Chapter 4

Analysis of Health Conditions Using Varma Entropy Based ID3 Algorithm

4.1 Introduction

Shannon (1948), introduced the concept of measure of information or entropy for a general finite complete probability distribution $P = (p_1, p_2, p_3, ..., p_n)$ associated with a discrete random variable $X = (x_1, x_2, x_3, ..., x_n)$ for $n \geq 2$ is given by

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i,\ 0 \leq p_i \leq 1\ and\ \sum_{i=1}^{n} p_i = 1. \tag{4.1}$$

One of the main drawback of (4.1) is, if $p_i = 1/n$ i.e. uniform distribution, then it is very difficult to make any prediction about such experiments. Moreover, different random variables with different probabilities associated may contribute same information. To remove such limitations, various generalizations of (4.1) have been introduced in literature, it mainly started

with Renyi's entropy (Renyi (1961)) of order *alpha*, given by

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_0^\infty p_i^\alpha\right), \alpha > 0 \ \ and \ \ \alpha \neq 1. \tag{4.2}$$

If $\alpha \to 1$ then (4.2) reduce to (4.1). Further, generalization process is extended by different researchers, refer to Mathai and Haubold (2007), Ochs (1976) etc and the generalized measures are widely used in various disciplines like Survival Analysis (refer to Ebrahimi (1996)), Pattern recognition (refer to Chen (1973)) etc.

An important two parametric generalization of the Shannon entropy (4.1) is Varma entropy (Varma (1966)), given by

$$H_{(\alpha,\beta)}(X) = \frac{1}{\beta-\alpha}\log\left[\sum_{i=1}^{n}(p_i)^{\alpha+\beta-1}\right], (\beta-1) \leq \alpha \leq \beta \ and \ \beta \geq 1. \tag{4.3}$$

When β=1 and $\alpha \to 1$, then (4.3) reduces to (4.1). It has similar properties as that of Shannon entropy, but it contains additional parameters 'α' and 'β' which can be used to make it more or less sensitive to the shape of probability distributions.

It can be observed form the work of different researchers like Jin et al. (2009), Kumar et al. (2015), Patel and Rana (2014), Xu (2006) that current active area of research is the application of entropy in the field of data mining, in which results can be enhanced with the help of improved classifying algorithms. ID3 (refer to Quinlan (1986)), C4.5 (refer to Quinlan (2014)), ASSISTANT (refer to Kononenko (1984)) etc. are mostly used algorithms of data mining with the mission of classification. In this chapter, we have considered ID3 algorithm, one of the key algorithms, that produces a decision tree by calculating entropies of all attributes and then the values of each attribute are used in order to derive generalized rules from a given sample set. Classification algorithms of data mining can be modified using different measures of information to get the desired results with more accuracy. An attribute is selected as root of the decision tree

on the basis of the maximum gained information and then the process is repeated again for its sub trees.

In the algorithm of decision tree, information gain plays important role in identifying appropriate attribute of every node of the tree. In ID3 algorithm, information gain is obtained by using the concept of Shannon entropy, refer to Quinlan (1986). In the preceding chapter, we have already applied the concept of Renyi entropy (Renyi (1961)) to obtain information gain in development of decision tree using ID3 algorithm (Kumar et al. (2012)). In this chapter, we have used Varma entropy (Varma (1966)) of two parameters α and β in ID3 algorithm and have derived some rules for the given data under consideration. We have applied the algorithm designed to classify the health conditions in India on the basis of the data for the year 2015. The work reported in this Chapter has been published in the paper entitled,**"Analysis of Health Conditions Using Generalized Information Measure Based ID3 Algorithm"**, in 4th Annual International Conference on Operations Research and Statistics (ORS 2016), Singapore, 2016.

4.2 Procedure To Generate Decision Tree

As already outlined in Chapter 3, the procedure to generate decision tree is given as follows:

Input: Specific data, having different properties displayed by discrete values and classes.

Procedure: Select the property with maximum information gain for first node known as root. For its various branches, which depends upon its different number of ranges, repeat the step until all the values are in atleast one class.

