A Comparison of Decision Tree Algorithms on Healthy Eating System

Kamble Rita R.
Department of CSE,
RSCOE, Tathawde,
Pune, India.

ABSTRACT

Now daysmany people suffers from various diseases due to poor eating habits. In the current scenario fast food become important food in daily routine because it is effortlessly available but taking fast food in routine may cause for disease like heart attack, diabetics etc. Healthier diets help us to maintain our health and keep us away from many diseases.A healthy diet may improve or maintain optimal health. In developed countries, affluence enables unconstrained caloric intake and possibly inappropriate food choices. Medical study has revealed that people set a bigger possibility of warding off illness by consumption of healthy foods and by increasing their resistant system. For better recovery from diseases or surgery etc individual have special needs according to their medical profile, cultural backgrounds and nutrient requirements. Design and implementation of healthy diet recommendation system is based on web data mining which is the application of data mining technique help us to determine pattern from web. In terms of accuracy and time performance analysis we are using two decision tree learning algorithm ID3 and C4.5 and apply it on healthy diet application.In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains.

Keywords

Decision tree, bagging, healthy diet, ID3, C4.5

1. INTRODUCTION

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The concept of *Data Mining* is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business *Data Mining* (e.g., Classification Trees). In a machine learning process the classification can be described as a supervised learning algorithm. Data records are belong to class on the bases of Knowledge of class it assign a class labels to data to co-design and co develop software and hardware, and hence, such components. However, incorporation of that deal with knowledge extraction from database records and prediction of class label from unknown Data set of records.

We can define classification is a development in which specified set of data records is Separated into training and test data sets. For validating the model we required the test data record and for constructing the classification model training data set is required. The constructed classification model is used for classifying and predicting new data set records. These new data set records are different from training and test data set. For getting higher classification accuracy or accurate prediction we required a prior knowledge of the class label data record which makes attribute selection effortless. For higher classification accuracy supervised learning algorithm (like classification) is preferred to unsupervised learning algorithm (like clustering). In current scenario, data mining technology has been widely used in education, real estate, stocks, health care and other fields. A number of widespread classification algorithms used in data mining and decision support systems is: neural networks, logistic regression,Decision trees etc. Among these classification algorithms decision tree algorithms is the most frequently used because of it is effortless to understand.

2. DATA MINING TECHNIQUES

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. In the data collection and preprocessing web server data base contains two types of data base one is content data base that contain the information like user information and other types of data and second is the server log data base for recording the HTTP transaction log records).Data collection or data acquisition module collect data from the external web atmosphere to provide resources and material for the latter data mining.From the web environment the data source we get the web pages data, hyperlinks data and history data of user visiting log. Data collection module composed by three independent processes that are data collection, data selection, data search.

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.Data preprocessing is preparation for data mining and it mainly includes data scrubbing, data integration, data conversion, data reduction, etc. Basically in the data preprocessing step convert the data into the form which is accepted by the data mining algorithm.

INFORMATION FILTERING:An Information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation to a human user. Information filtering is the main step of the recommendation system. In the existing association rules are applied in the content base filter. In the performance analysis of healthy diet recommendation system, introduced a new architecture based on data mining algorithm for constructing a healthy diet recommender system. A healthy diet recommender system is an intermediary program (or an agent) with a user interface that automatically and intelligently extracts the useful information of people‘s eating habit which suits an individual‘s needs.

CONTENT BASED FILTER: The content-based filtering (CBF) is a consequence and persistence of information filtering research. It constructs the recommendation based on the correlation between difference resources. In content-based recommendation systems, resources are described as a vector of attributes. The system then learns a profile of the user‘s interests based on the features presented in the objects the user has rated. When making a prediction on the customers‘ preferences, the system analyzes the relationship between the products rated by the users and other products by calculating the similarity between their attribute vectors. In our healthy eating recommendation system the healthy eating dataset first apply to the content base filter it analysis the user behavior or the content of dataset for example the whether the user is vegetarian or suffering from some kind of diseases. The content base filters analysis the user profile. For classifying data we apply the decision rule mining on user access pattern. We apply the ID3 algorithm for classify the data. Decision rule mining construct the rule that is apply on user access pattern and generate the result. The output of the content base filter is the food that is beneficial for your health. For improving the accuracy of the system we apply bagging.

BAGGING: Bootstrap aggregation, or shortly said bagging, is an ensemble meta-learning technique that includes training many classifiers on different partitions of the training data and using the majority vote on the results of all those classifiers to define the final answer for a test pattern. This technique was proposed by Breiman in 1994 and can be used with many classification methods. The final effect of this technique is to reduce the variance associated with prediction, and thereby improve the prediction process.

The steps are:

1. X bootstrap samples are drawn from the available training data,
2. A classification model is trained on each bootstrap sample,
3. All classification models are run on the test set
4. For each test pattern, the results are combined by simple voting: the class with the majority of results across all classifiers is chosen as the final class.

In healthy diet recommendation system framework, database using the Relational Database management System (RDBMS) is designed and constructed. This database stores the URLs (i.e., Web pages), keywords for the Web pages, the recommended set of rules from content -based filtering, user login information, and user profiles. MySQL provides a multi-threaded, multi-user, and robust SQL (Structured Query Language) database management system, which is suitable for the application of recommender systems.

