Final Project
(MATH 3210 – Joan Muliawan)

Executive Summary

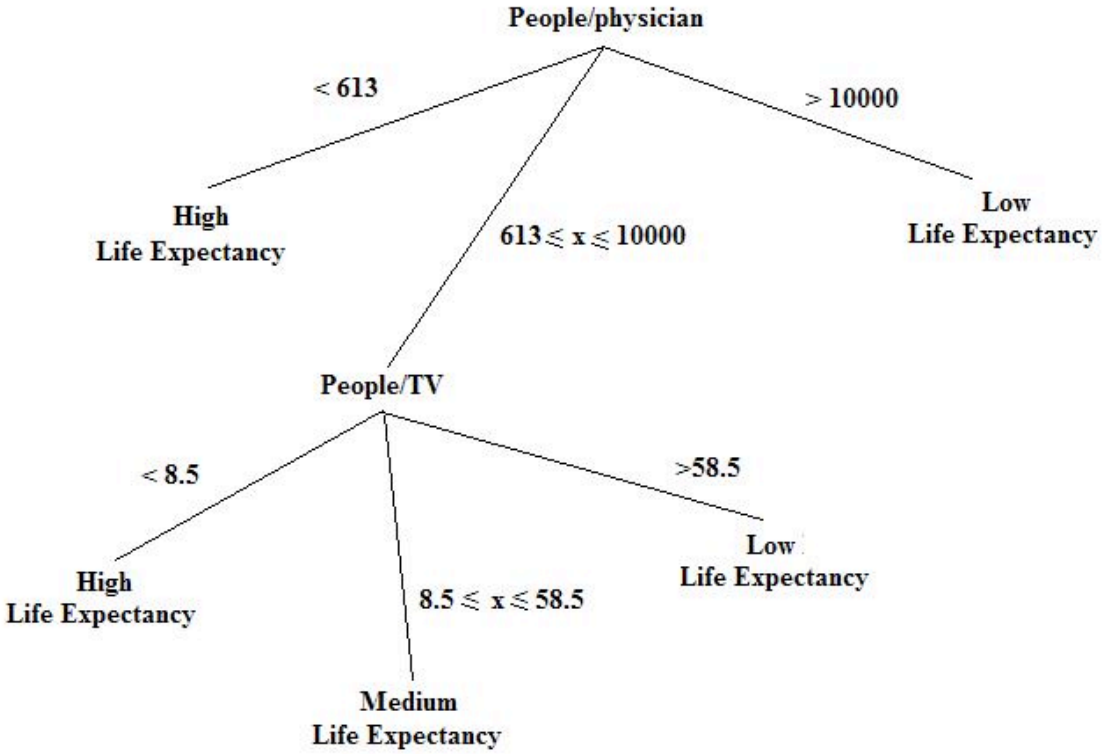
The data that I used for my final project have 6 attributes; country, people/TV, people/physician, female life expectancy, male life expectancy, and life expectancy. The life expectancy attribute is the mean value from female and male life expectancy. I split the life expectancy into three classes; low, medium, and high. Using correlation, ID3 entropy value and Self Organizing Map algorithm (also known as SOM), I would design an experiment using the dataset, determine which attribute is a better predictor of life expectancy, address the analysis technique and recommend a rule set.

SOM is an algorithm to visualize and interpret large high-dimensional data sets in an unsupervised learning category (*http://www.cis.hut.fi/research/som-research/som.shtml*). SOM uses four small programs; randinit.exe, vsom1.exe, vsom2.exe, and vcal.exe. The file reproduced by these 4 files would be in .cod extension. The main goal of this algorithm is to respond similarly to certain input patterns.

Correlation indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. In this research, I am going to use the correlation with aspect to the life expectancy.

ID3 is a non-incremental algorithm, meaning it derives its classes from a fixed set of training instances. A statistical property, called information gain, is used to decide which attribute is the better or even the best predictor. Gain measures how well a given attribute separates training examples into targeted classes.

By using correlation, ID3 entropy values, and information of gain, people/physician is a better predictor of life expectancy than people/TV. Thus, I use people/physician as the first splitting attribute. Below is a decision tree I recommend:



However, there is a 7.25% error for this rule set.

Problem Description

The problem is to analyze the dataset using correlation, ID3 entropy value, and SOM algorithm, and address the analysis technique. The other issue is I have two missing data for people/TV attributes in the dataset that I used. Sometimes, using the mean or mode value is the best thing to do to fill in the missing data. However, every country has different amount of population and TV, depends on its economic status, total birth, total death, etc. Thus, using a mean or mode value is not the best solution. I did some more research on the countries and fill in the missing data with the data that I have found in my sources.

Analysis Technique

The data that I used for my final project have 6 attributes; country, people/TV, people/physician, female life expectancy, male life expectancy, and life expectancy. The

life expectancy attribute is the mean value from female and male life expectancy. I split the life expectancy into three classes; low, medium, and high.

SOM is an algorithm to visualize and interpret large high-dimensional data sets in an unsupervised learning category (*http://www.cis.hut.fi/research/som-research/som.shtml*). The visualization process states by representing the central dependencies within the data on the map that consists of a regular grid of processing units called neurons. The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time, the models become ordered on the grid so that similar models are close to each other and dissimilar models are far from each other.