For more details about pseudo code and algorithm of decision tree, refer to Qian et al. (2007).

4.3 Algorithm of Information Gain for Varma Entropy

In the algorithm of decision tree, appropriate attribute is selected on the basis of information gain for every node. ID3 algorithm based upon the concept of information gain and selects the property having maximum information gain in given specific data set, which has some given ranges for further branches of decision tree. The selected attribute reflects the minimum randomness for classification, refer to Han and Kamber (2000). This information entropy reduces the number of steps which are needed for the classification and ensures to find a simple decision tree. Let us consider a specific data 'D' having 'N' number of different values with 'R_j' where $j = 1, 2, 3, \cdots, r$ number of ranges for each 'P_k' property, where $k = 1, 2, 3, \cdots, m$ and the data is to be divided into classes 'C_i' where $i = 1, 2, 3, \cdots, c$ number of classes. Now, the quantity of the information required for the object data is calculated using Varma entropy (Varma (1966)), as follows:

$$I(C_1, C_2, C_3, \cdots, C_c) = \frac{1}{\beta - \alpha} \log_2 \sum_{i=1}^{c} p_i^{\alpha+\beta-1} , \tag{4.4}$$

here p_i represents the probability of the values in particular class.

Let us assume that the property 'P_k' is selected from the set. Using this property, 'D' can be divided into 'R_j' number of sets which has same number of values as that of in 'D' and entropy for the particular property is calculated using equation (4.5) based upon the total number of the values of 'D'. If 'm_j' is the number of values in the range 'R_j' of property 'P_k' then entropy of 'P_k' that is 'E(P_k)'is given as:

$$E(P_k) = \frac{1}{\beta - \alpha} \sum_{j=1}^{r} \frac{m_j}{N} \log_2 \sum_{i=1}^{c} p_i^{\alpha+\beta-1} . \tag{4.5}$$

Here, 'r' represents the number of ranges, in particular property and thus

the net gained information $G(P_k)$ from the property 'P_k' is

$$G(P_k) = I(C_1, C_2, C, \cdots, C_c) - E(P_k). \tag{4.6}$$

Information gains and entropies are calculated for each property. The property having maximum value of information gain becomes the root of the decision tree. Further, branches of the tree are developed on the basis of classes of the root. Other properties are rearranged accordingly. This process is repeated for other sub trees until we get the leaves of the decision tree. The resulted decision tree helped in the development of some rules.

The classification mainly deals with extraction of information of the system and its systematic development, and to achieve this objective the best solution is the process in which entropy is the least.

4.4 Health Conditions in India in 2015

The data of different States and Union Territories of India has been collected from https://data.gov.in/ in March, 2015 and summarized in Table 4.1. The complete data is divided into two classes; one as C_1 which represents 'g' means 'good' and indicates good health conditions while another as C_2 which represents 'ng' means 'not good' and indicates bad health conditions. Mainly, four attributes have been selected form the source so that these can help in selecting appropriate health related projects to be implemented in regions based on the classification of health condition as 'good' or 'not good'. Each attribute is further divided into four ranges, which depend upon their values either less than or greater than national average based on per unit population. Corresponding average of particular region is considered as 'good' if it is more than that of national average, otherwise 'not good'. Further, we develop decision tree using modified ID3 algorithm based upon Varma entropy (Varma (1966)) measure having two parameters as α and β.

The details of the four attributes selected for classification are:

'PHC' which represents average population covered by Public Health Centre having four ranges as: less than 26000, 26000-47000, 48000-61000 and greater than 61000. Here, to make the continuity in the ranges selected, we have considered the values less than or equal to 47500 in the range 26000-47000, while the values greater 47500 has been considered in the next range i.e. 48000-61000. The same consideration has been done for the other properties also.

'SBC' represents average population covered by Sub Centre functioning at the end of Twelfth Plan (As on 31st March 2014)-(2012-2017), having four ranges as: less than 3600, 3600-4500, 4600-7100 and greater than 7100.

