associated with Mexico. While the traditional dish of Netherlands involves a lot of fish, some of which is eaten raw, the French diet is associated with lashings of rich food, cheese and wine and so on. Every country has their own variety of food which definitely is different from other countries of the world.

Considering the case of India, while travelling from North India to South India huge differences in the choices of food by different persons can be easily observed. The cuisines vary significantly from each other given the range of diversity in soil types, climate and occupations. The Indian cuisine changes across the geography of India as a result of variation in local culture, geographical location and economics. For instance, most of the people in coastal regions prefer eating fish and other sea foods. The food of Andhra Pradesh, located in the south part of India, is accused sometimes unfairly of using excessive amounts of spices and tamarind. In the coastal states of West Bengal and Kerala, the people consume a lot of

fish due to easy availability. All along the northern plain, from Punjab through Uttar Pradesh, to the eastern side of India, the main crop cultivated is wheat. People from these areas use wheat flour to make chapattis (Indian bread) and other closely related breads.

This was just considering the demography. Food habits vary according to age also. With varying age, there lies a variance in food intake and the choice of food by the individual. It has been observed that students and the youths have higher tendency to include junk foods in their diet due to the taste factors and easy availability. Some students may feel that their independence from parents equals the freedom to eat whatever they want, with little or no regard for possible consequences [2].The mature and the elderly people make healthier food choices. They prefer having food-at-home in contrast to the young aged adults who prefer to have food-away-from-home at restaurants and fast food places. The adolescent cohort showed a decrease in the intake of raw fruits, non-potato sources of vegetables, and dairy sources while there is an increase in high-fat potatoes and mixed dishes and particularly soft drink consumption.

Not only the region to where a person belongs shapes their liking towards a particular food item, it is also highly dependent on age and sex. It has been highlighted that nutrition could differently influence the health of male and female individual [3]. When it comes to choices of food between men and women, results are quite different on the basis of the gender of that person [4]. One theory says that the gender driven eating and access to different kinds of food can be explained as a result of evolution and differences in physiology [5]. Also according to some studies conducted in modern society, there are consistent associations between gender and specific foods, where meat, alcohol, and hearty portion sizes are associated with masculinity, while vegetables, fruit, fish and sour dairy products are associated with femininity [6]. Food choice is an area in which research has revealed consistent behavioural gender differences. In general, men mostly prefer eating meat and poultry items and women prefer to eat more of fruits and vegetables. There are a number of hypothesis on gender differences in diet. Girls are generally reported to have a more frequent consumption of healthy food items in contrast to boys who reported a higher consumption of unhealthy food items such as soft drinks and fast food. According to another explanation, females are more concerned about health considerations, have stronger beliefs in the importance of healthy foods, have a stronger desire to look after ones appearance and are more likely to translate their attitudes to action. Another explanation could be boys have higher energy requirements and thus control their food preferences toward more energy-efficient dense foods. In general, women have been frequently reported to engage in far more health-promoting behaviours than men and have healthier lifestyle patterns [7].

This paper basically focuses on the differences in food habits between male and female. Whatever be the reason, there is a vast difference between choices of food made by the two genders. It is possible to classify male and female easily on basis of choice they make after proper training of the machine.

2. METHOLODGY

Data is increasing with every passing day. Data, i.e., the raw facts and figures, are growing at a phenomenal rate and there is need of finding hidden information from the databases for the efficient use of the data. And here is where data mining comes into scene. Data Mining is a powerful technology with great potential that helps to find important information that is present in the databases but remains hidden to the common eyes.Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data, where the data is stored in databases, data warehouses, or any other information repositories [8].It is exploration and analysis of large data sets which helps in extraction of hidden predictive information from large database. In Data mining, large storages of data is searched in order to discover patterns and trends that cannot be discovered from simple analysis. Here, different mathematical algorithms are used to segment the data and evaluate the probability of future events. These tools can include statistical models, mathematical algorithms and machine learning methods. High performance data mining helps to explore full depth of any database. Data Mining automates the process of finding predictive information in large databases and this predictive information helps to track different behaviours of the data sets in an easy way.

