Chapter 1

Introduction

"Research is to see what everybody else has seen, and to think what nobody else has thought."

-Albert Szent-Gyorgyi

(1893-1986, Hungarian Biochemist)

1.1 Introduction and Literature survey

Information is basically the specific knowledge, refer to Thomas and Cover (2009) and Reza (1994), it is new to the receiver and uses symbols or letters for communication. Moreover, it is useful only if is best interpreted at the receiver's end. The information may be right or wrong and is generally interrupted by noise. Hence, accuracy in communication can be measured by the amount of information communicated. In the communication of generalized form of information, known as message, communication system plays a vital role. The system processes a message generated by the source and transmits it in encoded form through the channel to decoder and then to the destination. But, it is not completely reliable process due to bad effect of noise, which can decrease the amount of information to be delivered. Also, as per the findings of Shannon (1948) the fundamental problem of communication is of reproducing at one point either exactly or

approximately, a message selected at another point.

To handle the problems of communication, information theory came into existence. In telecommunication and computer networking it is a physical transmission medium e.g. wire, but in information theory is refers to a theoretical channel model with certain error characteristics. It is a mathematical approach, with probabilistic base, which concretize the concept of information. In a broader view, information theory handles the following problems concerning to any system:

Information Processing: The processing of information in any manner which is detectable and interpretable by an observer, in context to any problem or situation.

Information Storage: A systematic process of collection and cataloging data to smooth the process of extraction.

Information Retrieval: Area of study concerned with the extraction of information at destination.

Decision Making: The process of making a logical choice among all the available choices, along with the analysis of different methods and their results.

While in precise view, it deals with all theoretical problems connected with the transmission of information over communication channels and also studies the uncertainty measures, their generalizations, development as well as analysis of practical and economical methods of coding information for better communication.

The development in literature of communication dates back to 20th century with the work initiated by Nyquist (1924), who was the first to investigate the statistical nature of communication. Later Nyquist (1924) and Hartley (1928) recognized the logarithmic nature of the natural measure of information and introduced the entropy of a distribution of equally probable events. In 1948, a great American mathematician, electrical engineer and computer scientist, Claude E. Shannon (1948) published a landmark paper 'The Mathematical Theory of Communication' in the 'Bell System

Technical Journal' which laid the foundation of the modern day's information theory. Being an electrical engineer, his aim was to maximize the line capacity with minimum distortion. In Shannon model, refer to Shannon (1948), a randomly generated message, produced by a source of information, is transmitted in encoded form and is decoded at destination. He introduced a measure of information (or entropy) for a general finite complete probability distribution and its characterization theorem. Entropy, as defined by Shannon and added upon by other physicists, is closely related to thermodynamical entropy which represents the measure of randomness and the amount of information a message contains is measured by the extent it combats entropy. It has been observed that the less predictable the message carries more information.

The second half of the 20th century was characterized by the tremendous development of communication systems in which the coded information is transmitted in digital form. By this coding, the real nature of the information signal becomes secondary and as a result the same system becomes capable of transmitting simultaneously signals for very different nature e.g. data, audio, video etc. This development has been made possible by the use of more powerful integrated circuits. Although it is mainly during the last 30 years that the truly operational digital systems have been developed, the theoretical foundations for all these developments date back to the work of Shannon and others in the mid of the 20th century which led to the development of information theory as a field of mathematics.

Information Theory basically moves around following three fundamental questions:

Compression: How much data can be compressed so that another person can recover an identical copy of the uncompressed data?

Lossy Data Compression: How much data can be compressed so that another person can recover an approximate copy of the uncompressed data?

Channel Capacity: How quickly reliable communication is possible from the source to destination through a noisy medium?

The theory is concerned with the mathematical laws governing systems designed to communicate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store and process information.

