European Scientific Journal August 2015 edition vol.11, No.24 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

EVALUATING RISK FACTORS OF BEING OBESE, BY USING ID3 ALGORITHM IN WEKA SOFTWARE

Msc. Daniela Qendraj (Halidini)
Msc. Evgjeni Xhafaj
Department of Mathematics, Faculty of Information Technology,
University "Aleksandër Moisiu" Durrës, Durrës, Albania

Abstract

Id3 algorithm is used for building a decision tree from a fixed set of examples, next step is an iterative way that uses the resulting tree to classify future examples. Nowdays the large amount of data needs to be classified into useful information. Being obese refers to an excessive accumulation of body fat. The aim of this paper is to construct a decision tree with Id3 algorithm, by the data collect from Tirana Inter-medical Centre, analyzing the factors that makes the patients obese.

Keywords: Id3 algorithm, decision tree, information gain, Weka

Introduction

By the database of Tirana Inter-medical Centre, the number of obese patients is growing up, during the last 3 years. Obesity is a complex disease, present to children, young people, and older, which has many causes, but the main factors are: genetic, metabolic, physical activity, blood pressure etc. We have run Id3 algorithm in order to study which is the most important factor that influence to the obesity in population.

Id3 algorithm or Iterative Dichotometer 3, has been introduced by Ross Quinlan 1986, as an algorithm which produce reasonable trees. Id3 is an algorithm that constructs a decision tree from a fixed set of data, usually discrete attributes. The leaf node of the decision tree contains the class name, but a non leaf node is a decision node. The decision node is an attribute test with each branch, being a possible value to the attribute.Id3 uses information gain to decide which of the attribute goes into a decision node.

Ross Quinlan [1] has worked on this kinds of decision trees, which look simple and technically easily to use. If we have a sub-tree that leads to a unique solution, than all sub-tree may reduce to the simple conclusion. Quinlan has improved that this kind of tree does not change the final result.

Data

With the recorded data of Tirana Intermedical Centre, for us as researchers is easier to get access to the large data, to produce a tree that is useful for clinical diagnosis. We have collected some data from the patients as: exercise, smoker, fruits. The data set is the result of a study done in this clinic, for each of 100 patients. We will show the decision tree only for twenty of them.

Methods

The Id3 algorithm is implemented to handle discrete and continuous value. This algorithm uses the Shannon entropy and a statistical property called Information Gain. In thermodynamics, entropy measures how ordered or disordered is a system. In information theory, entropy is a measure of how certain or uncertain is the value of the random variable.

This entropy was introduced by Claude Shannon in 1948, which quantifies randomness , lower values implies less uncertainty , high value implies more uncertainty. If we have a set S of n attributes, that have different values, than the entropy is defined as :

$$Entropy(S) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

where p_i is the proportion of S belonging to the class i.

Information Gain measures effective change in entropy, after making a decision based on a value of an attribute. In the context of building a decision tree, we are interested in how much information about the output attribute can be gained by knowing the value of an attribute X.

$$Gain(S, A) = Entropy(S) - \sum_{j=1}^{n} \left[p_j \cdot Entropy\ (p_j) \right]$$

where p_j is the set of all possible values for the attribute A.

Gain (S, A) is used for ranking attributes and building the decision tree where at each node is located the attribute that has the highest information gain, compared to the other attributes that are not considered in the path from the root. Considering the data in Table 1, that has twenty data points of randomly persons with attributes named Exercise, Smoker, Fruits, Obese.

After calculating by hand the best attribute, we will run Weka Tushar Version 3.7.12 that generates the decision tree with Id3 algorithm automatically. We will visualize the tree by the software implemented.

Table 1

Persons	Exercise	Smoker	Fruits	Obese
Per1	No	Yes	No	Yes
Per6	Yes	Yes	No	No
Per11	No	Yes	No	Yes
Per17	Yes	Yes	No	Yes
Per20	No	No	Yes	Yes
Per23	No	Yes	Yes	No
Per27	No	Yes	No	Yes
Per36	Yes	Yes	No	Yes
Per40	Yes	Yes	No	Yes
Per48	No	Yes	Yes	Yes
Per51	No	No	Yes	Yes
Per59	Yes	No	No	No
Per62	Yes	Yes	No	Yes
Per65	No	Yes	Yes	No
Per74	No	Yes	Yes	No
Per85	Yes	Yes	No	Yes
Per88	Yes	No	No	No
Per90	Yes	Yes	Yes	Yes
Per97	No	No	Yes	Yes
Per100	No	No	No	Yes

constructing the decision tree, we have to find the root node of the tree, which is one of our attributes. Now let calculate the Entropy and Information Gain for each attribute, and which of them have the highest gain, that will be the root node. After that this attribute will be removed from the set, and the data set will be split on the values of this attribute. This theory is used recursively for each of the subtrees.