A. Classification

Classification is one of the most important methodologies of Data Mining. It is a function in data mining that assigns items in a collection to some target classes. Basically, classification recognizes the pattern that describes the group to which an item belongs [9]. It aims to achieve the goal of accurately predicting the target class for each case in the data. This isdone by examining existing items that has already been classified and characterized into well-defined classes, forming training set which consists of reclassified examples which helps in attaining the goal of building up a model that can be applied to unclassified class for its classification. In other words, Classification is the process of sorting and categorizing data into various types, forms or any other

distinct class. It allows most effective and efficient use of the data. A well-planned data classification system makes essential data easy to find and retrieve. Different classification algorithms use different techniques for finding the relationships.

One important type of classifier is Support Vector Machine (SVM).

B. Support Vector Machine

The support vector machine (SVM) [10] [11] [12] is a supervised training model with algorithms which is used for learning classification from data. SVM can be used to learn polynomial, radial basis function (RBF) and multi-layer perceptron (MLP) classifiers. SVMs are currently among the best performers for a number of classification tasks. SVM has been used successfully in many real world problems like text categorization, image classification, bioinformatics, and hand written character recognition. Basically SVM is used when our data contains exactly two classes. SVM is based on the concept of decision planes that defines decision boundaries. A decision plane (hyperplane) can be defined as the one which is used to separate different set of objects that have different class memberships. In other words, SVM classifies data by finding the best hyperplane that will separate all data points of one class from those of the other class. Hence, the best hyperplane for an SVM means the one with the largest margin between the two classes. Here, Margin means the maximal width of the slab which is parallel to the hyperplane that will have no interior data points.

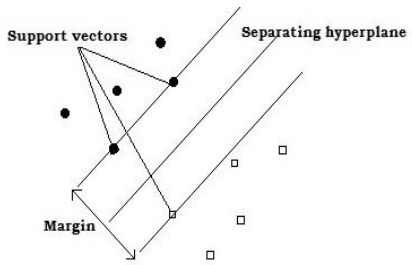

Figure 1. Figure showing Support Vectors and Hyperplane and margin

Here in Fig 1, the circles indicates data of Type 1 and squares indicates data of Type 2. Here the marked points on two parallel lines form the Support Vector. The separating hyperplane divides the two types of data in equal parts. The best hyperplane for an SVM is marked as Separating Hyperplane. The slab which is parallel to the Hyperplane that contains no interior data points, marked here as 'Margin', represents maximum width of two different class.

Support Vectors are the data points that lie closest to the decision surface and are most difficult to classify. These points have direct effect on the optimum location of the decision surface and are located on the boundary of the slab. In other words, support vectors are the elements of the training set that would change the position of the dividing hyperplane, if removed and hence are termed as the critical elements of the training set. SVM is used to find the vectors or we can say support vectors that define the separators which are used to give the widest separation of classes.

SVM provides the advantage of finding flexibility while choosing similarity functions and feature selection. It also has the ability to handle large feature spaces and provides sparseness of solution when dealing with large data sets. SVM also provides advantages of producing very accurate classifiers, less over fitting, robustness to noise. It is very effective in high dimensional spaces and used for the purpose of pattern recognition.

C. Google Forms

Google Form is a free tool provided by the GOOGLE which allows the creation of forms, surveys and quizzes. It provides the facility of editing the documents online while collaborating with other users in real time. This suite is tightly integrated with GOOGLE DRIVE. All files created with the apps and all the responses are by default saved to Google Drive. A form can be created from Google Drive or from an existing spreadsheet that can record responses to the form. Google Form allows selection from multiple question types, drag and drop to reorder questions and customizing values as easily as pasting a list. Google forms allows for the data to be submitted from anywhere. People from around the globe can complete and submit the form and share with others too. Forms can be shared via e-mail, link, or a website. It can be made to reach a broad category of audience by embedding it on website or sharing via various social networking sites like Google+, Facebook or Twitter etc. All the responses collected are stored in a spreadsheet which helps in correct analysing of data. The data or the results stored in Google Spreadsheets can also be exported for the purpose of analysing it with other software. Google form also makes it easy to control who is able to view and edit the forms and can also provide helpful summaries of the collected data with the help of various pie charts and line graphs.

3. RESULT AND DISCUSSION

The survey was conducted by the usage of Google Form. There was a complete online survey conducted in region of West Bengal, especially Durgapur. The form consisted of 15 questions spread over choices of different food items ranging

from daily vegetables like cabbage, cauliflower, potato and ladyfinger to sweets like Jalebi, KajuBarfi and MistiDahi,Drinks, choices between popcorn and nachos to chocolates and pastries and several other fast foods.

