Chapter 3

Classification of Census Using Renyi Entropy Based ID3 Algorithm

3.1 Introduction

Shannon (1948) introduced the concept of measure of information or entropy for a general finite complete probability distribution $P = (p_1, p_2, p_3, ..., p_n)$ given by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i,\ 0 \leq p_i \leq 1\ and\ \sum_{i=1}^{n} p_i = 1. \tag{3.1}$$

The notation of entropy is of fundamental importance in different areas such as physics, probability and statistics, communication theory and economics. Shannon entropy plays an important role in the context of information theory. To date, one of the most widely benefiting application has been for data compression and transmission, refer to Ash (1990).

Since the pioneering work of Shannon (1948), the concept of entropy has been generalized in a number of different ways by different researchers, refer to Taneja (2001). Entropy and its various generalizations are widely used in mathematical statistics, communication theory, physical and com-

puter sciences for characterizing the amount of information in a probability distribution, refer to Thomas and Cover (2009). The Renyi entropy (Renyi (1961)) is an additive generalization of Shannon entropy defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left[\sum_{i=1}^{n} (p_i)^\alpha \right] ; \ \alpha \neq 1, \ \alpha \geq 0. \tag{3.2}$$

It has similar properties as that of Shannon entropy, but it contains additional parameter 'α' which can be used to make it more or less sensitive to the shape of probability distributions, refer to Renyi (1961). If it has large positive value this measure is more sensitive to events that occur often, while for large negative values of 'α', it is more sensitive to the events which happen seldom, refer to Maszczyk and Duch (2008).

An active area of current research in the application of entropy is in data mining, refer to Marakas (2003). In the algorithms of data mining, the classification is an important mission. Decision tree algorithm, one of the key algorithms, is commonly used to build predictive model for classification. Using the different concepts of measurement of information the classification algorithms of data mining can be modified to get desired results with more accuracy. On the basis of gained information one property is selected as root of the tree 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. This information gain is obtained by using the concept of Shannon entropy in the different algorithms employed, refer to Akgobek (2011). The mostly used decision tree producing algorithms are ID3 and C4.5. These algorithms produce a decision tree by calculating entropies of attributes and the values of each attribute are then used in order to derive generalized rules from a given sample set. Its objective is the analysis of data with specific constraints to learn a given model, and then to compare the examples of unknown classes, refer to Han and Kamber (2000). Maszczyk and Duch, refer to Maszczyk

and Duch (2008), compared the Shannon, Renyi and Tsallis entropy in context to their use in decision trees, for different parametric values and found that Renyi entropy results good for $\alpha = 2$. In this chapter also, we have used the concept of Renyi entropy for $\alpha = 2$ in ID3 Algorithm and have applied this algorithm for classification of Census 2011 of India. The work reported in this chapter has been published in the paper entitled, **"Classification of Census Using Information Theoretic Measure Based ID3 Algorithm."** in International Journal of Mathematical Analysis, Vol. 6, no. 51, 2012, pp. 2511-2518.

3.2 Pseudocode for ID3 Algorithm

Pseudocode of ID3 Algorithm can be discussed as follows:

Algorithm: Procedure $Build\_Decision\_Tree$ from given sample data to generate a decision tree.

Input: D-the Specific data, displayed by discrete value, $attribute\_list$ - candidate property set.

The following is the pseudo code of decision tree algorithm:

Create Node N;

Procedure $Build\_Decision\_Tree()$

If D belongs to Class S then

Return N as Leaf Node and labeled by S;

If $attribute\_list$ is null then

Return N as Leaf Node and labeled by Class U in D;

For each attribute in $attribute\_list$, compute information gain G

Select the attribute which has max (G) in $attribute\_list$ as the $test\_attribute$ of N;

Set Node N as $test\_attribute$;

Set s_i as the aggregate of $test\_attribute = a_i$ in the Example D;

If s_i is null then

Add a Leaf Node, label as normal Class in D;

Else

Recursive procedure *Build_Decision_Tree* (s_i, *test_attribute*);

For more details about this algorithm, refer to Qian et al. (2007).

3.3 Algorithm of Information Gain for Renyi Entropy

Information gain is used in the algorithm of decision tree to identify appropriate attribute of every node. ID3 algorithm by the concept of information entropy theory selects the property of maximum information gain in given specific data set, which has some given classes with complete observing values, as the test property. The selected attribute makes the minimum value of information gain in the result of classification of example data, and reflects the minimum randomness, refer to Han and Kamber (2000). This information gain criteria reduces the number of steps which is needed to the classification and ensures to find a simple tree. Let us consider a specific data 'D' having 'n' number of different values with 'R_i' where i=1, 2, 3,. . . , r number of ranges for each 'P_i' property, where i=1, 2, 3,. . . , m and the data is to be divide into 'C_i' where i=1, 2, 3,. . . , c number of classes. Now, the quantity of the information required for the object data is calculated using Renyi entropy (Renyi (1961)), as follows:

$$I(C_1, C_2, C_3, \cdots, C_c) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{c} p_i^{\alpha} \ , \tag{3.3}$$

where 'p_i' represents the probability of each range of property under consideration in different classes.