$$Entropy\ (S) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i = -\frac{14}{20}\log_2\frac{14}{20} - \frac{6}{20}\log_2\frac{6}{20} = 0.86$$

For the first attribute named Exercise, we calculate the information gain:

$$Entropy(S_{Yes}) = -\frac{6}{9}\log_2\frac{6}{9} - \frac{3}{9}\log_2\frac{3}{9} = 0.933$$

$$Entropy(S_{No}) = -\frac{3}{11}\log_2\frac{3}{11} - \frac{8}{11}\log_2\frac{8}{11} = 0.863$$

$$Gain(S, Exercise) =$$
$$Entropy(S) - \frac{9}{20}Entropy(S_{Yes}) - \frac{11}{20}Entropy(S_{No}) =$$
$$= 0.86 - 0.45 \cdot 0.933 - 0.55 \cdot 0.863 = 0.148$$

For the second attribute named, Smoker we calculate the information gain:

$$Entropy(S_{Yes}) = -\frac{10}{14}\log_2\frac{10}{14} - \frac{4}{14}\log_2\frac{4}{14} = 0.84$$

$$Entropy(S_{No}) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.9$$

$$Information\ Gain(S, Smoker) = \\ = Entropy(S) - \frac{14}{20}Entropy(S_{Yes}) - \frac{6}{20}Entropy(S_{No}) \\ = 0.876$$

For the last attribute named Fruits, we calculate the information gain:

$$Entropy(S_{Yes}) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.9$$

$$Entropy(S_{No}) = -\frac{3}{12}\log_2\frac{3}{12} - \frac{9}{12}\log_2\frac{9}{12} = 0.8$$

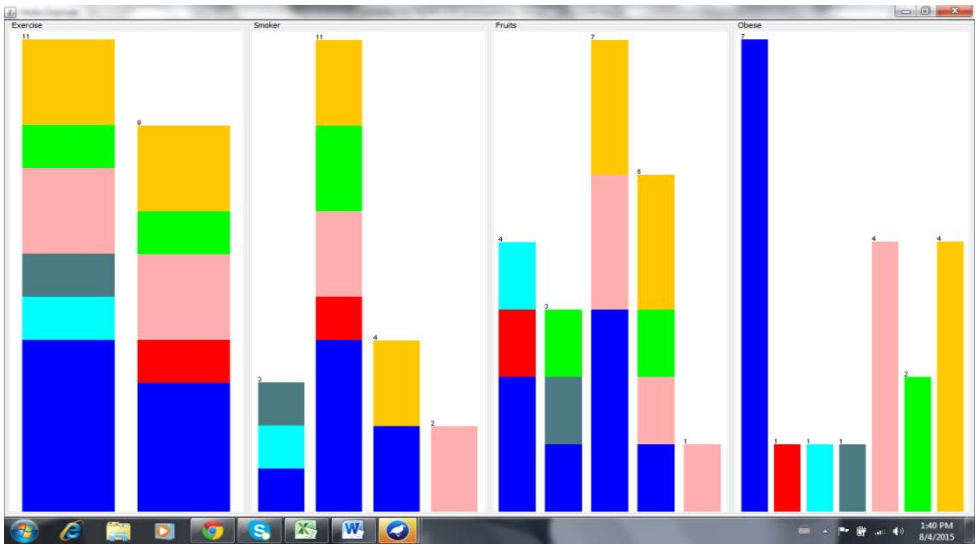
$$Gain(S, Fruits) = Entropy(S) - \frac{8}{20}Entropy(S_{Yes}) - \frac{12}{20}Entropy(S_{No}) \\ = 0.799$$

Smoker is the attribute that has the highest information gain so it is used as the root node of the tree.

