International Journal of Advanced Smart Convergence Vol.6 No.3 22-28 (2017)
https://doi.org/10.7236/IJASC.2017.6.3.22

IJASC 17-3-4

Correlation Analysis of the Frequency and Death Rates in Arterial Intervention using C4.5

Yong Gyu Jung [1]*, Sung-Jun Jung [2], Byeong Heon Cha[3]

[1]*Dept. of Medical IT, Eulji University, Korea*
[2]*Dept. of Electronic Engineering, Sogang University, Korea*
[3]*Dept. of Biomedical Laboratory Science, Eulji University, Korea*
E-mail: ygjung@eulji.ac.kr, sean121@naver.com, jobogy@eulji.ac.kr

Abstract

With the recent development of technologies to manage vast amounts of data, data mining technology has had a major impact on all industries.. Data mining is the process of discovering useful correlations hidden in data, extracting executable information for the future, and using it for decision making. In other words, it is a core process of Knowledge Discovery in data base(KDD) that transforms input data and derives useful information. It extracts information that we did not know until now from a large data base.
In the decision tree, c4.5 algorithm was used. In addition, the C4.5 algorithm was used in the decision tree to analyze the difference between frequency and mortality in the region. In this paper, the frequency and mortality of percutaneous coronary intervention for patients with heart disease were divided into regions.

Key words*: Decision Tree, C4.5 Algorithm, Percutaneous Coronary Intervention, Heart Disease, Mortality*

1. Introduction

As technology evolves, existing data is not just numerical data, but it can predict other information and derive new information accordingly. Data mining is the modeling of information by analyzing the relationship and characteristics of each property based on the vast amount of data. Data mining can be described as an automated, semi-automated technique for exploring and modeling useful relationships that were previously unknown in the vast amounts of data through processes such as data refinement and summarization, pattern and rule discovery and extraction. Examples of massive data are databases, billing information, purchase information, and personal information that we produce and collect to do things such as decision support, competitive advantage, and forecasting of future information. Data mining is used in various fields such as military, information, security, and medicine.

In this paper, we combine data mining with medical field. Recently, data mining has been widely used to derive new medical facts about gene information and various symptoms of various organisms. Human medical information is the most valuable one among all biological data. In this paper, Decision trees were applied to enhance the accuracy of predicted values. C4.5 algorithm used in decision trees, were compared and evaluated

Manuscript Received: 20 June, 2017 / Revised: 5 July, 2017 / Accepted: 17 July, 2017
Corresponding Author: ygjung@eulji.ac.kr
el:+82-31-740-7235, Fax: +82-31-740-7190
Department of Medical IT, Eulji University, South Korea

2. Related Research

2.1 Decision trees

Decision trees are not efficient systems in collecting, storing, and distributing information, but algorithms for supporting decision making and improving the effectiveness of decision making. That is, it is a tool for choosing wise alternatives. But it is not an algorithm that can replace the decision-making system, but an algorithm that helps decision-making.

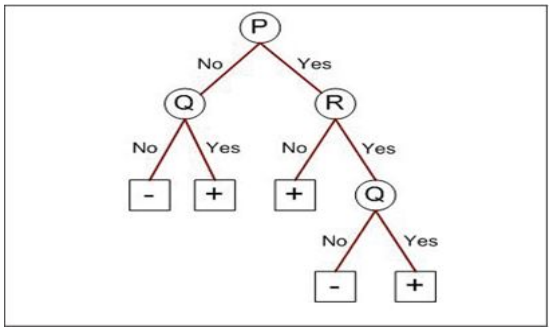

Figure 1. Example of Decision Trees

As shown in Figure 1, the nodes in the decision tree involve verifying specific attributes. In general, the verification scheme consists of comparing one attribute value with an arbitrary constant. In the beginning, the root node, P, is divided into yes and no. If no, Q is divided into no and yes, and execution stops at the last node. And then classified according to the class assigned to the node. Decision trees are graphical representations of procedures for classifying or evaluating interesting items. For example, given a patient's symptoms, it can be used to determine a likely diagnosis and recommend a treatment method. In other words, a decision tree can be used to obtain a diagnosis of hepatitis in the stored data and patient symptoms, and to predict the likelihood of death.