Let us consider, the property 'P_i' is selected from the set. Using this property, 'D' can be divided into 'R_i' number of sets which has same number of values as that of in 'D' and entropy for the particular property is calculated by following equation based upon the total number of the values of 'D'. If 'r_i' is the number of values in the range 'R_i' of property 'P_i' then

entropy of 'P_i' is given as:

$$E(P_i) = \frac{1}{1-\alpha} \sum_{i=1}^{r} \frac{r_i}{n} \log_2 \sum_{i=1}^{c} p_i^{\alpha} \ . \tag{3.4}$$

Here, 'r' represents the number of ranges and 'r_i' number of different values, in particular property and thus the net gained information from the property 'P_i' is

$$G(P_i) = I(C_1, C_2, C_3, \ldots, C_c) - E(P_i). \tag{3.5}$$

Information acquisitions are calculated individually and accordingly entropy values are calculated. The property with the highest knowledge acquisition is selected as the root of the tree. Other properties are rearranged accordingly. The same procedure is repeated for other sub sets. Rules are formed on the basis of the final decision tree. The classification mainly deals with extraction of information of the system and its systematic development. To achieve this objective the best solution is the process in which entropy is the least.

3.4 Classification of Census 2011 of India using ID3 algorithm

In this section, we have applied the Renyi entropy based ID3 algorithm to classify the data from Census 2011 of India. For $\alpha = 2$, using the Renyi entropy measure (3.2), on the data given in table 3.1, we have some results for decision tree.

Table 3.1:: Census INDIA-2011

Sr. no.	Density	Population	Sex Ratio	Literacy	$S \setminus U$
1	< 120	< 14	< 900	> 85	U
2	$271 - 550$	> 605	> 974	< 72	S
3	< 120	< 14	$900 - 946$	< 72	S

4	$271-550$	$66-330$	$947-963$	$72-78$	S
5	$551-2170$	>605	$900-946$	<72	S
6	>2170	<14	<900	>85	U
7	$120-270$	$66-330$	>974	<72	S
8	$551-2170$	<14	<900	$72-78$	U
9	$551-2170$	<14	<900	>85	U
10	$271-550$	$14-65$	$964-974$	>85	S
11	$271-550$	$331-605$	$900-946$	$79-81$	S
12	$551-2170$	$66-330$	<900	$72-78$	S
13	$120-270$	$66-330$	$964-974$	$82-85$	S
14	$120-270$	$66-330$	<900	<72	S
15	$271-550$	$66-330$	$947-963$	<72	S
16	$271-550$	>605	$964-974$	$72-78$	S
17	$551-2170$	$331-605$	>974	>85	S
18	$551-2170$	<14	$900-946$	>85	U
19	$120-270$	>605	$900-946$	<72	S
20	$271-550$	>605	$900-946$	$82-85$	S
21	$120-270$	$14-65$	>974	$79-81$	S
22	$120-270$	$14-65$	>974	$72-78$	S
23	<120	<14	>974	>85	S
24	<120	$14-65$	$900-946$	$79-81$	S
25	>2170	$66-330$	<900	>85	U
26	$120-270$	$331-605$	>974	$72-78$	S
27	>2170	<14	>974	>85	U
28	$271-550$	$66-330$	<900	72 78	S
29	$120-270$	>605	$900-946$	<72	S
30	<120	<14	<900	$82-85$	S
31	$551-2170$	>605	>974	$79-81$	S

32	$271-550$	$14-65$	$947-963$	>85	S
33	$551-2170$	>605	$900-946$	<72	S
34	$120-270$	$66-330$	$947-963$	$79-81$	S
35	$551-2170$	>605	$947-63$	$72-78$	S

Table 3.1 contains the reformed data of Census 2011 of India taken from the source http://www.censusindia.gov.in. The complete data is divided into two parts; one as States ('S') and another is Union territory ('U') on the basis of four properties: Density, Population, Sex ratio and Literacy having different set of ranges 'R_i' as '$< 120, 120-270, 271-550, 551-2170$ and > 2170' for Density, '$< 14, 14-65, 66-330, 331-605$ and > 605' for Population, '$< 900, 900-946, 947-963, 964-974$ and > 974' for Sex ratio and '$< 72, 72-78, 79-81, 82-85$ and > 85' for Literacy. These ranges for each property has been set randomly, above and below near the center of the table.