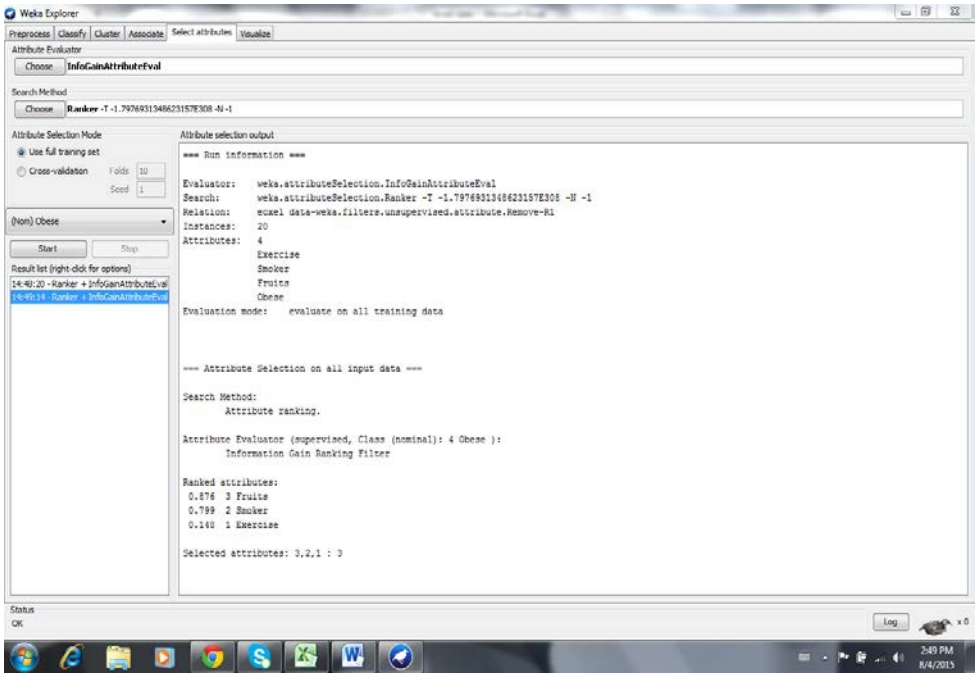
Run Weka 3.7.12.Tushar

Commands : Open file (File has been saved in type csv. file format)
Select Classify, trees, j48 pruned
Use training set, start. Visualize tree.

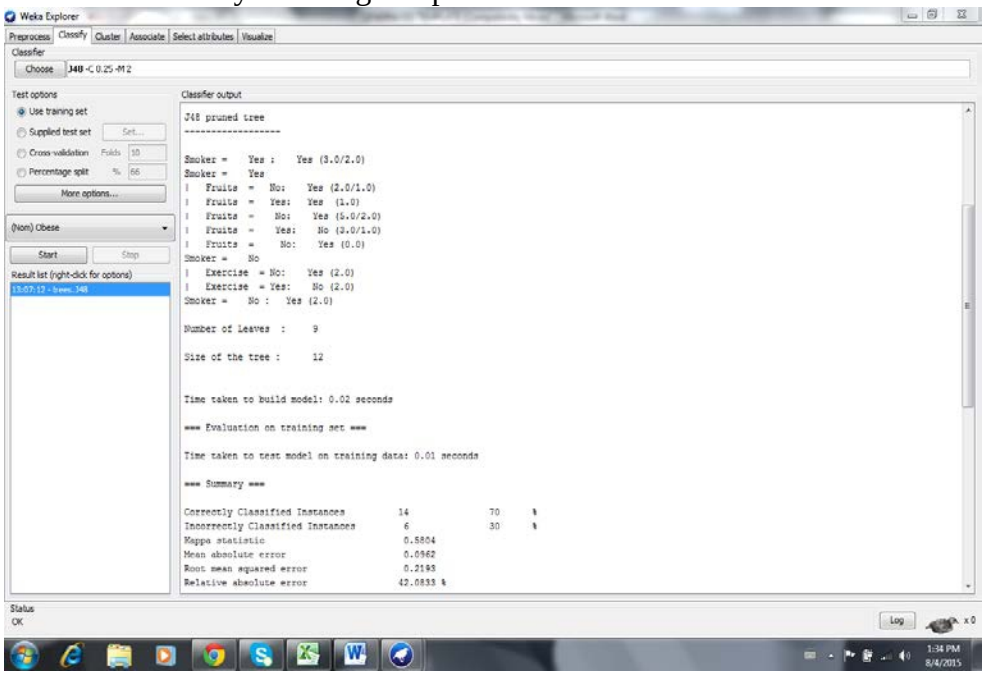
Here are the attributes clusters.



