A SURVEY ON FOOD RECOMMANDATION SYSTEM USING DATA MINING CONCEPTS

Rimoni Patel, Mitesh Patel

Department of computer engineering ,silver oak engineering college

Abstract- **In this paper we are focusing on healthy diet recommendation concepts Good eating habits are important for maintaining a healthy life and preventing the lifestyle-related disease epidemic. Researches about menu recommendation or diet planning are thus attracting much attention recently. A key factor toward a successful diet planning is an individual's food preference instead of dogmatic nutrition pattern since it is unlikely that an individual would accept the meal plan merely based on the nutrition supplements. However, the extraction of personal preference is definitely not a trivial matter.**

Index Terms **– web data mining, recommdation system; personal health management , healthy eating ,decision algorithm**

I. INTRODUCTION

The measurement of population intakes of foods and nutrients is central to the science of human nutrition. At present, patterns of dietary intake are studied on a food-by-food basis, given that the base units for analysis using food composition databases are the individual food components of every meal. Whereas the use of individual foods for the study of dietary patterns has served us well, it remains possible that parallel analysis using the nutritional composition of meals might increase our ability to study diet-disease patterns. The concept of analyzing food combinations at the meal level is not entirely new. The examination of food combinations at the meal level provides an approach to deal with the complexity and unpredictability of the diet and aims to overcome the limitations of the study of nutrients and foods in isolation [1].

Data mining is a process that uses a variety of data analysis tools to discover patterns and relations in data that may be used to make predictions. Supervised data mining techniques are used to model an output variable based on one or more input variables, and these models can be used to predict or forecast future cases. The present article describes and compares 2 supervised methods, artificial neural networks (ANNs) and decision trees. In the past decade, the use of artificial intelligence has been explored in almost every field of medicine [2]. It has been suggested that physicians will not use any method that purports to improve diagnostic accuracy unless it is easy to use and dramatically and consistently improves their performance [3]. With regard to ANNs, success in the medical field has been widely demonstrated in the diagnosis and prediction of many illnesses (eg, coronary heart disease and cancer). For example, an ANN was used to estimate the risk of acute coronary death during 10-y follow-up in the Prospective Cardiovascular Munster Study. Overall, their analysis suggested that the use of the ANN might allow prevention of 25% of all coronary events in middle-aged men, compared with 15% with logistic regression. Decision tree algorithms have also been widely applied in the medical field. However, despite both of their widespread use in the medical field, thus far there have been no reports in the literature [3, 4].

Data mining [5] is the process of selecting, exploring and modeling large amounts of data. This process has become an increasingly pervasive activity in all areas of medical science re-search. Data mining has resulted in the discovery of useful hid-den patterns from massive databases. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and statistics.

The following six types of treatments were identified in the 2005 World Health Organization’s NCD report of Ministry of Health, Saudi and are discussed below:

(a) Drug

(b) Diet

(c) Weight reduction

(d) Smoke cessation

(e) Exercise

(f) Insulin

A. Drug: Oral medications, in the form of tablets help to control blood sugar levels in patients whose bodies still produce some insulin (the majority of people with type 2 diabetes). Drugs are usually prescribed to patients with diabetes (type 2) along with recommendations for making specific dietary changes and getting regular exercise. Several drugs are often used in combination to achieve optimal blood sugar control.

B. Diet: Patients with diabetes should maintain consistency in both food intake timings and the types of food they choose. Dietary consistency helps patients to prevent blood sugar levels from extreme highs and lows. Meal planning includes choosing nutritious foods and eating the right amount of food at the right time. Patients should consult regularly with their doctors and registered dieticians to learn how much fat, protein, and carbohydrates are needed. Meal plans should be selected to fit daily lifestyles and habits.

C. Weight reduction: One of the most important remedies for diabetes is weight reduction. Weight reduction increases the body's sensitivity to insulin and helps to control blood sugar levels.

D. Smoke cessation: Smoking is one of the causes for uncontrolled. Smoking doubles the damage that diabetes causes to the body by hardening the arteries. Smoking augments the risk of diabetes.

E. Exercise: Exercise is immensely important for managing diabetes. Combining diet, exercise, and drugs (when pre-scribed) will help to control weight and blood sugar levels. Exercise helps control diabetes by improving the body's use of insulin. Exercise also helps to burning excess body fat and control weight.