It expresses the function by mapping the element of the range to the element of the area with the letter or numerical value representing the class. In the inner node of the tree, find one test that produces a small number of possible outputs. Depending on the results of each test, you will reach a leaf that contains the same class character or numeric value as the item you want to know.

Each leaf shows the number of instances of the class falling into that leaf. These leaves are not usually one class. Therefore, the most common class character is selected.

Rule Set

r1: (P=No,Q=No) ==> -
r2: (P=No,Q=Yes) ==> +
r3: (P=Yes,R=No) ==> +
r4: (P=Yes,R=Yes,Q=No) ==> -
r5: (P=Yes,R=Yes,Q=Yes) ==> +

Figure 2. Decision tree represented by rules

The pruning procedure uses an estimate of the expected error for the verification data per node. First, the mean value of the absolute values of the difference between the prediction value and the actual class value is obtained for each of the other training instances in the corresponding node. Because the tree is explicitly constructed for this data set, this average value will underestimate the expected error for unseen cases. To compensate for this, the average value is multiplied by (n+v)(n-v). Where n is the number of different training instances at the node and v is the number of parameters present in the linear model providing the class value for that node.

2.2 C4.5

C4.5 is an innovative decision tree, a machine learning algorithm that is used in almost every practical mining field today. C4.5 uses the concept of entropy according to information theory to construct a simple decision tree that can classify a given pattern correctly. C4.5 correctly classifies a given pattern and uses the concept of entropy according to information theory to construct a simple decision tree. Chaotic diagram quantitatively shows the degree of mixing of different kinds of patterns. The more random data of different classes are mixed in a subtree corresponding to a certain node (or path) of a decision tree, the higher the degree of disorder. On the contrary, if the data is a single class in a sub tree corresponding to a certain node, the degree of disorder is low.

The C4.5 algorithm constructs a simple decision tree by expanding the tree from the root node by choosing the property that can minimize the disorder. In a tree structure, a colon is assigned a class label that is assigned to a specific leaf node, followed by the number of different instances of that leaf node. The algorithm is expressed in decimal because of the way it uses instances of fractions to handle missing attribute values. The algorithm is expressed in decimal because of the way it uses instances of fractions to handle missing attribute values. If there are inappropriate instances, the number of instances is also displayed, and the accuracy is displayed.

Classification and Regression Trees (CART), which is one of the decision tree construction algorithms, forms a binary tree structure by forming a binary partition at each node, while C4.5 is composed of trees each having a structure of dodge separation.

A typical CART (Classification and Regression Trees) of the configuration algorithm of the decision tree creates a separation tree structure that does not form a binary division at each node, whereas C 4. 5 creates a separation tree structure of the tree. For continuous variables, variance is used, but for categorical variables CART uses the Gini index as a divisor and C4.5 as an entropy index.

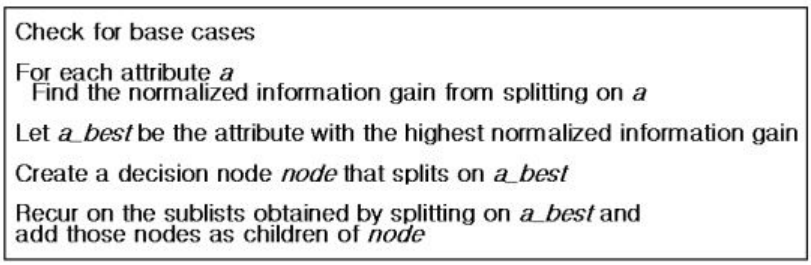

Figure 3. C4.5 Pseudocode of the algorithm.