'CHC' represents average population covered by Community Health Centre functioning at the end of Twelfth Plan (2012-2017) as on 31st March 2014 having ranges as: less than 140000, 140000-220000, 230000-320000 and greater than 320000.

'RPHC' represents average Rural Population covered by Primary Health Centre having ranges as: less than 22000, 22000-31000, 32000-48000 and greater than 48000.

The next step is to select appropriate values for the parameters α and β of Varma entropy (Varma (1966)). As per the requirement of ID3 algorithm the value of information gain must be positive. So, we have drawn the graph (Fig. 4.1) for the information measure of example under consideration, where $x = (\alpha+\beta-1)$ is along X-axis and information required is along Y-axis. It is clear from the graph that values of two parameters α and β must satisfy $0 \leq (\alpha + \beta - 1) \leq 1$ along with the conditions by definition of Varma entropy i.e. $(\beta - 1) < \alpha < \beta$ and $\beta \geq 1$. So, we have considered α=1/10 and β=1 to measure entropy or information gain.

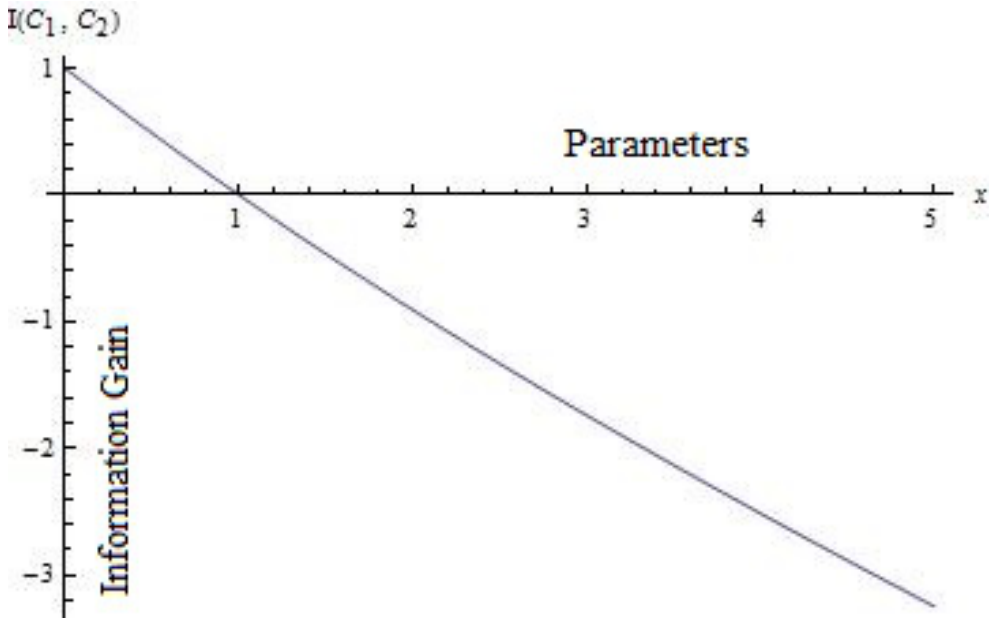

Figure 4.1: Plot for Information Gain

Table 4.1:: Health Conditions in INDIA-2015

Sr. no.	PHC	SBC	CHC	RPHC	g and ng
1	< 26000	< 3600	< 140000	< 22000	ng
2	$48000 - 61000$	$4500 - 7100$	$230000 - 320000$	$32000 - 48000$	g
3	< 26000	$4500 - 7100$	< 140000	< 22000	ng
4	$26000 - 47000$	$4500 - 7100$	$140000 - 220000$	$22000 - 31000$	ng
5	$48000 - 61000$	> 7100	> 320000	> 48000	g
6	< 26000	> 7100	> 320000	< 22000	ng
7	$26000 - 47000$	$4500 - 7100$	$140000 - 220000$	$22000 - 31000$	ng
8	$48000 - 61000$	$4500 - 7100$	> 320000	$22000 - 31000$	g
9	> 61000	> 7100	< 140000	< 22000	ng
10	> 61000	> 7100	< 140000	> 48000	g
11	> 61000	$4500 - 7100$	> 320000	$22000 - 31000$	g
12	$48000 - 61000$	> 7100	$140000 - 22000$	$22000 - 31000$	ng
13	$48000 - 61000$	> 7100	$230000 - 320000$	$32000 - 48000$	g