F. Insulin: Many people with diabetes must take insulin to manage their disease.

II. WEB DATA MININIG

The characteristics of web data determine the immense challenge for effective data mining. According to the characteristics of web data and combining the general process of data mining, the web data mining process can be described as the five functional modules, namely the data acquisition, data pre-processing, data mining, analysis and evaluation and knowledge formulation modules [6]. The functions of each module are shown as under.

A. Data acquisition

Generally speaking, in terms of functionality, data acquisition module selectively obtains data from the outside web environment to provide material and resources for the latter data mining. The data source that the web environment provided includes the web pages data, hyperlinks data and the history data of user visiting log. This module is composed by three relatively independent processes which are data search, data selection and data collection.

B. Data preprocessing

Data preprocessing mainly processes and reconstructs the source data acquired in data acquisition phase and builds the data warehouse of related themes to create basic platform for data mining process. Data preprocessing is preparation for data mining and it mainly includes data scrubbing, data integration, data conversion, data reduction, etc.

C. Data mining

Data mining module is the core of the whole system. Data mining is the process of extracting patterns from data, which is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. Generally speaking, the ultimate goals of data mining only are description and prediction, the so-called description is that using a comprehensible model to express the attributes and characteristics information contained in the data; and the prediction is to find the discipline of the attributes according to their existing data value and then speculate a possible attribute value in the future. Classical data mining techniques include classification of users, finding associations between different product items or customer behavior, and clustering of users [7].

1) Classification arranges the data into predefined groups. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification and neural networks.

3) Clustering is like classification but the groups are not predefined, so the algorithm will try to group similar items together. *Regression* attempts to find a function which models the data with the least error.

4) Association rule learning searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

D. Analysis and evaluation

Analysis and evaluation module is to analyze the credibility and effectiveness of the knowledge model the data mining obtained, and to reduce evaluated conclusions to provide information support for the management and decision-making of users.

E. Knowledge formulation

Knowledge expression module refers to the knowledge models mined from the web data by using data mining tools, and it will be shown with appropriate form to facilitate user acceptance and mutual exchange.

III. METHODOLOGY

Food consumption data

The data presented [8, 9] are derived from the North-South Ireland Food Consumption Survey (NSIFCS). The NSIFCS was a randomized cross-sectional study of food and nutrient intakes over 7 consecutive days in 1379 adults aged 18 – 64 y from the Re-public of Ireland and Northern Ireland. Each respondent was required to record their dietary intake for each eating occasion per day for the duration of the 7 d, and the definition and time of each eating occasion were also recorded. Respondent’s meal definitions were grouped into 1 of 6 categories: breakfast, light meal, main meal, snacks, alcoholic beverages, and nonalcoholic beverages. For the 1379 subjects in the database, a total 49 671 eating occasions were identified from the original 222,404 records. For ease of analysis, each of the unique food codes consumed by the respondents (3000 codes) was recorded into 1 of 62 food groups. These groups represent an aggregated classification system of similar food types consumed in the database.

A. Clustering Analysis

Clustering analysis [10] is a common technique for statistical data analysis that used in various fields including machine learning, pattern recognition, and data mining. Clustering is a method of unsupervised learning which groups similar objects on the basis of their attributes into a same group called cluster. The purpose of clustering is to group the objects based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. In other words, the greater the similarity within a group and the greater the difference between the groups, the better the clustering. Nowadays, there are many clustering algorithms which can be classified into the following categories; Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods, and Model-based methods. The techniques used in this study are the competitive learning called Self-Organizing Map and K-mean clustering which are described in the following parts.

B. Self-Organizing Map

The Self-Organizing Map (SOM) [11] or commonly known as Kohonen network is a type of artificial neural network that is trained using unsupervised learning for the visualization and analysis of high-dimensional data purpose. It was invented by a Finnish professor named Teuvo Kohonen. The SOM composed of map units called nodes or neurons which are connected to adjacent neurons by a neighborhood relation. In the two-dimensional map, the neurons can be arranged in geometrical shape such as rectangle or hexagon as shown in Fig.1. The SOM algorithm will compute a model, m_i, for each node. These models are the representation of the input space of the training samples and organized into an n-dimensional ordered map in which similar models are closer to each other in the grid than the more dissimilar one. In this concept, the SOM is a similarity graph, and a clustering diagram, too.