The SOM algorithm has four small programs, randinit.exe, vsom1.exe, vsom2.exe, and vcal.exe. After I defined the size from the input, randinit.exe initializes a 2-dimentional grid. Next, the vsom1.exe and vsom2.exe will do the e vsom1.exe and vsom2.exe then performs a training in 1000 steps and the finally, the vcal.exe will generates the visual mapping. For the dataset that I used, the training is performed with a grid size of 5x5. The map is described below:

Argentina	Brazil	Colombia	Iran	Myanmar
Romania	S Korea	India		
		Morocco	Kenya	Sudan
		Indonesia		Tanzania
	Bangladesh			Ethiopia

High	Medium	High	Medium	Low
High	High	Medium		
		Medium	Medium	Low
		Medium		Low
	Low			Low

As you see on the table above, it showed that the data clustered the country with a similar characteristic close together. Countries with low life expectancy are mapped close together, so does medium life expectancy and high life expectancy.

Correlation indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. I am going to use the correlation with aspect to the life expectancy. However, I am only going to analyze the correlation of people/TV and people/physician, because life expectancy is a mean from female and male life expectancy. Thus, the female and male life expectancy has the highest correlations among all other attributes. The correlation for people/TV and people/physician respectively are -0.623944 and -0.666.

ID3 is a non-incremental algorithm, meaning it derives its classes from a fixed set of training instances. A statistical property, called information gain, is used to decide which attribute is the better or even the best predictor. Gain measures how well a given attribute separates training examples into targeted classes. In order to define gain, we first borrow an idea from information theory called entropy. Entropy measures the amount of information in an attribute. Below are the formula for entropy and gain.

Entropy value = $\sum$ *P (i) log10 (1/P (i))*
(http://mercury.webster.edu/Aleshunas/MATH%203210/MATH%203210%20Home.htm)

$\sum$ is a symbol for summation

P (i) is the probability of the instance

Gain = entropy class – ((sample/population) * attribute entropy)
(http://mercury.webster.edu/Aleshunas/MATH%203210/MATH%203210%20Home.htm)

P (low) = 7/40

P (medium) = 13/40

P (high) = 20/40

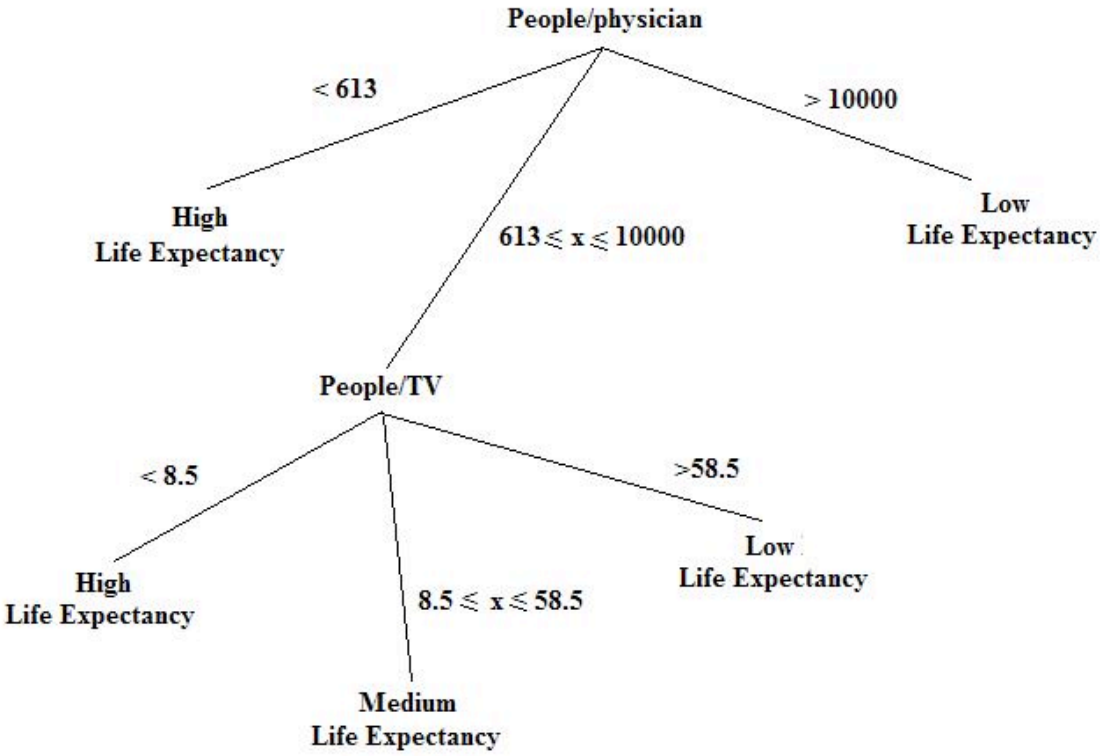
$$\textbf{Class Entropy} = \frac{7}{40}\log\frac{40}{7} + \frac{13}{40}\log\frac{40}{13} + \frac{1}{2}\log 2$$

$$= 0.298$$

Gain by using people/TV = 0.222

Gain by using people/physician = 0.283

Since using people/physician attribute is causing a greater gain, it is a better predictor of life expectancy than people/TV. A classification tree is below:



However, there are three data that is not mapped into the right class. Since there are 40 data in total, the accuracy of this classification is 92.5% (by using this formula: ((total data-wrong class)/total data) * 100%).