Different properties like 'Density' is the population density of person per square kilometers, 'Population' is in Lakhs, 'Sex ratio' represents females per 1000 males and 'Literacy' is the rate of literacy in percentage, of literate for 7 years and older.

In Table 3.1, the complete data is divided into two classes i.e. c=2 having 28 number of 'S' and 7 number of 'U'. Therefore the required information is;

$$I(C_1, C_2) = \frac{1}{1-2}\left(\log_2\left\{\frac{7}{35}^2 + \frac{28}{35}^2\right\}\right)$$

Further, the information associated with the four properties using Renyi entropy for $\alpha = 2$ is given as follows:

$$
\begin{aligned}
E(P_1) = E(Density) &= -\frac{5}{35}\log_2\left\{\frac{1}{5}^2 + \frac{4}{5}^2\right\} - \frac{3}{35}\log_2\left\{\frac{3}{3}^2\right\} \\
&\quad -\frac{9}{35}\log_2\left\{\frac{9}{9}^2\right\} - \frac{9}{35}\log_2\left\{\frac{9}{9}^2\right\} \\
&\quad -\frac{9}{35}\log_2\left\{\frac{3}{9}^2 + \frac{6}{9}^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
E(P_2) = E(Population) &= -\frac{9}{35}\log_2\left\{\frac{6}{9}^2 + \frac{3}{9}^2\right\} - \frac{9}{35}\log_2\left\{\frac{9}{9}^2\right\} \\
&\quad -\frac{5}{35}\log_2\left\{\frac{5}{5}^2\right\} - \frac{3}{35}\log_2\left\{\frac{3}{3}^2\right\} \\
&\quad -\frac{9}{35}\log_2\left\{\frac{1}{9}^2 + \frac{8}{9}^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
E(P_3) = E(Sex\ ratio) &= -\frac{9}{35}\log_2\left\{\frac{5}{9}^2 + \frac{4}{9}^2\right\} - \frac{9}{35}\log_2\left\{\frac{1}{9}^2 + \frac{8}{9}^2\right\} \\
&\quad -\frac{9}{35}\log_2\left\{\frac{1}{9}^2 + \frac{8}{9}^2\right\} - \frac{5}{35}\log_2\left\{\frac{5}{5}^2\right\} \\
&\quad -\frac{3}{35}\log_2\left\{\frac{3}{3}^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
E(P_4) = E(Literacy) &= -\frac{9}{35}\log_2\left\{\frac{9}{9}^2\right\} - \frac{10}{35}\log_2\left\{\frac{6}{9}^2 + \frac{4}{9}^2\right\} \\
&\quad -\frac{8}{35}\log_2\left\{\frac{1}{8}^2 + \frac{7}{8}^2\right\} - \frac{5}{35}\log_2\left\{\frac{5}{5}^2\right\} \\
&\quad -\frac{3}{35}\log_2\left\{\frac{3}{3}^2\right\}
\end{aligned}
$$