1.2 Areas of Intersection

Various concepts and properties of information theory overlaps with the theories of some other fields of studies, like in computer science Kolmogorov (1968) and Solmonoff (1964) resembles the concept of entropy with the concept of Kolmogorov Complexity, as both aim to provide means for measuring information in bits, Jaynes (1983) and Ellis (2012) described that the notion of Shannon is the generalized form of the notion of entropy that first exist in thermodynamics physics, the information and entropy can be also be considered as measures of uncertainty of probability distribution, thus it relates with probability and statistics also, refer to Lazo and Rathie (1978), moreover the concept of entropy is also helpful in predicting the future behavior/trends of customer/market which are similar to the aim of economics and philosophy, refer to Karlin (1992), Schnorr (1977) and Chaitin (1966).

Zhu and Wen (2010) founds that in past sixty years, the literature on information theory has grown quite voluminous and apart from its applications in communication theory it is also being used for the purpose of information retrieval and purposeful gathering. Also, it has found deep applications in many fields and remains an active area of research. In the information age, it provides insight and direction for the design and analysis of communication as well as computer systems, refer to Wen and Zhu (2012).

1.3 Shannon Entropy

Claude Elwood Shannon (1948), being an electrical engineer was working on the goal to get maximum line capacity with minimum distortion and as a result of his efforts the base of information well known as *Information Theory*, came into existence. Complete Shannon information theory is about measurement of 'information', which exists if there is some prior uncertainty and also information gain measured from an experiment/observation is equal to the amount by which the uncertainty has been reduced.

Shannon (1948) conceived the statistical nature of the communication signal with that of the random variable $X = (X_1, X_2, X_3, ..., X_n)$ having probability distribution $P = (p_1, p_2, p_3, ..., p_n)$, where $p_i = Pr\{X = X_i\}$ and introduced the **measure of information (or uncertainty)** as

$$H(X) = -\Sigma p_i log p_i, \tag{1.1}$$

such that $0 \leq p_i \leq 1$ and $\Sigma_{i=1}^{n} p_i = 1$, associated with this experiment. This is also called the **Shannon's measure of entropy**. Here the base of the logarithm, unless it is mentioned otherwise, is generally taken as 2 and the units are then is in bits, a short form of the binary digit. Also, we define $0log0 = 0$.

1.3.1 Properties of the Uncertainty Function

a) **Non-negativity:** $H(P)$ is always non-negative, that is,

$$H(P) = -\Sigma p_i log p_i \geq 0.$$

b) **Maxima:** $H(p_1, p_2, ..., p_n) \leq log\ n$ with equality when $p_i = \frac{1}{n}$ for all i.

c) **Continuity:** $H(p_1, p_2, ..., p_n)$ is continuous function of p_i's.

d) **Symmetry:** $H(p_1, p_2, ..., p_n)$ is a symmetric function of p_i's, that is, it remains invariant with respect to the order of the outcomes.

e) **Additivity:** If $P = \{p_1, p_2, ..., p_n\}$ and $Q = \{q_1, q_2, ..., q_n\}$ are two independent probability distributions, then we have, $H(P*Q) = H(P)+H(Q)$, where $P*Q$ denotes the joint probability distributions of P and Q.

A few more properties have also been given for this measure as follows, for these properties and proofs we can refer to Aczel and Daroczy (1975) and Mathai and Rathie (1975):

1) *The entropy value doesn't change on addition of an impossible event.*

2) *On maximizing Shannon entropy, under some linear constraints, resultant probabilities are non-negative.*

3) *It is always positive.*

4) *It is minimum if p_i's represents a degenerate distribution.*

5) *Being a concave function of p_i's, local maximum is the global maximum.*

6) *For uniform distribution, its value is maximum.*

7) *Strong additivity: The joint entropy of two not compulsory independent probability distributions is the entropy of the first distribution added to the expected value of the conditional entropy of the second distribution.*

8) *Sub additivity: The joint entropy of two not compulsory independent distributions is less than or equal to the sum of the entropies of the two distributions.*

In case, X is a continuous random variable, say denoting the lifetime of a 'unit' with probability density function $f(x)$ then the measure of Shannon uncertainty associated with X is given by

$$H(X) = -\int_0^\infty f(x) log f(x) dx. \tag{1.2}$$

This is a measure of uncertainty of the lifetime of a unit, also known as differential entropy, refer to McEliece (2002). Also, Rao et al. (2004) and Wang and Vemuri (2005) have studied various properties and applications of the measure (1.2).