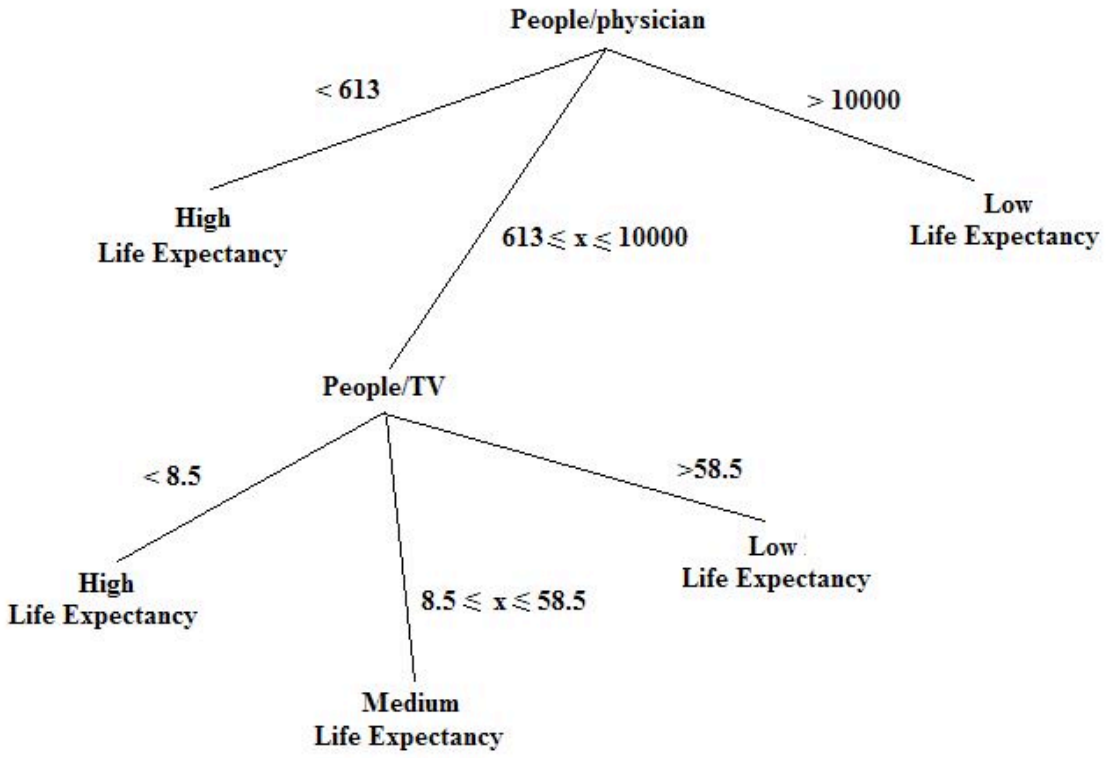
Assumptions

The dataset is a representative for classifying life expectancy.

Results

By using correlation, ID3 entropy values, and information of gain, people/physician is a better predictor of life expectancy than people/TV because it is

having a greater gain than people/TV attribute. Furthermore, I chose people/physician attribute to be the first splitting attribute. A classification tree is below:



Issues

I have two missing data values in the people/TV attribute for Tanzania and Zaire. Missing data values cause problems during both the training phase and the classification process itself. There are three ways of handling missing values:

- Ignore missing data
- Assume a value for the missing data
- Assume a special value for the missing data

In this case, I would use a special value for the missing data because every country has different amount of population and TV, depends on its economic status, total birth, total death, etc. Thus, using a mean or mode value is not the best solution. I did

some more research on the countries and fill in the missing data with the data that I have found in my sources.

Appendices

There is no appendix.

References

University of Missouri St. Louis, Retrieved April 28, 2008, from http://www.umsl.edu/services/govdocs/wofact93/wf940222.txt

University of Missouri St. Louis, Retrieved April 28, 2008, from http://www.umsl.edu/services/govdocs/wofact98/239.htm

University of Missouri St. Louis, Retrieved April 28, 2008, from http://www.umsl.edu/services/govdocs/wofact93/wf940250.txt

Altapedia, Retrieved April 28, 2008 from http://www.atlapedia.com/online/countries/DemRepCongo.htm

Math forum, Retrieved February 24, 2008 from http://mathforum.org/workshops/sum96/data.collections/datalibrary/data.set6.html

Laboratory of Computer and Information Science Lab, Retrieved April 23, 2008, from http://www.cis.hut.fi/research/som-research/som.shtml

Aleshunas, John. Retrieved April 30, 2008 from http://mercury.webster.edu/Aleshunas/MATH%203210/MATH%203210%20Home.htm