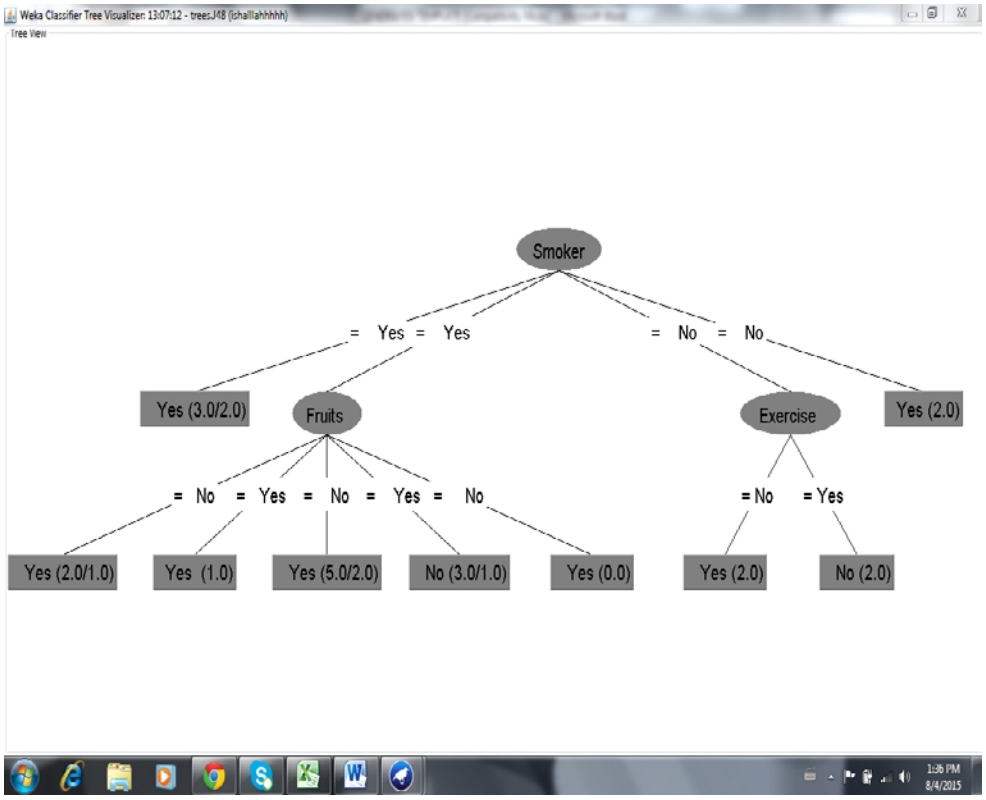
Here we have run Information Gain, for ranking attributes



Then we run Id3 by selecting J48 pruned of trees



Here is the tree visualized.



During the construction of the tree, some data that have a missing attribute value are eliminated from the next iteration, so if a data point has missing values for a given attribute that entry is removed from the construction of the subtree.

Conclusion

This application of Id3 algorithm using Weka software is built for ranking the attributes showed in Table 1, as so on if we have a new case and we want to classify it that is Obese or no, we will follow the path from the root node to a leaf node using the point's attribute values. Potential areas as economy or meteorology, where data are Boolean , are the best cases for using Weka.

References:

J.R.Quinlan "Induction of decision trees",Machine Learning ,vol 1, pp 81-100,mar 1986.

Schmidt.R.Montani, S.Bellazi, R.Portinale, Case based reasoning for Medical Knowledge-Based systems, International Journal of Medicine Info 2001.

Q.Zhang et al, "Application of ID3 Algorithm in Exercise Prescription", in Proccedings of the International Conference on Electric and Electronics, 2011 pp 669-675

Mark Hall et al, " The Weka Data Mining Software". Sigkdd Explorations, vol 11,issue 1, 2009.

Sunita Soni, Jyothi Pillai " An expert case-based system using decision tree Induction for Weight Management Conseling to Obese Children", International Journal of Computer science and Applications, vol1, nr2, august 2008.

T.Nelson, Java Universal Network/ Graph Framework [Online] http://jung.sourceforge.net/index.html