"Probability Entropy" Method for Data Mining: PIE Method - "A New Approach"

Ibrahem Tadros, Mohammad Al-Hiyassat, and Mohammad Al-Laham
Al-Balqa Applied University, Salt, Jordan

Tadros12@yahoo.com; mohiyassat@yahoo.com; Laham1st@yahoo.com

Abstract

Data mining is the process of analyzing data and summarizing it into useful information that can be used in decision making or to increase revenue, cuts costs, or both. Decision tree is a widely used approach in data mining and machine learning for classification problems, which is considered to be self-explained models and easy to follow when compacted, namely if it has a reasonable number of leaves. Statistical Data mining methods such as Bayes' Theorem or Conditional Probability Theory (special case) is widely used method. Introducing the PIE Method that is a combination of Bayes' theorem and Entropy concept from information theory together to improve Bayes' Theory method by minimizing number of possible combination needed to solve a problem, without losing Information. A comparison of the results of the 3 methods will be done on an example that represents a small database.

Keywords: Bayes' Method, Conditional Probability Method, Data Mining, Entropy, PIE Method, Statistical Data Mining.

Introduction

Generally, data mining or knowledge discovery is the process of analyzing data from different point of views and summarizing it into useful information that can be used to make decision or to increase revenue, cuts costs, or both. Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Hence, Data mining is the science and technology of exploring data in order to discover previously unknown patterns, the accessibility and abundance of information today makes data mining a matter of considerable importance and necessity. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases using some special data mining software's and tools used for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Definitions

Data Mining: "is the process of sifting through large amounts of data to produce data content relationships" (*Data mining*, n.d.). Data Mining consists of five major elements:

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

Data mining enable companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics.

Decision Tree: A graphical representation of all possible outcomes and the paths by which they may be reached; often used in classification tasks. The top layer consists of input nodes. Decision nodes determine the order of progression through the graph. The leaves of the tree are all possible outcomes or classifications, while the root is the final outcome (Decision Tree, n.d.). Decision trees are one of the most common data mining techniques and are by far the most popular in tools aimed at the business user. They are easy to set up, their results are understandable by an end user, they can address a wide range of classification problems, they are robust in the face of different data distributions and formats, and they are effective in analyzing large numbers of fields. A decision tree algorithm works by splitting a data set in a process called iteration in order to build a model that successfully classifies each record in terms of a target field or variable. On each iteration we need to choose the independent variable that most effectively splits the data. This means that the subsets produced by splitting the data according to the value of the independent variable should be as "homogeneous" as possible, with respect to the dependent variable. We shall compare the results of our 3 methods for the decision tree technique (Maki & Teranishi, 2001).

Entropy: "A measure used to determine the disorder in the population," according to Shannon Information Theory (Shannon & Weaver, 1959). Shannon's formula for calculating entropy is Entropy = $- (\Sigma\ p_i \log_2 p_i)$ Where

$$\log_2 p_i = (\log_{10} p_i / \log_{10} 2).$$

Bayes' Method: Bayesian Theory that was named after mathematician Thomas Bayes is applied to revise probabilities of an event after we obtain more information. The following is Bayes' formula for a two event case:

$$P(A_1/B) = C / (C + E)$$

$$C = P(A_1) * P(B/A_1),\ E = P(A_2) * P(B/A_2)$$

The Conditional Probability (as a special case): is the probability that an event will occur given that another event has already occurred (Michael & Linoff, 1997) if A and B are two events, then the conditional probability of A is written as P(A\B) and read as "the probability of A given that B has already occurred", i.e. If A and B are two events, then:

$P(B\backslash A) = P(A \text{ and } B) / P(A)$ and $P(A\backslash B) = P(A \text{ and } B) / P(B)$,Given that $P(A) \neq 0$ and $P(B) \neq 0$ (Rolfe & Brown, 1997).

Example

Lets assume that we are in a company that sells cars as shown in Table 1 and our database consists of 4 columns (Size, Color, Manufacture Place, Satisfaction), the manager wants to know

"according to customer Satisfaction which is measured by the "time needed to sell this car"" is it good for us to buy large, green car from Europe? Or medium, red cars from Japan? Or small, blue cars from USA?