14	< 26000	< 3600	< 140000	< 22000	ng
15	< 26000	$4500 - 7100$	$140000 - 220000$	< 22000	ng
16	> 61000	> 7100	$140000 - 220000$	> 48000	g
17	$26000 - 47000$	$4500 - 7100$	> 320000	< 22000	ng
18	$26000 - 47000$	> 7100	$140000 - 220000$	< 22000	ng
19	< 26000	$4500 - 7100$	< 140000	< 22000	ng
20	> 61000	> 7100	$140000 - 220000$	$32000 - 48000$	g
21	> 61000	> 7100	$230000 - 320000$	$32000 - 48000$	g
22	$26000 - 47000$	$4500 - 7100$	$140000 - 220000$	$22000 - 31000$	ng
23	$26000 - 47000$	$4500 - 7100$	< 140000	< 22000	ng
24	< 26000	< 3600	< 140000	< 22000	ng
25	< 26000	$4500 - 7100$	< 140000	< 22000	ng
26	$26000 - 47000$	$4500 - 7100$	< 140000	$22000 - 31000$	ng
27	$48000 - 61000$	> 7100	> 320000	< 22000	g
28	> 61000	> 7100	$140000 - 220000$	$32000 - 48000$	g
29	$26000 - 47000$	$4500 - 7100$	< 140000	$22000 - 31000$	ng
30	< 26000	$3600 - 4500$	$230000 - 320000$	< 22000	ng
31	$48000 - 61000$	> 7100	$140000 - 220000$	$22000 - 31000$	ng
32	$26000 - 47000$	$3600 - 4500$	$140000 - 220000$	$32000 - 48000$	ng
33	$48000 - 61000$	> 7100	$230000 - 320000$	$32000 - 48000$	g
34	$26000 - 47000$	$4500 - 7100$	$140000 - 220000$	$22000 - 31000$	ng
35	> 61000	> 7100	$230000 - 320000$	> 48000	g

In Table 4.1, the complete data is divided into two classes i.e. c=2 and as there are 13 number of 'g' in C_1 and 22 number of 'ng' in C_2. Therefore the required information $I(C_1, C_2)$ in bits is;

$$I(C_1, C_2) = \left\{ \frac{1}{1 - \frac{1}{10}} \right\} \log_2 \left\{ \left\{ \frac{13}{35} \right\}^{\frac{1}{10}+1-1} + \left\{ \frac{22}{35} \right\}^{\frac{1}{10}+1-1} \right\}$$

Further, the entropy measure i.e. $E(P_k)$ in bits associated with the four different attributes at $\alpha = \frac{1}{10}$ and $\beta = 1$ are given as follows:

$$\begin{aligned}
E(P_1) &= E(PHC) \\
&= \left\{\frac{1}{1-\frac{1}{10}}\right\}\left[\frac{9}{35}\log_2\left\{\left\{\frac{9}{9}\right\}^{\frac{1}{10}+1-1}\right\}\right. \\
&\quad +\frac{8}{35}\log_2\left\{\left\{\frac{1}{8}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{7}{8}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad +\frac{10}{35}\log_2\left\{\left\{\frac{10}{10}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad \left.+\frac{8}{35}\log_2\left\{\left\{\frac{2}{8}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{6}{8}\right\}^{\frac{1}{10}+1-1}\right\}\right]
\end{aligned}$$