List of questions involved in the survey-

A. Preference for vegetable
B. Liking for Spinach
C. Choice for vegetarian food
D. Choice for non-vegetarian food
E. Choice between tea, coffee and green tea
F. Choice between chocolate and pastries
G. Choice for beverage
H. Choice for sweets
I. Choice for evening snacks
J. Choice between ice cream and soft drink
K. Choice between popcorn and nachos
L. Choice for favourite chips
M. Choice between milk chocolate and dark chocolate
N. Choice between rice and chapatti
O. Choice of preferred fast food

Approx. 1009 were collected and stored in Spreadsheet where they were assigned a corresponding value and the entire data was converted into Value Matrix.

After the conversion into value matrix, SVM classifier was applied on the value matrix to obtain the results.

SVM Classifier provides 6 types of Kernels, i.e., rbf kernel, linear kernel, quadratic kernel, polynomial (degree 1) kernel, polynomial (degree 2) kernel and polynomial (degree 3) kernel. On application of code on these kernel types we obtained values for:

- Sensitivity which is the measure of actual positives in the data sets which were correctly identified. The formula used for calculating Sensitivity is-

Sensitivity= TP/ (TP+FN)

Where TP stands for True Positive which is correctly identified data and FN stands for False Negative which is incorrectly rejected data from the dataset.

- Specificity is the measure of the proportion of negatives which were correctly identified. The formula for calculating Specificity is-

Specificity= TN/ (FP+TN)

Where TN stands for True Negative which is correctly rejected data and FP which stands for False Positive which is incorrectly identified data from the dataset.

- PPV which stands for Precision Predictive Value or Positive Predicted Value which defines the ratio of true positives (TP) to combined true and false positives. The formula used for calculating PPV is-

PPV= TP/ (TP+FP)

Where TP stands for True Positive which is correctly identified data and FP stands for False Positive which is incorrectly identified data from the dataset.

- Accuracy represents the accuracy of data involved and the number of correct guesses. It is the ratio of sum of true positive and true negative values to the sum of total positive and total negative. The formula used for calculating Accuracy is-

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

Where TP stands for True Positive which is correctly identified data, TN stands for True Negative which is correctly rejected data, FP stands for False Positive which is incorrectly identified data and FN stands for False Negative which is incorrectly rejected data from the dataset.

After proper calculations the following results were obtained for different types of kernels -

Table 1-
Table showing values obtained by using different SVM kernels for same data set.

Kernel/ Parameters	Mean Sensitivity	Mean Specificity	Mean PPV	Accuracy
RBF	69.8890	98.3535	97.4171	85.2287
Linear	90.9621	89.1347	87.9645	89.9911
Quadratic	88.6216	88.4175	86.9154	88.5030
Polynomial1	90.9621	89.1347	87.9645	89.9911
Polynomial2	88.6216	88.4175	88.9154	88.5030
Polynomial3	83.6633	78.1414	76.8671	80.6762

Now, as clear from the table, when we compare the values we find that the values obtained for linear and polynomial (Order 1) gives the highest accuracy. Hence for our dataset the Sensitivity is 90.96%, the Specificity is 89.13%, the PPV is 87.97% and the accuracy is 89.99%

The correlation between the choices of male and female for different questions in the survey was also calculated. The correlation was calculated considering a sample space of 450

for both male and female respectively. The values then obtained are-

Table 2-
Table showing correlation between data values

Question No	Correlation Coefficient (CC)
A	-0.07278
B	-0.01033
C	-0.03451
D	+0.04566
E	-0.06884
F	-0.06035
G	-0.01827
H	-0.04566
I	+0.04289
J	-0.06075
K	-0.00135
L	-0.08712
M	-0.09199
N	-0.07443
O	+0.01689

Now, as we know that the correlation coefficient value ranges from +1 to 0 where +1 represents highly correlated data and 0 represents highly uncorrelated data. Hence, from the table we can infer that the choice of non-vegetarian has a high correlation between that of male and female and the choice of chocolates, i.e., in between milk chocolate and dark chocolate has least correlation. In other words, the choices for non-vegetarian food are similar for male and female but the choices for chocolates are highly dissimilar.

Also, presence of negative sign before most of the values of correlation coefficient confirms that the choices of food between two different genders are different.