Table 1: A Company Sale

Size	Color	Manufacture Place	Satisfaction
S	R	USA	NO
M	R	ASIA	YES
S	R	ASIA	YES
L	B	USA	NO
S	B	JAP	YES
S	G	USA	NO
M	G	JAP	YES
M	B	EU	YES
M	R	ASIA	NO
L	G	JAP	YES
M	B	JAP	NO
S	B	USA	YES
S	R	EU	YES
M	R	ASIA	YES
L	G	EU	NO
M	G	USA	YES
S	R	EU	YES
S	R	USA	NO
M	R	USA	YES
L	G	EU	NO
L	B	ASIA	YES
L	G	EU	NO
M	B	USA	YES
M	R	JAP	YES
S	R	JAP	NO
S	G	USA	YES
S	B	JAP	YES
M	R	ASIA	NO
S	R	JAP	YES
M	R	USA	YES

Entropy Method: Start from Satisfaction, split the table according to satisfaction and repeating this until reaching the end to obtain the sequence of splitting for each branch according to satisfaction to get the following tree with the following rules (Shown in Figure 1):

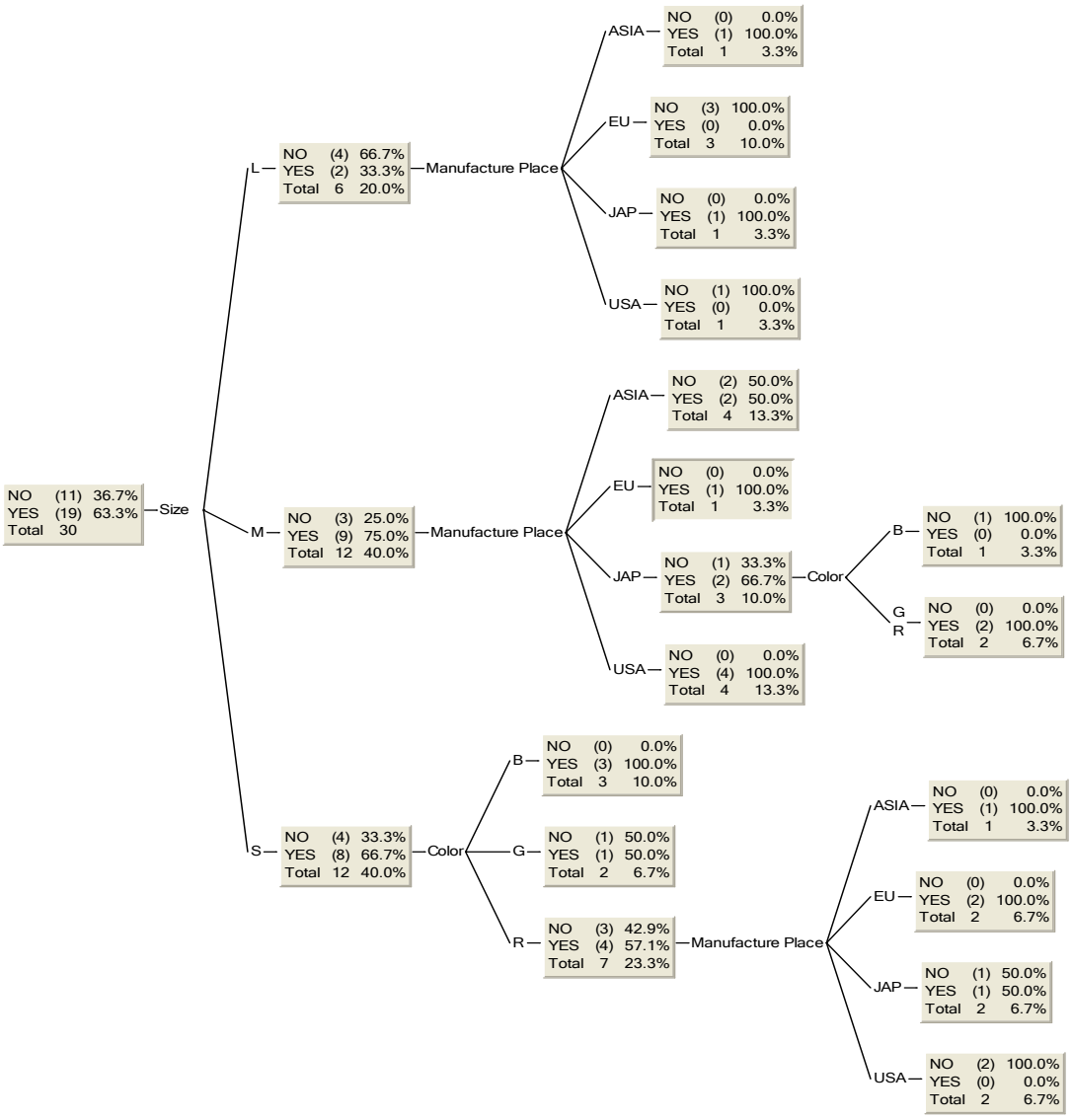

Figure 1: Decision tree graph using Entropy Method.

Satisfaction = YES 100.0%
RULE_1 IF
Manufacture Place = ASIA
Size = L
RULE_2 IF
Manufacture Place = JAP
Size = L
RULE_3 IF
Manufacture Place = EU
Size = M
RULE_4 IF
Color = G or R
Manufacture Place = JAP
Size = M
RULE_5 IF
Manufacture Place = USA
Size = M
RULE_6 IF
Color = B
Size = S
RULE_7 IF
Manufacture Place = ASIA
Color = R
Size = S
RULE_8 IF
Manufacture Place = EU
Color = R
Size = S

Satisfaction = NO 100.0%
RULE_1 IF
Manufacture Place = EU
Size = L
RULE_2 IF
Manufacture Place = USA
Size = L
RULE_3 IF
Color = B
Manufacture Place = JAP
Size = M
RULE_4 IF
Manufacture Place = USA
Color = R
Size = S

Satisfaction = NO 50.0% **Satisfaction = YES 50.0%**
RULE_1 IF
Manufacture Place = ASIA
Size = M
RULE_2 IF
Color = G
Size = S
RULE_3 IF
Manufacture Place = JAP
Color = R
Size = S