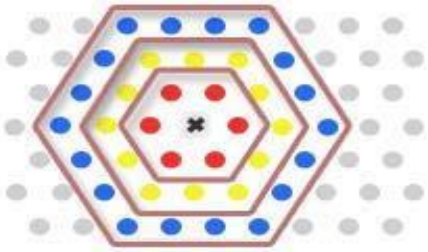

Figure 1. Neighborhood size 1, 2, and 3 of the centered node in a two-dimensional

hexagonal grid.

The difference between the Self-Organizing Map and other artificial neural networks is that SOM use a neighborhood function to preserve the topological properties of the input space.

The SOM training can be considered to be a competitive learning since when the input vector is presented to the network, the Euclidean distance to all nodes in the map is calculated to find the node that gives the smallest distance is called best matching unit (BMU). Then, the weights of the BMU and its neighbor are adjusted towards the input vector. This adjustment process stretches the prototypes of the BMU and its topological neighbors towards the input vector. The BMU's local neighborhood can be determined by using the neighborhood radius which will shrink with time.

C. K-mean clustering

K-means clustering [12] is one of the most well known and commonly used partitioning clustering methods. The k-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into non-overlapping *k* clusters where $k<n$. This method aims to minimize the sum of squared distance between an object to the centroid which is called sum of squared error. The K-means algorithm proceeds as follows. First is to randomly select the k centroids from a dataset. The centroid represents the mean value of all the objects in the cluster. The next step is to assign the remaining object to the nearest cluster based on the distance between the object and the cluster mean, which is the centroid and then calculates the new mean for each cluster. This process iterates until there is no change of the centroids values. In other words, until the criterion function is convergence [13].

D. Diabetes and Diet

Diabetes Mellitus is a chronic disease that occurs when the body does not produce enough insulin or the cells ignore the insulin. The effect of uncontrolled diabetes is hyperglycemia or high blood glucose which can lead to many serious diabetes complications especially cardiovascular diseases but it can be prevented and controlled by managing the nutrition intake which is called nutrition therapy. Among the different components of nutrition, carbohydrates have the greatest influence on blood glucose levels so people with diabetes need to essentially concern the total amounts of carbohydrates consumed in each day [14]. For dietary purposes, carbohydrates can be classified as simple and complex carbohydrate. Simple carbohydrates or simple sugars are typically high in glycemic index which is the measure of the effects of carbohydrates on the blood glucose levels so they will cause a rapid rise in the blood glucose levels. Simple carbohydrates can be found in refined sugars and some fruits as well. Complex carbohydrates which are also called starch are usually found in grain product such as rice, noodles, and bread and starchy vegetables. Diabetic patients are able to consume complex carbohydrates food but with limitation and should avoid food with simple carbohydrates.

A. Food Dataset

This study is based on the dataset "Nutritive values for Thai food" provided by Nutrition Division, Department of Health, Ministry of Public Health (Thailand). The dataset gives eighteen nutrient values of food in 100 grams edible portion including Energy, Water, Protein, Fat, Carbohydrate, Fiber, Ash, Calcium, Phosphorus, Iron, Retinol, Beta-carotene, Vitamin A, vitamin E, Thiamine, Riboflavin, Niacin, and Vitamin C. The total number of data we used in this study is 290 and most of them are Thai local mixed food dishes and one plate dishes.

We have categorized the dataset into groups by two different ways.

1) Categorized by food characteristic: The dataset was divided into 22 groups based on their characteristic and shortly named A to V. The 22 groups consist of noodles, round rice noodles, Thai curry, northeastern food, namprik, rice, fried food, condiments, snack, fruits, Thai dessert, jam, milk and yogurt, juice, tea and coffee, soft drink, alcoholic drinks, chocolate/cocoa, vegetarian, nuts and beans, sausages, and miscellaneous. The first letter in the FoodID represents the group of that item e.g. A001 indicates that this item belongs to group A (noodles) and numbered 001. In Fig.2 shows the dataset.