1.4 Characterizations and Generalizations of Shannon Entropy

Ever since introduction, based on the different set of properties of Shannon measure (1.1), it has been characterized by many researchers like Kinchin (1957), Aczel and Daroczy (1975) etc. Further, on considering various values of parameters it has also been generalized such that these tends to Shannon entropy for limiting cases of the parameters. More details can be found in Renyi (1961), Havrda and Charvat (1967), Varma (1966), Arimoto (1971), Ferreri (1980), Sharma and Mittal (1975) etc. Next, we discus the details of a few characterizations and generalizations of Shannon entropy:

1.4.1 Renyi Entropy

An generalization of the measure (1.1) was given by Renyi as

$$H_\alpha(X) = 1/(1-\alpha) log[\Sigma_{i=1}^n p_i^\alpha]; \alpha \neq 1 \; and \; \alpha > 0. \tag{1.3}$$

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).

1.4.2 Havrda and Charvat's Information Measure

A non-additive generalization of the measure (1.1) is given by Havrda and Charvat (1967)

$$H^\alpha(X) = \frac{1}{\alpha - 1}[\Sigma_{i=1}^n (p_i^\alpha) - 1]; \alpha \neq 1 \; and \; \alpha > 0. \tag{1.4}$$

This measure satisfies the non-additivity property

$$H(X * Y) = H(X) + H(Y) + (\alpha - 1)H(X)H(Y).$$

It is more general than Shannon entropy as well as simpler than Renyi entropy. When $\alpha \to 1$, both measure (1.3) and (1.4) reduce to (1.1).

1.4.3 Cumulative Residual Entropy

Rao et al. (2004) introduced an alternative measure of uncertainty called cumulative residual entropy, of a random variable X, defined as

$$H(X) = -\Sigma P(X > x)\, logP(X > x). \tag{1.5}$$

This measure is based on cumulative distribution function rather than probability density, and is thus, in general more stable, since the distribution function is more regular because it is defined in an integral form unlike the density function, which is defined as the derivative of the distribution.

1.4.4 Varma Entropy

Varma (1966) introduced another two parameter generalized entropy of order α and type β defined by

$$H^{(\alpha,\beta)}(X) = \frac{1}{\beta - \alpha} \log(\sum_{i=1}^{n} p_i^{\alpha+\beta-1}),\ (\beta - 1) < \alpha < \beta\ and\ \beta \geq 1. \tag{1.6}$$

We get Renyi entropy (1.3) for β=1 in (1.6), and in addition, if also $\alpha \to$1, then $H^{(\alpha,\beta)}(X)$ reduces to Shannon Entropy (1.1).

1.4.5 Relative Information Measure

Kullback and Leibler (1951) gave the statistical view for information and called it the discrimination function (other authors also called it cross entropy, relative information etc.). They considered a random experiment with two associated probability distributions. Let $P = (p_1, p_2, ..., p_n)$ and

$Q = (q_1, q_2, ..., q_n)$ be the actual and predicted probability distributions respectively, associated with the outcomes $X = (X_1, X_2, ..., X_n)$ of a random experiment then relative information measure given by Kullback and Leibler (1951) is

$$H(P/Q) = \sum_{i=1}^{n} p_i log \frac{p_i}{q_i}. \tag{1.7}$$

such that $p_i \geq 0, q_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1 = \sum_{i=1}^{n} q_i$. It quantities discrimination between two populations and also whenever $q_i = 0$, the corresponding p_i is also zero and $0log\frac{0}{0} = 0$. It also satisfies a number of properties like non-negativity, additivity etc. For details refer to Kullback and Leibler (1951).