$$\begin{aligned}
E(P_2) &= E(SBC) \\
&= \left\{\frac{1}{1-\frac{1}{10}}\right\}\left[\frac{3}{35}\log_2\left\{\left\{\frac{3}{3}\right\}^{\frac{1}{10}+1-1}\right\}\right. \\
&\quad +\frac{15}{35}\log_2\left\{\left\{\frac{10}{15}\right\}^{\frac{1}{10}+1-1}\left\{\frac{5}{15}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad +\frac{2}{35}\log_2\left\{\left\{\frac{2}{2}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad \left.+\frac{15}{35}\log_2\left\{\left\{\frac{3}{15}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{12}{15}\right\}^{\frac{1}{10}+1-1}\right\}\right]
\end{aligned}$$

$$\begin{aligned}
E(P_3) &= E(CHC) \\
&= \left\{\frac{1}{1-\frac{1}{10}}\right\}\left[\frac{11}{35}\log_2\left\{\left\{\frac{1}{11}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{10}{11}\right\}^{\frac{1}{10}+1-1}\right\}\right. \\
&\quad \left.+\frac{6}{35}\log_2\left\{\left\{\frac{4}{6}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{2}{6}\right\}^{\frac{1}{10}+1-1}\right\}\right.
\end{aligned}$$

$$+\frac{12}{35}\log_2\left\{\left\{\frac{3}{12}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{9}{12}\right\}^{\frac{1}{10}+1-1}\right\}$$
$$+\frac{6}{35}\log_2\left\{\left\{\frac{5}{6}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{1}{6}\right\}^{\frac{1}{10}+1-1}\right\}\Bigg]$$

$$\begin{aligned}
E(P_4) &= E(RPHC) \\
&= \left\{\frac{1}{1-\frac{1}{10}}\right\}\Bigg[\frac{14}{35}\log_2\left\{\left\{\frac{1}{14}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{13}{14}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad +\frac{4}{35}\log_2\left\{\left\{\frac{4}{4}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad +\frac{10}{35}\log_2\left\{\left\{\frac{2}{10}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{8}{10}\right\}^{\frac{1}{10}+1-1}\right\} \\
&\quad +\frac{7}{35}\log_2\left\{\left\{\frac{1}{7}\right\}^{\frac{1}{10}+1-1}+\left\{\frac{6}{7}\right\}^{\frac{1}{10}+1-1}\right\}\Bigg]
\end{aligned}$$

Net gained information i.e. $Gain(P_k)$ for different attributes is given by;

$$Gain(PHC) = I(C_1, C_2) - E(P_1) = 0.55605\ bits.$$
$$Gain(SBC) = I(C_1, C_2) - E(P_2) = 0.15524\ bits.$$
$$Gain(CHC) = I(C_1, C_2) - E(P_3) = 0.03534\ bits.$$
$$Gain(RPHC) = I(C_1, C_2) - E(P_4) = 0.16591\ bits.$$

We notice that gain information of 'PHC' is largest, therefore it is considered as the root of the tree. As there are four different ranges of 'PHC' therefore we get four different sub-trees. Further, the above process of information gain is repeated for subtrees again, until we get leaves as 'g' or 'ng', associated calculations are given in *Appendix B*. Resultant decision tree for the data in Table 4.1, obtained by using modified ID3 algorithm is given in Fig. 4.2.