2) Categorized by nutrition for diabetes: We have asked the nutritionist to help dividing the dataset based on the nutrients for diabetes.

I. Normal Food (NF): Food in this group are very low in carbohydrate and have the least effect on the blood sugar level of the diabetic patients. Meat, poultry, meat fats, fish and seafood, vegetables excluding starchy vegetables, condiments with no sugar, and sugar-free herbal drinks are in this group.

II. *Limited Food (LF):* Foods in this group have large amount of carbohydrate but mostly are complex carbohydrates. Diabetic patients should carefully control the food intake in this group. Starchy food like rice, noodles, bread, and also starchy vegetables and grain are placed in this group.

III. *Avoidable Food (AF)*: Foods in this group also have large amount of carbohydrate especially simple carbohydrate or simple sugar which is considered to primarily affect the blood sugar level. Diabetic patients should avoid food in the avoidable group e.g. sweets and chocolates, sweet fruits, and drinks that contain sugar.

IV. DECISION TREE LEARNING ALGORITHM

'Decision tree learning [15, 16] is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Decision tree learning is one of the most widely used and practical. Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules.

- ID3 Basic

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric---information gain.

To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

- Information Gain --- measuring the expected reduction in Entropy

To minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice.

We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node.

V. CONCLUSION

Survey contain study of k-mean , clustering analysis and self-organizing map as a methodology after that we concentrate on three parameters normal food, limited food and avoidable food . based on study of various classification algorithm decision tree learning ID3 algorithm is analyses now. Still for accurate food recommendation it is require to have less and more accurate parameter and also based on complexity and prediction we can select other classification algorithm. Also the valid dataset plays a vital role in food recommendation system.

REFERENCES

[1] Aine P Hearty and Michael J Gibney ,'Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees' Am J Clin Nutr 2008;88:1632– 42. Printed in USA. © 2008 American Society for Nutrition

[2] Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 2002;15:11–39.

[3] Baxt WG, Shofer FS, Sites FD, Hollander JE. A neural computational aid to the diagnosis of acute myocardial infarction. Ann Emerg Med 2002; 39:366 –73

[4]A Data Mining Framework for Building A Web-Page Recommender System Choochart Haruechaiyasak , Klong Luang, Pathumthani, Mei - Ling Shy, Shu-Ching Chen

[5] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui ,' Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University – Computer and Information Sciences (2013) 25, 127–136

[6] Sarwar, B., Karypis, G., Konstan, J.A., & Reidl, J., Item-based Collaborative Filtering Recommendation Algorithms, Proceedings of the Tenth International Conference on World Wide Web, 2001, pp. 285 - 295.

[7] J. Han and M. Kamber, Data Mining: Concepts

and Techniques, Morgan Kaurmann Publishers, 2003.

[8] Harrington KE, Robson PJ, Kiely M, Livingstone MB, Lambe J, Gibney MJ. The North/South Ireland Food Consumption Survey: survey design and methodology. Public Health Nutr 2001;4:1037– 42.

[9] Kiely M, Flynn A, Harrington KE, Robson PJ, Cran G. Sampling de-scription and procedures used to conduct the North/South Ireland Food Consumption Survey. Public Health Nutr 2001;4:1029 –35

[10] L. Chang-Shing , W. Mei-Hui, L. Huan-Chung and C. Wen-Hui, "Intelligent Ontological Agent for Diabetic Food Recommendation", IEEE Xplore., pp. 1803-1810, 2008

[11] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, pp. 383-403, 2006.

[12] T. Kohonen, "The self-organizing maps," Proc. IEEE, vol. 30, no. 9, pp. 1464-1480, 1990.

[13] J. B. MacQueen : "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, vol. 1, pp.281-297, 1967.

[14] K. Vipin, An Extensive Survey of Clustering Methods for Data Mining. http://www-users.cs.umn.edu/~han/dmclass/

[15] Wei Peng, Juhua Chen and Haiping Zhou ‘ An Implementation of ID3 --- Decision Tree Learning Algorithm’

[16] Paul E. Utgoff and Carla E. Brodley, (1990). 'An Incremental Method for Finding Multivariate Splits for Decision Trees', Machine Learning: Proceedings of the Seventh International Conference, (pp.58). Palo Alto.