1.4.6 Inaccuracy Measure

As a generalization of Shannon entropy, Kerridge (1961) introduced the notation of inaccuracy that can take account of the errors that occurs due to insufficient data and wrong specification of the model in deciding about the probability of happening of different events in an experiment and is defined as

$$H(P; Q) = -\sum_{i=1}^{n} p_i log\ q_i. \tag{1.8}$$

where $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ are the actual and predicted probability distributions respectively, associated with the outcomes $X = (X_1, X_2, ..., X_n)$. It reduces to the Shannon entropy if $p_i = q_i$ for all values of i. The measures of information, relative information and inaccuracy is given by

$$H(P; Q) = H(P) + H(P/Q). \tag{1.9}$$

Thus, inaccuracy can also be measured as the sum of information and relative information.

An event, in a particular experiment with reference to a specific context may play more important role as compare to other events. In these types of situations, more weightage should be given to such events. With this point

of view, next we discuss weighted information measure to take account of such aspect.

1.5 Weighted Information Measures

If $P = (p_1, p_2, ..., p_n)$, $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$ be the probability distribution associated with an experiment $X = (X_1, X_2, ..., X_n)$ and $U = (u_1, u_2, ..., u_n)$, where $u_i > 0$ is the weight associated with the outcome X, which is equivalent to its importance to the experimenter's context Belis and Guiasu (1968) introduced a weighted information measure, well known as 'useful' information measure, given by

$$H(P;U) = -\sum_{i=1}^{n} u_i p_i log\ p_i, \tag{1.10}$$

where $u_i \geq 0$, is the weight assigned to the i^{th} outcome X_i, $i = 1, 2, ..., n$. This measure reduces to (1.1) for equal u_i's. It is also known as the quantitative and qualitative measure of information. A quantitative and qualitative measure of relative information, introduced by Taneja and Tuteja (1984), is given by

$$H(P/Q;U) = \sum_{i=1}^{n} u_i p_i log\frac{p_i}{q_i}, \tag{1.11}$$

where the weight u_i is associated to X_i and is independent of probability p_i or q_i. The measures (1.11) reduces to the Kullback measure (1.7) when all the weight $u_i's$ are equal.

Generalization of the measures (1.10) and (1.11) and their applications to coding theory have been studied extensively by many authors including Aggarwal and Picard (1978), Gurdia and Pessoa (1977), Hooda and Tuteja (1981), Taneja Hooda and Tuteja (1985) etc.

Information theoretic measures in lifetime distributions also form an important part of this research work, so next we will consider some basic concepts in reliability.

1.6 Survival Function

There exists various functions like survival function, hazard rate function, and mean residual life function which can completely specify the distribution function of the lifetime of a unit. Information associated with these functions differ and behave differently under different situations. The survival function is also known as the reliability function and represents the probability of performance of a system without failure for a given time period under given conditions. Let us consider a non-negative continuous random variable X which represents the lifetime of a unit or system, with distribution function $F(x)$, then the survival function is given by

$$\overline{F}(x) = Pr(X > x) = \int_x^\infty f(x)dx, \tag{1.12}$$

where $f(x)$ is the probability density function of X and also $\overline{F}(x) = 1 - F(x)$. Further, It is decreasing function of x satisfying $\overline{F}(0) = 1$ and $lim_{x\to\infty}\overline{F}(x) = 0$.

1.7 Mean Residual Life Function

Another main component in survival analysis is the mean residual life of a system or a component, which provides information about the survival age of the component and further it helps in the improvement of the average life time of the system. Let us consider a continuous random variable X with $E(X) < \infty$, then mean residual life function is given by

$$\delta(t) = E[X - t|X > t] = \frac{\int_t^\infty F(x)dx}{\overline{F}(t)}. \tag{1.13}$$

The information obtained from this measure can improve the maintenance policies. The survival function can be represented as a function of

the mean residual life, as

$$\overline{F}(t) = \frac{\delta(0)}{\delta(t)} \exp\left[-\int_0^t \frac{dx}{\delta(x)} \right]. \tag{1.14}$$

For more properties and applications of the mean residual life function, refer to Whittle and Tennakoon (2005), Asadi and Baryamoglu (2005) etc.