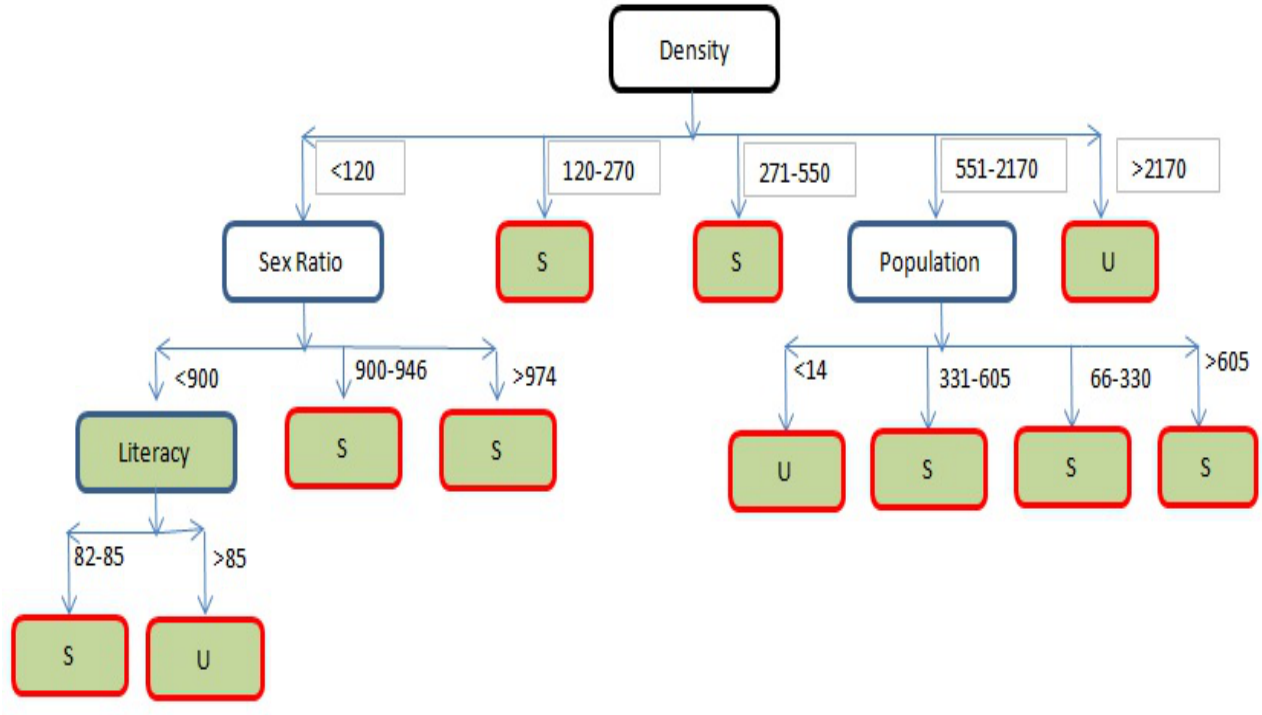

Figure 3.1: Decision Tree using Renyi Entropy

Net gained information;

$$\begin{aligned} Gain(Density) &= I(C_1, C_2) - E(P_1) = 0.258852\ bits. \\ Gain(Population) &= I(C_1, C_2) - E(P_2) = 0.256699\ bits. \\ Gain(Sexratio) &= I(C_1, C_2) - E(P_3) = 0.140526\ bits. \\ Gain(Literacy) &= I(C_1, C_2) - E(P_4) = 0.205441\ bits. \end{aligned}$$

As we can notice that gain information of 'Density' is the largest and hence is the root of the tree. By repeating above process again for the different sub-trees, associated calculations are given in *Appendix A*, we get the decision tree as a result, given in Fig. 3.1.

3.4.1 Rules

On the basis of decision tree some conditions can be set or rules can be formed using 'if-then' so that the correct decision can be made during implementation of different policies or projects of development. There are 11 rules out of which only 3 rules are classifying the data in favor of Union Territories (i.e. U) and rest are in favor of State (i.e. S). These rules can be described as follows:

1. If Density is '< 120', Sex Ratio is '< 900' and Literacy is '$82 - 85$' then 'S'; It means that for any value of Population in case of small Density and Sex ratio with Literacy between 82 and 85 percent, the policy must be implemented in 'State'.

2. If Density is '< 120', Sex Ratio is '< 900' and Literacy is '> 85' then 'U'; It means that for any value of Population in case of small Density and Sex ratio with Literacy between more than 85 percent, the policy must be implemented in 'Union Territories'.

3. If Density is '< 120', Sex Ratio is '$900 - 946$' then 'S'; It means that for any value of Population and Literacy in case of small Density and Sex ratio within 900 to 946 per 1000 males, the policy must be implemented in 'State'.

4. If Density is '< 120' and Sex Ratio is '> 974' then 'S'; It means that for any value of Population and Literacy in case of small Density and high Sex ratio, the policy must be implemented in 'State'.

5. If Density is '$120 - 270$' then 'S'; It means that for any value of Population, Sex ratio and Literacy in case of Density within 120 to 270, the policy must be implemented in 'State'.

6. If Density is '$271 - 550$' then 'S'; It means that for any value of Population, Sex ratio and Literacy in case of Density within 271 to 550, the policy must be implemented in 'State'.

7. If 'Density' is '$551-2170$', Population is '< 14' then 'U'; It means for any value of Sex ratio and Literacy in case of Density within 551 to 2170 and low Population, the policy must be implemented in 'Union Territories'.

8. If 'Density' is '$551-2170$', Population is '$66-330$' then 'S'; It means that for any value of Sex ratio and Literacy in case of Density within 551 to 2170 and Population within 66 to 330, the policy must be implemented in 'State'.

9. If 'Density' is '$551-2170$', Population is '$331-605$' then 'S'; It means that for any value of Sex ratio and Literacy in case of Density within 551 to 2170 and Population within 331 to 605, the policy must be implemented in 'State'.

10. If 'Density' is '$551-2170$', Population is '> 605' then 'S'; It means that for any value of Sex ratio and Literacy in case of Density within 551 to 2170 and high Population, the policy must be implemented in 'State'.

11. If 'Density' is '> 2170' then 'U'; It means that for any value of Population, Sex ratio and Literacy in case of high Density, the policy must be implemented in 'State'.

3.5 Conclusion

More output with minimum input is expected from every policy or project of development. In these types of situations, decision tree provides different modes of classification and ensures to find a right decision. By classifying a data purposely, we find some commercial valuable and potential information. We have applied Renyi entropy for α=2 in ID3 algorithm to develop a decision tree, which can help in following the concept of right policy for right people.