4. CONCLUSIONS

This survey on food habits variation of different gender is a noble work and no such work has been done earlier in knowledge of the authors. As we can see that the machine if properly trained can easily differentiate between male and female just on the basis of their choices of menu items. The future scope of this study can involve the survey results collected from different parts of world to check if the results are similar or not. Moreover, the success of this study can help in pre determining the choices of different people. This study can also be extended to classify between food habits of different regions of same country or of different countries. The breaches of this study are that food habits of individuals is highly dependent on age group to which the person belongs. It is also dependent upon the mood of the person at that particular point of time. This study was done using SVM Classifier. Other classifying methods can also be applied to obtain results and comparisons can be made for the results obtained from different methods using same data set.

ACKNOWLEDGMENT

In all humility and with much fervor, I owe my deep and sincere gratitude to Bengal college of Engineering, Durgapur, CSE department, for the enlightened guidance, continuous encouragement, estimated supervision and paternal affection throughout the period of this research. Key improvements in the proposed research work would not be possible without the valuable suggestion and feedback of my guides.

REFERENCES

[1] Harris, M. (1985). Good to Eat: Riddles of Food and Culture (Previously published as The Sacred Cow and the Abominable Pig). 2nd Edition 1998, Waveland Press Inc., Long Grove, Illinois USA
[2] Stockton S, Baker D. (2013). College students' perceptions of fast food restaurant menu items on Health. *American Journal of Health Education* 44(2), 74-80.
[3] Marino, M., Masella, R., Bulzomi P., Campesi, I., Malorni, W., &Franconi, F. (2011). Nutrition and human health from asex–gender perspective. Molecular Aspects of Medicine. Vol. 32, No. 1, pp. 1-70
[4] International Conference on Emerging Infectious Diseases in Atlanta, Georgia, March 19, 2008
[5] David Katz, Prevention Research Center, Yale University
[6] Jensen & Holm, 1999; Sobal, 2005
[7] Courtenay, 1998, 2000; Gough & Conner, 2006; Kandrack et al., 1991; Lonnquist et al., 1992; Roos et al., 2001
[8] Han, J., Kamber, M. (2000) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco
[9] Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C., eds. 1984. Classification and Regression Trees. CRC Press, Boca Raton, FL.
[10] B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Fifth Annual Workshop on Computational Learning Theory, pages 144–152. ACM, 1992.
[11] C. Cortes, V. Vapnik."Support Vector networks" Machine Learning 20(3):273. doi : 10.007/BF00994018
[12] Hsu,Chih-Wei; Chang, Chih-Chung; and Lin, Chih-Jen(2003). A Practical Guide to Support Vector Classification. Department of Computer Science and Information Engineering, National Taiwan University.

Authors

Akansha Sinhaacquired the B. Tech (2015) in Computer Science Engineering and currently working with TCS. Her work has spanned a seeming widely diverse set of topics, Data mining, Data classification.

Prasenjit kumar patra, is an Asst. Professor in the Department of computer science in Bengal College of Engineering and Technology, Durgapur, India,acquired the B. Tech (2009) and M. Tech (2013) Degree both in Computer Science Engineering. His work has spanned a seeming widely diverse set of topics, Data mining, scheduling techniques in distributed environments, Fault tolerance in distributed computing environments, availability for services and high-performance with the help of proper resource allocation, sensor networks.

NilanjanDey, PhD., is an Asst. Professor in the Department of Information Technology in Techno India College of Technology, Rajarhat, Kolkata, India. He holds an honorary position of Visiting Scientist at Global Biomedical Technologies Inc., CA, USA and Research Scientist of Laboratory of Applied Mathematical Modeling in Human Physiology, Territorial Organization Of-Sgientifig And Engineering Unions, BULGARIA. He is the Managing Editor of *International Journal of Image Mining (IJIM)*, Inderscience, Regional Editor-Asia of *International Journal of Intelligent Engineering Informatics (IJIEI)*, Inderscience and Associated Editor of *International Journal of Service Science, Management, Engineering, and Technology*, IGI Global. His research interests include: Medical Imaging, Soft computing, Data mining, Machine learning, Rough set, Mathematical Modeling and Computer Simulation, Modeling of Biomedical Systems, Robotics and Systems, Information Hiding, Security, Computer Aided Diagnosis, Atherosclerosis.