1.8 Length Biased Model

The concept of weighted distribution introduced by Rao (1965), is widely used in statistics and other applications. These distributions are the result of the weighted observations of a stochastic model. Let X be a non-negative continuous random variable with probability density function $f(x)$, and let X_w be a weighted random variable corresponding to X with weight function $w(x)$, which is positive for all values of $x \geq 0$. Then the probability density function $f_w(x)$ of the weighted random variable X_w is given by

$$f_w(x) = \frac{w(x)f(x)}{E(w(X))}, 0 \leq x < \infty, \tag{1.15}$$

with $0 < E[w(X)] < \infty$. Obviously $f_w(x) \geq 0$ and $\int_0^\infty f_w(x)dx = 1$. X_w is said to be a length biased random variable if $w(x) = x$, and its p.d.f. based on equation (1.15) becomes

$$f_w(x) = \frac{xf(x)}{E(X)}. \tag{1.16}$$

Length-biased sampling occurs in different fields in the absence of proper sampling criteria. As a solution, sampling is proportional to the length of quantities, in other words large quantities are sampled with higher probabilities. We have detailed and applied the concept of length biased random variable to inaccuracy measures in Chapter 6.

1.9 Residual Entropy

It is a type of dynamic information theoretic measures that arise for the left truncated data. If X is the lifetime distribution of a component, then the residual lifetime of the component which has survived for time t is represented by the random variable $[X - t|X > t]$. In this direction, Ebrahimi (1996) proposed a dynamic measure of entropy based on Shannon entropy known as residual entropy, given by

$$H(X;t) = -\int_t^\infty \frac{f(x)}{\overline{F}(t)} log(\frac{f(x)}{\overline{F}(t)})dx, \tag{1.17}$$

where $f(x)$ denotes the probability density function of the random variable $[X - t|X > t]$. The concept and generalization of residual entropy has been detailed, in Chapter 6.

According to Ash (1990), information theory is extensively studied subject since the introduction of Shannon entropy. Due to the simplicity of information theoretic measures, it gives researcher a unique possibility to explore this subject in depth. Zhu and Wen (2010) found it as the basic study of the amount of information stored in data for communication. This study also gives us valuable information about the identification of more theoretic measures and their generalization with suitable applications. As a result, new possibilities have also emerged in different areas of applications. Next, we have outlined some areas of application.

1.10 Applications of Shannon Entropy

The Shannon's measure of entropy and its generalizations, finds applications in coding theory, marketing and in almost all other social and physical sciences like: Genetic Algorithms (refer to Yang et al. (2001)), Finance (refer to Zhou et al. (2013), Ou (2005) and Gulko (1997) and Buchen and Kelly (1996)), Industrial Engineering (refer to Zamiri (2013) and Kullback and Leibler (1951)), Stochastic Processes (refer to Haken (2003)

and Karmeshu (2003)), Non-random Functions and Complex Fractals (refer to Jumarie (2003)), Data Mining (refer to Yao (2003)), Queuing System and Networks (refer to Kouvatsos (2003)), Artificial Societies (refer to Tang and Mao (2014) and Tseng and Tuszynski (2014)) and in Biological systems (refer to Baez and Pollard (2015)) etc.

An active area of current research, in the application of entropy, is in data mining, refer to Chen Han and Yu (1996), which is the process of extracting some potential but valuable information from massive, noiseful, incomplete, fuzzy or random data. In the algorithms of data mining, Gondek and Hofmann (2007) and Zhang et al. (2006) found the classification as an important aspect. Tadros et al. (2011) and Wang and Vemuri (2005) explained classification as the process of finding the common feature of data object belonging to same class and its aim, via analyzing the training set to learn a classified model, which can be used to compare the examples of unknown classes. Among algorithms of classification, decision tree is a common feature to build predictive model. By classifying a host of data purposely, we find some valuable and potential information from commercial point view. In the algorithms of decision tree, Zhu and Wen (2010) found information gain as the key of identifying appropriate property of every node. Some algorithms have employed the concept of Shannon entropy for classification of data. Keeping in view the role of entropy in data mining, in next section we discuss basics of data mining.

1.11 Introduction to Data Mining

Generally, solution to most of the real world problems are based on the earlier data available under some constraints. These solutions are based upon the pattern discovery of data set which is expected to help in prediction of future behavior of the problem under consideration. This have helped in answering various business questions, which earlier were time consuming. Moreover, data mining concepts can easily be applied on soft as well as

hardware platforms, which increases its credibility and accuracy.