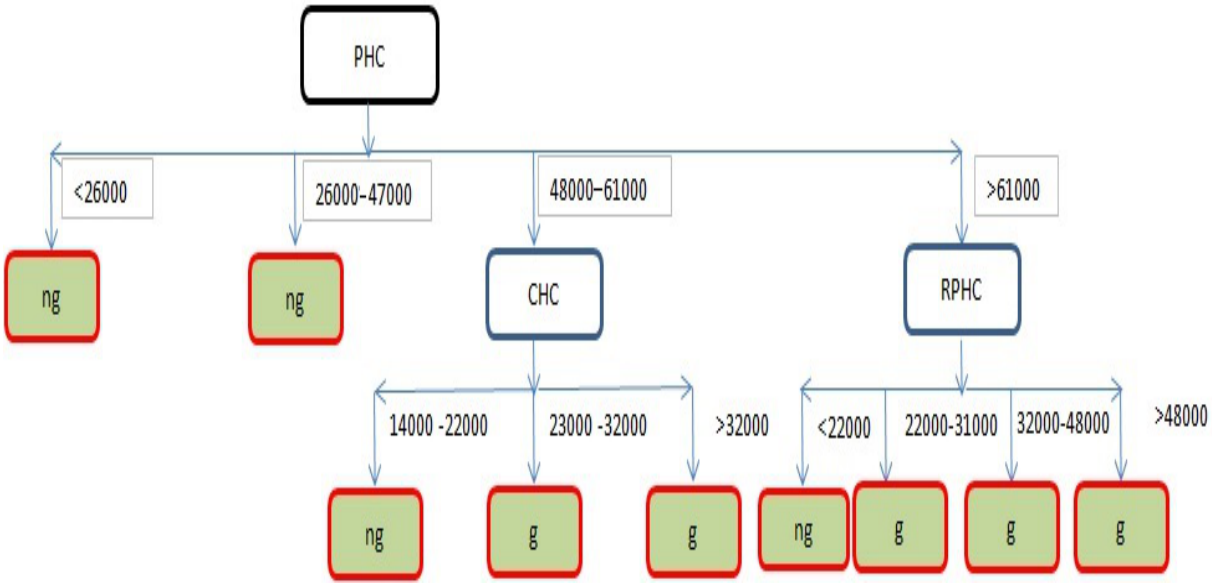

Figure 4.2: Varma Tree

4.5 Rules Induced

Some rules have been induced, on the basis of resultant decision tree, using 'if-then' to improve the decision making criteria during implementation of different health policies or projects of development. There are total 9 rules out of which only 4 rules are classifying the health condition as 'not good' (i.e. ng) and rest in all the regions have good health related services. These can be described as follows:

1. If PHC are '< 26000', then 'ng'; It means that for any value of SBC, CHC and RPHC, if PHC are less than 26000 then health conditions are 'not good'.

2. If PHC are '$26000 - 47000$', then 'ng'; It means that for any value of SBC, CHC and RPHC, if PHC are within 26000 to 47000 then health conditions are 'not good'.

3. If PHC are '$48000 - 61000$' and CHC are '$14000 - 22000$', then 'ng'; It means that for any value of SBC and RPHC, if PHC are within 48000 to 61000 and CHC are with in 14000 to 22000 then health conditions

are not good.

4. If PHC are within '48000−61000' and CHC are within '23000−32000', then 'g'; It means that for any value of SBC and RPHC, if PHC are within 48000 to 61000 and CHC are with in 23000 to 32000 then health conditions are good.

5. If PHC are within '48000 − 61000' and CHC are greater than '32000', then 'g'; It means that for any value of SBC and RPHC, if PHC are within 48000 to 61000 and CHC are more than 32000 then health conditions are good.

6. If PHC are greater than '61000' and RPHC are less than '22000', then 'ng'; It means that for any value of SBC and CHC, if PHC are greater than 61000 and RPHC are less than 22000 then health conditions are not good.

7. If PHC are greater than '61000' and RPHC are within '22000−31000', then 'g'; It means that for any value of SBC and CHC, if PHC are greater than 61000 and RPHC are within 22000 to 31000 then health conditions are good.

8. If PHC are greater than '61000' and RPHC are within '32000−48000', then 'g'; It means that for any value of SBC and CHC, if PHC are greater than 61000 and RPHC are within 32000 to 48000 then health conditions are good.

9. If PHC are greater than '61000' and RPHC are greater than '48000', then 'g'; It means that for any value of SBC and CHC, if PHC are greater than 61000 and RPHC are greater than 48000 then health conditions are good.

4.6 Conclusion

A project is said to be successfully implemented if its target is achieved in least possible time with desired results. But it is possible only if right decision is taken at right time. Here, the decision tree has been obtained by improved ID3 algorithm based upon generalized information theoretic measure having two parameters α =1/10 and β =1. The collected data is classified into two different classes, which helps us in developing more refined rules in comparison, if modified ID3 algorithm is used for classification. Thus, we can get more clear idea while making any final decision about implementation of, or analysis of, any implemented project.