Buy Cars if they are: (Small, Europe, Red), (Small, USA, Black), (Small, Japan, Black), (Small, ASIA, Red), (Medium, USA *), (Medium, Europe, Black), (Medium, JAP, Red), (Medium, JAP, Green), (Large, JAP, Green), (Large, ASIA, Black). Don't Buy Cars if they are: (Small, USA, Red), (Medium, Japan, Black), (Large, Europe, Green), (Large, USA, Black). The Benefits of Entropy Method is that it tells you about data that should be removed from the original database (those having 50% for both Satisfaction (NO) and (YES)), the Draw back of this method is the little bit complicated calculations (Baldwin, n.d.).

Bayes' Method: Start from Satisfaction and then consider all possible combination of the other factors (Size, Color, and Manufacture Place), this will give the following probabilities:

1. (Size, color, Manufacture Place)
2. (Size, Manufacture Place, color)
3. (Color, Manufacture Place, size)
4. (Color, size, Manufacture Place)
5. (Manufacture Place, color, size)
6. (Manufacture Place, size, color)

This should be done for both “Satisfaction attributes (YES, NO)”, (i.e. we will have 12 probability tree structure). Calculate the conditional probability for each branch of the tree to get rules for purchasing new cars using this method as follows (note that I will consider only the sets with high probability and not all possibilities, sets will be placed in order):

Buy Car's if they are: (Medium, Asia, Red)*, (Small, Europe, Red), (Small, Japan, Black) or (Medium, USA, Red). Don't Buy Cars if they are: (Large, Europe, Green), (Medium, Asia, Red)*, (Small, USA, Red) or (Jap, Medium, Red). The Benefits of the Probability Method is that it tells you about new markets (i.e. what options are not been selected in your database) and it can be calculated easily, the Draw back of this method is that it's too long, have many options and it didn't distinguish redundant sets of elements.