In pattern discovery for best results, it would be beneficial to extract all the possible and essential structures without modeling any noisy data. To achieve this objective, information from all the parameters should be maximized. According to Brand (1999), one of the best possible solution for this objective is entropy optimization in which entropy minimization maximizes the role of each parameter. In all, according to Chaovaliwongse (2003), data mining can be considered as an process of extraction of knowledge from the base of data.

Manual process of extracting patterns from the data is known for centuries. With the increase in data ware houses, demand for new techniques also increases with the imminent need for turning such data into useful information and knowledge. Among old methods of pattern discovery, Baye's Theorem and Regression Analysis are well known methods. Further, some discoveries in the field of computer science like Clustering, Genetic algorithms, Decision trees etc. also supported in uncovering hidden patterns. These methods have been used in different fields by governments, businessmen etc. to develop market reports or to analyze Census.

Methods of data mining techniques, generally are of two types, first type of methods are *predictive methods*, used to predict future behavior of some parameters and these methods form the base of present research work. While, second type of methods are *descriptive methods*, used to describe the data in understandable manner.

Many authors like Berry and Linoff (2004) have described the role of data mining as per following six tasks:

1.11.1 Classification, Estimation and Prediction

Classification is one of the example of direct mining which deals with discrete outcome like: yes or no; class 1 or class 2 etc. and helps in easy communication to increase understanding with the world. In some examples of classification, we observe, dogs in breeds, people in races, exam scores

in grades, high or low risk credit applicants, spotting fraudulent insurance claims etc. Classification is characterized by well-defined set of classes and a set of pre defined examples, on these basis it develops a model that can be applied to classify an undetermined data. Some well known methods of classification are Decision Trees, Nearest Neighbor Techniques, Neural networks and link analysis.

Next, estimation has continuously valued outcomes obtained from unknown continuous variables like income, height etc. It is used to perform a classification task, for example as if some gift coupons are to be given to some potential customer then the same can be decided upon there earlier continuous record like shopping record, shopping limit, estimating the number of persons in family, family income etc. Regression models, Survival Analysis and Neural Networks are some methods used for estimation.

Apart from classification and estimation, in the task of prediction, records are classified as per the requirement of particular domain of the data under observation. Thus, it differs from classification and estimation in terms of the input variables. Predicting customers will leave in next few months, subscriber will add on particular channels, students will join computer science etc. are some examples of this task. Most of the data mining techniques are capable of prediction. The choice of any technique applied will depend upon the nature of input variables and importance of the prediction.

1.11.2 Clustering and Association Rules

In clustering the heterogeneous data is divided into more homogeneous subgroups known as *clusters*. Unlike classification, it does not has predefined set of classes and examples. Data provided is segmented as per similarity between properties associated with each value. It can be considered as initial step for data mining. Self organizing maps, is one of the technique of clustering.

In super markets a lot of packages or offers are given to the potential customers, but to decide about the contents of the package the role of

association rules come into action. Sometimes it is to be checked if some one buys toothbrush, will they also buy toothpaste, in the same way some more items can be associated to make association rules so that effective offers or packages can be prepared or the things can be arranged in shelves of departmental store.

1.11.3 Profiling

The description of data is the task of profiling, which helps in deciding which part of the data is important for the desired results. The description of winning of a political party in an election will draw the attention towards particular section of society about their decision, is one of the example of profiling. Decision trees are one of the powerful tool for profiling. Even techniques of clustering and association rules can be used to make profiles.

In order to improve the quality of results, some standards have been defined for data mining, like the 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0), refer to Chapman et al. (2016), and the 2004 Java Data Mining Standard (JDM 1.0), refer to Java (2004). Latest versions of these standards are still in under progress. Beside these, there are open-source software systems like RapidMiner, Weka, R etc. to speed up the process of data mining, these are explained in chapter 2. Moreover, an independent group known as, Data Mining Group, of various data companies, has released an XML based language Predictive Model Markup Language of version 4.2 (PMML 4.2) in February, 2014, refer to PMML (2014), which represents the data in a sharable standard way and can import and export data from most of the data mining softwares.

