

Prediction of Missing Associations Using Rough Computing and Bayesian Classification

D. P. Acharjya, Debasrita Roy, Md. A. Rahaman

School of Computing Science & Engineering, VIT University, Vellore, TamilNadu, India

Email:dpacharjya@gmail.com; roy.debasrita1@gmail.com; a.rahaman89@gmail.com

Abstract— Information technology revolution has brought a radical change in the way data are collected or generated for ease of decision making. It is generally observed that the data has not been consistently collected. The huge amount of data has no relevance unless it provides certain useful information. Only by unlocking the hidden data we can not use it to gain insight into customers, markets, and even to setup a new business. Therefore, the absence of associations in the attribute values may have information to predict the decision for our own business or to setup a new business. Based on decision theory, in the past many mathematical models such as naïve Bayes structure, human composed network structure, Bayesian network modeling etc. were developed. But, many such models have failed to include important aspects of classification. Therefore, an effort has been made to process inconsistencies in data being considered by Pawlak with the introduction of rough set theory. In this paper, we use two processes such as pre process and post process to predict the output values for the missing associations in the attribute values. In pre process we use rough computing, whereas in post process we use Bayesian classification to explore the output value for the missing associations and to get better knowledge affecting the decision making.

Index Terms— Rough Set, Order Relation, Almost Indiscernibility, Fuzzy Proximity Relation, Missing Data, Bayesian Classification

I. Introduction

The amount of data collected across a wide variety of fields today far exceeds our ability to reduce and analyze without the use of automated analysis techniques. There is much information hidden in the accumulated voluminous data. It is very hard to obtain this information. So, it is essential for a new generation of computational theories and tools to assist human in extracting knowledge from the rapidly growing voluminous digital data. Knowledge discovery in databases (KDD) is the field that has evolved into an important and active area of research because of theoretical challenges associated with the problem of discovering intelligent solutions for huge data.

Knowledge discovery and data mining are the two rapidly growing interdisciplinary fields which merge database management, probability theory, statistics, computational intelligence and related areas. The basic aim of all these is to extract useful knowledge and information from voluminous data.

The process of knowledge discovery in databases and information retrieval appear deceptively simple from the perspective view of the terminological definition [1]. Knowledge discovery in databases is defined as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. It tells that knowledge discovery process consists of several stages: data selection, cleaning of data, enrichment of data, coding, data mining and reporting. However, the data mining phase has become one of the most popular areas of recent research. The closely related process of information retrieval and data mining is defined [2] as “the methods and processes for searching relevant information out of information systems that contain extremely large numbers of documents”. In execution, however, these processes are not simple at all, especially when executed to satisfy specific personal or organizational knowledge management requirements.

The earliest and the most successful technique being used in data mining is the notion of fuzzy sets by Zadeh [3] that captures impreciseness whereas rough sets of Z. Pawlak [4] is another attempt that captures indiscernibility among objects to model imperfect knowledge [5, 6, 7]. There were many other advanced methods such as rough set with similarity, fuzzy rough set, rough set on fuzzy approximation spaces, rough set on intuitionistic fuzzy approximation spaces, dynamic rough set, were discussed by different authors to extract knowledge from the huge amount of data [8,9,10,11,12].

Missing data is a common problem in knowledge discovery, data mining and statistical inference. Several approaches to missing data have been used in developing trained decision systems. Little and Rubin [13, 14] have studied and categorized missing data into three types: missing completely at random, missing at random, and not missing at random. The easiest way to handle missing values is to discard the cases with missing values and do the analysis based only on the complete data. But, the absence of missing associations

among attribute values and missing data may have information value to predict the decision.

In this paper, we use two processes such as pre process and post process to predict the decision values for the missing associations in the attribute values. In pre process we use rough set on fuzzy approximation spaces with ordering rules to find the suitable classification of data set, whereas in post process we use Bayesian classification to explore decision values for the missing associations in the attribute values. Rest of the paper is organized as follows: Section 2 presents the basics of rough set on fuzzy approximation space. We present the basic idea of order information system in Section 3. In Section 4, we discuss the Bayesian classification whereas in Section 5 we propose prediction model. In Section 6, an empirical study on cosmetics companies were considered to analyze our proposed model. This is further followed by a conclusion in Section 7.

II. Rough Set on Fuzzy Approximation Space

Convergence of information and communication technologies brought a radical change in the field of decision making. It is a well established fact that right decision at right time provides an advantage in decision making. But the real challenge lies in converting huge data collected across various domains into knowledge, and to use this knowledge to make informed business decisions. Classical set has been studied and extended in many directions so as to model business decisions. Later the notion of fuzzy set by Zadeh [3], its generalizations and the notion of rough set by Pawlak [4] were the major research in this direction. The rough set philosophy is based on the concept of indiscernibility relation. The basic idea of rough set is based upon the approximation of sets by pair of sets known as lower approximation and upper approximation with respect to some imprecise information. However, indiscernibility relations in real life situations are relatively rare in practice. Therefore, efforts have been made to make the relations less significant by removing one or more requirements of an indiscernibility relation. A fuzzy relation is an extension of the concept of a relation on any set U . Fuzzy proximity relations on a universal set U are much more general and abundant than equivalence relations. The concept of fuzzy approximation space which depends upon a fuzzy proximity relation defined on a universal set U is a generalization of the concept of knowledge base. Thus, rough sets defined on fuzzy approximation spaces extend the concept of rough sets on knowledge bases as discussed by Acharjya and Tripathy [9].