Figure 3 shows the pseudocode of the C4.5 algorithm. Divide and Conquer operations are performed to construct decision tree with C4.5 algorithm. Construct a tree until it consists of cases where one class belongs to all subsets so that the input training set is successfully partitioned. The information gain ratio is used as a criterion for separating nodes. This is to separate the current training set in a way that can minimize the Average Information required to classify the given example. Suppose that the current training set is S and the number of cases belonging to the class Ci (i = 1, 2, ..., N) is Freq(Ci, S), The average information (Entropy) required to identify the class. The criteria using the gain ratio present much better results in the experiment than the gain. Therefore, profit ratio is used as separation criterion.

3. Experiment

3.1 Experimental data

In the United States, there are 60 million people with cardiovascular disease, of which one million people die from cardiovascular disease a year. About 100,000 (25%) of the 400,000 patients undergoing cardiac surgery in about a year and 1.5 million (5%) suffer perioperative cardiovascular disease costing over $ 2 billion a year among the 30 million who undergo extra-cardiac surgery. In particular, coronary artery disease

is one of the cardiovascular diseases that anesthesiologists commonly encounter. In the United States, the number of patients with coronary artery disease is increasing by 1.3 million people every year. They survive for a considerable period of time due to improved quality of care and about 1 million people experience anesthesia for noncardiac surgery every year. In this paper, we aim to find the correlation between the frequency of percutaneous coronary intervention and the mortality rate by region. Therefore, we extract the detailed region, number of cases, and number of death among 14 data values and find the association through the algorithm. The attributes selected in the experimental data are as follows.

No.	Label	Count
1	Capital District	5
2	Western NY - Rochester	2
3	Manhattan	7
4	Bronx	3
5	Kings	4
6	Western NY - Buffalo	4
7	Central NY	3
8	Queens	4
9	NY Metro - New Rochelle	2
10	NY Metro - Long Island	5

Figure 4. Detailed Region attributes

4. Experimental Results and Discussion

4.1 C4.5 Algorithm Results

Experiments were carried out by setting the Detailed Region attribute as a dependent variable and the number of cases and number of Death as independent variables based on the data Cardiac Surgery and Percutaneous Coronary Interventions. When analyzing the value, we performed the cross-validation with the fold value of 10. Figure 4 shows the analysis of C4.5.

```
Number of Deaths <= 8
|   Number of Cases <= 405: Central NY (11.0/8.0)
|   Number of Cases > 405
|   |   Number of Deaths <= 4
|   |   |   Number of Deaths <= 3: NY Metro - Long Island (3.0/1.0)
|   |   |   Number of Deaths > 3
|   |   |   |   Number of Cases <= 683: Capital District (3.0/1.0)
|   |   |   |   Number of Cases > 683: Manhattan (2.0/1.0)
|   |   Number of Deaths > 4: NY Metro - Long Island (4.0/2.0)
Number of Deaths > 8
|   Number of Cases <= 1334
|   |   Number of Deaths <= 9: Kings (4.0/2.0)
|   |   Number of Deaths > 9
|   |   |   Number of Deaths <= 11: Western NY - Buffalo (2.0)
|   |   |   Number of Deaths > 11: Capital District (2.0/1.0)
|   Number of Cases > 1334: Manhattan (8.0/3.0)
```

Figure 5. C4.5 Analysis table

Figure 5 shows that 17 nodes and 9 leaf nodes are derived. Nine leaf nodes were analyzed by region, frequency and number of deaths, and the corresponding region was derived at the last leaf node. Through this, it is possible to estimate the number of deaths, the frequency of operations, and the number of deaths and the frequency of operations in each region. Figure 6 is a visualization of the above table.