1.12 Applications and Current Issues

Although applications of data mining are not limited to certain fields only, with the increase in the number of data warehouses and competitive markets with demand of high quality and accuracy, the areas of application

of data mining is continuously increasing. Some fields of application are; Health Care Management, Business , Financial Data Analysis, Telecommunication Industry, for more details refer to Han Kamber and Pei (2011), Apte et al. (2002), Han and Kamber (2000) etc.

Al-Attar (1998) explained three main issues during data mining for any organization as: *First* is Methodology, which includes analysis of problem, data preparation/exploration, Pattern generation and its development and monitoring, while *second* is easy understanding and application of tools to be applied and their support to all steps of data methodology in pattern discovery and reporting, and *third* issue is performance and scalability with due availability of large amount of data.

1.13 Motivation and Organization of the Thesis

Based upon the above discussion and literature review, we were motivated to apply the concept of entropy in finding the solution of data mining and in life time distributions. Thus, the thesis is organized into two parts, in first part, due to big role of classification in data mining, we found interest in studying the generalization and characterization of the information theoretic measures and then their applications in classification algorithms. Considering different types of data, we found it worthwhile to study the effect of various information measures in decision tree induced algorithms and their comparison, while in second part we have studied weighted generalized two parametric residual information measure and its characterization theorem. Thesis comprises seven chapters including the current chapter on introduction and literature survey and a bibliography. The chapters have been organized as follows:

Chapter 2 give the details of the methodology used in present work. As classification play important role in data mining to draw conclusions, we

have discussed different types of classification algorithms along with various decision tree induction algorithms. After a brief discussion on software environments, that are in use now a days to develop decision trees, methodology for this present work, is outlined.

Chapter 3 deals with the classification of census using Renyi's entropy based ID3 Algorithm. As in the algorithms of data mining, the classification is an essential step, using an information theoretic measure in ID3 algorithm, one of the key algorithms of decision tree algorithms, we have discussed the different steps of the development of decision tree so that the best classification criteria can be developed which is helpful in making good decisions. From the data under consideration having a set of values, a property with maximum information gain is selected as the root of the tree and the process is repeated to develop complete decision tree. This method is applied to the data, a part of Census-2011 of India, to get some values worth in improving or implementing a policy with the view of right policy for right people. The work reported in this chapter has been published as research article 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.

In **Chapter 4**, we have modified ID3 algorithm and used generalized entropy of order α and type β to derive some more refined rules from the resultant decision tree in context with the analysis of the health conditions in the States and Union Territories of India. The work reported in this chapter has been published as research article entitled, **"Analysis of Health Conditions Using Generalized Information Measure Based ID3 Algorithm"**, in Proceedings of 4th Annual International Conference on Operations Research and Statistics (ORS-2016), pp. 33-37, held at Hotel Fort Canning, Singapore from 18th to 19th January, 2016.

A decision tree induced algorithm is applied in classification of data to develop a set of rules for classification in **Chapter 5**. To achieve this objective we have applied C4.5 algorithm to develop decision tree and have compared it with decision tree of modified ID3 algorithm based upon entropy introduced by Varma (1966). We get a set of rules that will be helpful in analysis of health services in particular region.

In **Chapter 6**, We have studied weighted generalization of information theoretic measures. Considering a particular condition, when a system has survived for some units of time, generalized residual entropy plays an important role in the field of information theory. In this chapter, we have proposed the concept of weighted generalized residual entropy of order α and type β, and shown that the proposed measure characterizes the distribution function uniquely. The work reported in this chapter has been published as research article entitled, **"On Weighted Generalized Residual Information Measure"**, in Mathematical Journal of Interdisciplinary Sciences, Vol. 4, No. 1, Sept. 2015, pp. 1-14.

Finally, **Chapter 7**, presents the summary and outlook of the work reported in the thesis with future scope. In the end, we have summarized with some calculations of chapters in appendices and bibliography of the thesis.