PIE Method : In this method I have tried to combine Probability and Entropy together to improve Bayes' Method (Paolo, 2003) the aim of this method is minimize the number of all possible combination of our factors (Size, Color, and Manufacture Place) in the Probability Method (those were 12 possibility) without losing information.

The Algorithm states that:

Use the Entropy Method to eliminate redundant records (Baldwin, n.d.).

Split the first step according to your need, and calculate the Entropy to determine the next splitting attribute.

Consider each main group as a separate group and start splitting according to the second step mentioned in 2.

Next, take all the rest of attribute combination into consideration.

Calculate the Entropy to determine the next step for each attribute.

Go to 4 and repeat until no other choices are left.

Once finished Check the path for the finish line and take decision.

Finish

---

* **Using Entropy Method this set should be eliminated**

Applying this method to our database I have found the following results: Start from Satisfaction and Calculate Information gain that will lead you to Size. Construct your Tree as having 2 main branches "Satisfaction (YES)" and "Satisfaction (NO)". Start with "Satisfaction (YES)" you will get the following tree that tells you the path you should take in the Probability Tree with the following rules (shown in Figure 2).

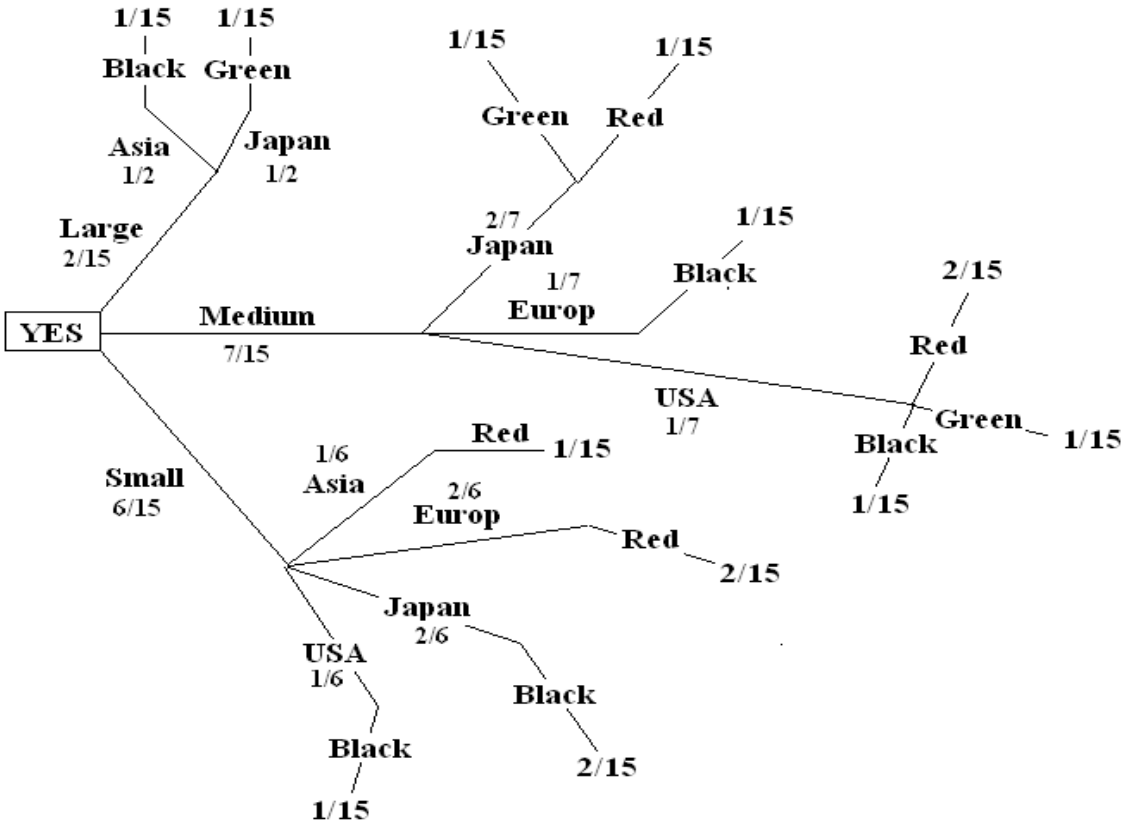

Figure 2: Probability tree using Entropy to determine path.