3. DECISION TREE

We defined decision tree is a tree in which each branch node symbolize a preference between a number of substitute, and each leaf node correspond to a decision. Decision tree are generally used for gaining information for the reason of decision -making. Recommendation systems are used to predict the desire value. By applying the data mining algorithm on data set in recommendation system predict the data according to the user preference. Prediction can be categorized into: classification, density estimation and regression. In classification, the predicted variable is a binary or categorical variable. various well-liked decision tree classification methods include decision trees, logistic regression and support vector machines. It starts with a root node on which it is for users to acquire actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. There are various decision tree classification algorithm are used like ID3, C4.5, C5.0 etc we work on ID3 and C4.5 the basic decision tree learning algorithm used for classify data.

ID3 builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. The example has several attributes and belongs to a class (like yes or no). The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, elicits knowledge from an expert. J. Ross Quinlan originally developed ID3 at the University of Sydney. He first presented ID3 in 1975 in a book, *Machine Learning*, vol. 1, no. 1. ID3 is based off the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C:

Step 1: If all instances in C are positive, then create YES node and halt.
If all instances in C are negative, create a NO node and halt.
Otherwise select a feature, F with values v1, ...,vn and create a decision node.

Step 2: Partition the training instances in C into subsets C1, C2, ..., Cn according to the values of V.

Step 3: apply the algorithm recursively to each of the sets Ci.

ID3 is a nonincremental algorithm, meaning it derives its classes from a fixed set of training instances. An incremental algorithm revises the current concept definition, if necessary, with a new sample. The classes created by ID3 are inductive, that is, given a small set of training instances, the specific classes created by ID3 are expected to work for all future instances. The distribution of the unknowns must be the same as the test cases. Induction classes cannot be proven to work in every case since they may classify an infinite number of instances. Note that ID3 (or any inductive algorithm) may misclassify data.

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan.[1] C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

This algorithm has a few base cases.

1. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
2. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
3. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

C4.5 made a number of improvements to ID3. Some of these are:

1. Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
2. Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ?for missing. Missing attribute values are simply not used in gain and entropy calculations.
3. Handling attributes with differing costs.

Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

APPLY DECISION TREE RULE MINING ON SYSTEM:The performance of healthy diet system used the ID3 and C4.5 decision tree classification algorithm for classify the healthy diet data set. First the content base filters analysis the user access pattern. Content base filter analyzed the user profile whether the user vegetarian or non vegetarian, suffering from some kind of diseases etc are analyzed. Then according to the user profile healthy diet data set is classified by the decision rule mining. It trains the data set and generate rule according to the user access pattern. In recommendation system we use the ID3 decision rule mining for mining the data and generate rule. These rules are applied on healthy diet data set and suggest food which is beneficial for your health. For performance analysis we calculate the accuracy of the system with ID3 and then compare the accuracy of ID3 with C4.5. For improving the performance of the system we apply bagging with ID3.

4. RESULT ANALYSIS

In the performance analysis of diet system decision tree first get the data from content base filter. In the implementation phase we first select the data set then the generated rule. Then these rules are applied into the data set. After applying the rule admin selects the profile where we want to apply rule. Oncethe profile selected the rules are applied and according to the user profile the food is suggested. Then we apply the rules on and analysis the system. The result analysis shows that ID3 works in each instance of data it‘s also work properly when the number of instance are increase. As compare to C4.5 and ID3 with bagging provide more accurate result Classification accuracy is higher. C4.5 construct tree in less time as compare to ID3 but will not work on each instance. ID3 work on each instance and gives more accurate result after applying bagging accuracy is increased.First the recommendation system suggests the food that is beneficial for your health then show the comparative analysis of two decision tree classification algorithms in terms of accuracy. For improving the performances of the system bagging is applied. The comparative study of the system shows that after applying bagging it gives more accurate result.

5. CONCLUSION

In this paper we conclude that result analysis of healthy diet recommendation system recommends the food that is beneficial for your health. We acquire people eating habit data in the database which could track people‘s recipe record. Then we introduce a web data mining solution to e-commerce to discover hidden patterns and business strategies from their customer and web data, implement a new framework based on data mining technology and build healthy eating recommendation system. Finally we give out personalized recommendations for each person. The system contains two domains Administrator and member. At the administrator‘s end we design a model that helps to get rules for the large data set for foodtest it over random values. System applies these rules over the all user‘s profile to get suggestion for healthy food. User manage their account and medical profilefinally get suggested with properdiet.If member update their health and medical profile the diet suggestion also update according to attributes to improve the health for all users by suggesting proper diet.

REFERENCES

[1] A Comparison of Decision Tree Ensemble CreationTechniques Robert E. Banfield and Lawrence O. Hall, levin W. Bowyer, W. Philip Kegelmeyer.

[2] Re Optimization of ID3 and C4.5 Decision Tree Devashsish Thakur, NisargaMarkandaiah, Sharan Raj Int‘l Conf. on Computer & Communication Technology(IEEE2010).

[3] An Implementation of ID3 Decision Tree Learning Algorithm‖ Wei Peng, Juhua Chen and Haipingzhou[1997].

[4] Xiaocheng Li, Xinliu, Zengjie Zhang, Yongming Xia, Songrong Qian Design of Healthy Eating System based on Web Data Mining 2010 WASE International Conference on Information Engineering.