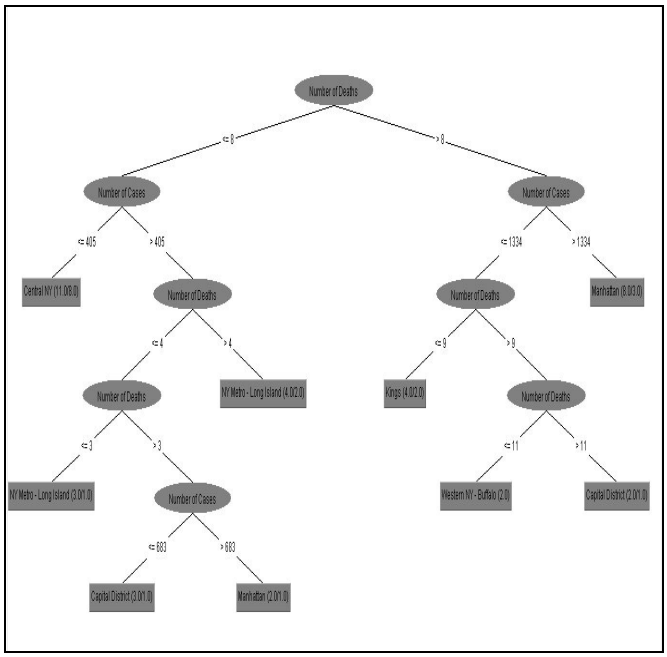

Figure 6. C4.5 Algorithm visualization

In Figure 7, a capital district is not classified as a, but is classified as two of c and one of two i of g. In other words, there was no case of a correctly classified capital district. b was classified as western NY-Rochester and c and I, respectively. c was mistakenly classified as a and f in Manhattan distance, but the remaining four were classified and showed higher accuracy than other values. As a result of classification to j in this way, the accuracy of the total data was about 20%.

```
a b c d e f g h i j   <-- classified as
0 0 2 0 0 0 2 0 1 0 | a = Capital District
0 0 1 0 0 0 0 0 0 1 | b = Western NY - Rochester
2 0 4 0 0 1 0 0 0 0 | c = Manhattan
0 0 2 0 1 0 0 0 0 0 | d = Bronx
0 0 0 0 0 1 1 1 0 1 | e = Kings
0 0 1 0 1 0 1 0 0 1 | f = Western NY - Buffalo
1 0 1 0 1 0 0 0 0 0 | g = Central NY
0 0 1 0 0 0 1 0 0 2 | h = Queens
0 0 0 0 1 0 0 0 0 1 | i = NY Metro - New Rochelle
2 1 1 0 0 0 0 1 0 0 | j = NY Metro - Long Island
```

Figure 7. Classification accuracy of C4.5

Figure 8 shows a visualization of the misclassified values represented by the matrix above. The color represented by the region is divided into the area where the color is drawn, and the square drawn on the area graph shows the position of the misclassified values.

Figure 8. Classifier Visualize errors using C4.5

5. Conclusion

Analysis of the results of this study showed that central NY and NY metro-Long Island had the lowest mortality rate compared to frequency, through decision tree and cluster analysis, and showed excellent survival rate of percutaneous coronary intervention. Manhattan has the highest mortality compared to other regions, indicating the lowest survival rate. This will lead to a better quality of medical treatment by analyzing the trend of medical treatment in regions where there is a high level of survival rate.

References

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, Morgan Kaufmann Publishers, 2011

[2] Yong-Gyu Jung, Jae-Jun Nam, Jae-kang Won, "Diabetes Management System using Decision Trees", IEIE 2017 FALL CONFERENCE, Vol.2012 No.11, pp. 796-799, Nov. 2012.

[3] Kurt, Imran, Mevlut Ture, and A. Turhan Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease." Expert systems with applications 34.1 (2008): 366-374.

[4] Acharya, U. Rajendra, et al. "Linear and nonlinear analysis of normal and CAD-affected heart rate signals." Computer methods and programs in biomedicine113.1 (2014): 55-68.

[5] Faghri, Pouran D., et al. "Circulatory hypokinesis and functional electric stimulation during standing in persons with spinal cord injury." Archives of physical medicine and rehabilitation 82.11 (2001): 1587-1595.

[6] Wolf, Aizik, et al. "Operative management of bilateral facet dislocation." Journal of neurosurgery 75.6 (1991): 883-890.http://blog.daum.net/hazzling/17067790

[7] Jang, Jae-Won, et al. "Vertebral artery injury after cervical spine trauma: a prospective study using computed tomographic angiography." Surgical neurology international 2 (2011).