With “Satisfaction (NO)” (Figure 3)

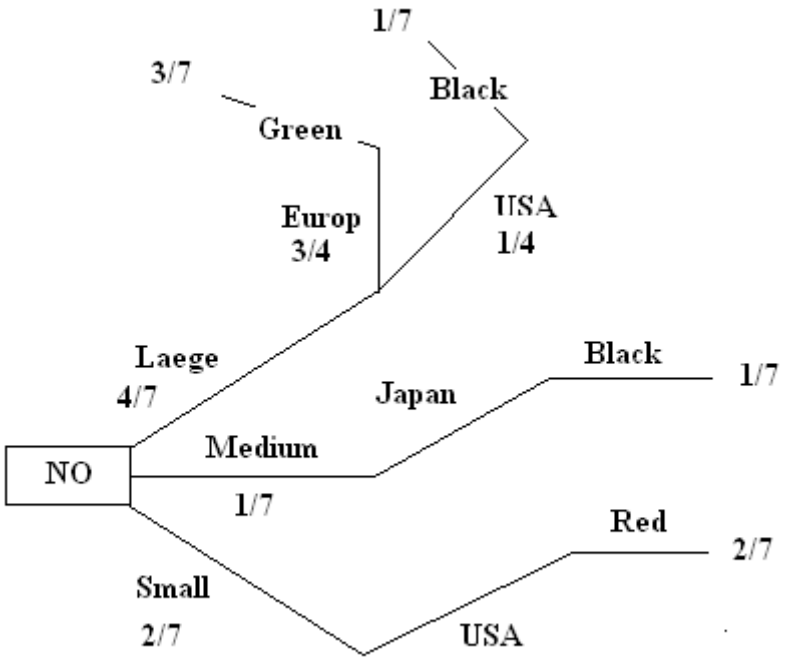

Figure 3: Probability tree using Entropy to determine path

Conclusion

The result using Entropy Method was:

> Buy Cars if they are: (Small, Europe, Red), (Small, USA, Black), (Small, Japan, Black), (Small, ASIA, Red), (Medium, USA, *), (Medium, Europe, Black), (Medium, JAP, Red), (Medium, JAP, Green), (Large, JAP, Green), (Large, ASIA, Black).
>
> Don't Buy Cars if they are: (Small, USA, Red), (Medium, Japan, Black), (Large, Europe, Green), (Large, USA, Black)

The result using Probability Method was:

> Buy Car's if they are: (Medium, Asia, Red), (Small, Europe, Red), (Small, Japan, Black) or (Medium, USA, Red).
>
> Don't Buy Cars if they are: (Large, Europe, Green), (Medium, Asia, Red), (Small, USA, Red) or (Jap, Medium, Red)

The result using PIE Method was:

> Buy Car's if they are: (Small, Europe, Red) or (Small, Japan, Black) or (Medium, USA, Red)
>
> Don't Buy Cars if they are: (Small, USA, Red) or (Large, Europe, Green)

References

Baldwin, J. (n.d.). *Rules for data mining*. Retrieved from http://www.enm.bris.ac.uk/teaching/enjfb/emat31600/Rules.pdf

Data mining. (n.d.) Retrieved from http://www.microstrategy.com/News/Glossary/Letter_d.htm

Decision Tree. (n.d.) Retreived from http://amsglossary.allenpress.com/glossary/browse?s=d&p=8

Shannon, C. E., & Weaver, W. (1959). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Maki, H. & Teranishi, Y. (2001). *Development of automated data mining system for quality control in manufacturing*, DaWaK, 93-100.

Michael, J.A. & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons.

Paolo, G. (2003). *Applied data mining: Statistical methods for business and industry.*

Rolfe, D. F. S., & Brown, G.C. (1997). Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiol. Rev., 77*, 731-758.