Let U be a universe. We define a fuzzy relation on U as a fuzzy subset of $(U \times U)$. A fuzzy relation R on U is said to be a fuzzy proximity relation if $\mu_R(x, x) = 1$ for all $x \in U$ and $\mu_R(x, y) = \mu_R(y, x)$ for $x, y \in U$. Let R

is a fuzzy proximity relation on U . Then for a given $\alpha \in [0, 1]$, we say that two elements x and y are α -similar with respect to R if $\mu_R(x, y) \geq \alpha$ and we write $xR_\alpha y$ or $(x, y) \in R_\alpha$. We say that two elements x and y are α -identical with respect to R if either x is α -similar to y or x is transitively α -similar to y with respect to R , i.e., there exists a sequence of elements $u_1, u_2, u_3, \dots, u_n$ in U such that $xR_\alpha u_1, u_1R_\alpha u_2, u_2R_\alpha u_3, \dots, u_nR_\alpha y$. If x and y are α -identical with respect to fuzzy proximity relation R , then we write $xR(\alpha)y$, where the relation $R(\alpha)$ for each fixed $\alpha \in [0, 1]$ is an equivalence relation on U . The pair (U, R) is called a fuzzy approximation space. For any $\alpha \in [0, 1]$, we denote by R_α^* , the set of all equivalence classes of $R(\alpha)$. Also we call $(U, R(\alpha))$, the generated approximation space associated with R and α . Let us consider $X \subseteq U$. Then the rough set on fuzzy approximation space of X in $(U, R(\alpha))$, is denoted by $(\underline{X}_\alpha, \overline{X}_\alpha)$, where \underline{X}_α is the α -lower approximation of X and \overline{X}_α is the α -upper approximation of X . We define \underline{X}_α and \overline{X}_α as follows:

$$\underline{X}_\alpha = \cup \{Y : Y \in R_\alpha^* \text{ and } Y \subseteq X\} \quad (1)$$

$$\overline{X}_\alpha = \cup \{Y : Y \in R_\alpha^* \text{ and } Y \cap X \neq \emptyset\} \quad (2)$$

Then X is said to be α -discernible if and only if $\underline{X}_\alpha = \overline{X}_\alpha$ and X is said to be α -rough if and only if $\underline{X}_\alpha \neq \overline{X}_\alpha$.

2.1 Almost Indiscernibility Relation

An information system is one that provides all available information and knowledge about the objects under certain consideration. Objects are only perceived or measured by using a finite number of properties. At the same time, it does not consider any semantic relationships between distinct values of a particular attribute [15]. Different values of the same attribute are considered as distinct symbols without any connections, and therefore on simple pattern matching we consider horizontal analyses to a large extent. Hence, in general one uses the trivial equality relation on values of an attribute as discussed in standard rough set theory [4].

However, in many real life applications it is observed that the attribute values are not exactly identical but almost identical [9, 16, 17, 18]. At this point we generalize Pawlak's approach of indiscernibility. Keeping view to this, the almost indiscernibility relation generated in this way is the basis of rough set on fuzzy approximation space as discussed in the previous section [9]. Generalized information table may be viewed as information tables with added semantics. For

the problem of predicting missing associations, we introduce order relations on attribute values [19]. However, it is not appropriate in case of attribute values that are almost indiscernible.

Let U be the universe and A be a set of attributes. With each attribute $a \in A$ we associate a set of its values V_a , called the domain of a . The pair $S = (U, A)$ will be called an information system. Let $B \subseteq A$. For $\alpha \in [0,1]$, we denote a binary relation $R_B(\alpha)$ on U defined by $xR_B(\alpha)y$ if and only if $x(a)R(\alpha)y(a)$ for all $a \in B$, where $x(a) \in V_a$ denotes the value of x in a . Obviously, it can be proved that the relation $R_B(\alpha)$ is an equivalence relation on U . Also, we notice that $R_B(\alpha)$ is not exactly the indiscernibility relation defined by Pawlak [9]; rather it can be viewed as an almost indiscernibility relation on U . For $\alpha=1$ the almost indiscernibility relation, $R_B(\alpha)$ reduces to the indiscernibility relation. Thus, it generalizes the Pawlak's indiscernibility relation. The family of all equivalence classes of $R_B(\alpha)$ i.e., the partition generated by B for $\alpha \in [0,1]$, will be denoted by $U/R_B(\alpha)$. If $(x, y) \in R_B(\alpha)$, then we will say that x

and y are α - indiscernible. Blocks of the partition $U/R_B(\alpha)$ are referred as B_α - elementary concepts. These are the basic building concepts of our knowledge in the rough set on fuzzy approximation space.

III. Order Information System

Let $I = (U, A, V_a, f_a)$ be an information system, where U is a finite non-empty set of objects called the universe and A is a non-empty finite set of attributes. For every $a \in A$, V_a is the set of values that attribute a may take and $f_a : U \rightarrow V_a$ is an information function. A special case of information systems called information table or attribute value table where the columns are labeled by attributes and rows are by objects. Consider the information system given in Table 1. Here, we have $A = \{3G, \text{Touch screen } (ts), \text{Screen size } (ss), \text{Camera resolution } (cr), \text{Sim } (s), \text{Price}\}$ and $V_{3G} = \{Yes, No\}$. Similarly, we get $V_{ts} = \{Yes, No\}$, $V_{ss} = \{1.8'', 2.2'', 4.3'', 2.4'', 3.5''\}$, $V_s = \{Dual, Single\}$, $V_{cr} = \{\text{VGA}, 3.2\text{MP}, 8\text{MP}, 3.2\text{MP}, 2\text{MP}\}$, and $V_{\text{Price}} = \{3500, 13300, 29300, 13000, 30000\}$.

Table 1: Information System

Object	3G	Touch screen	Screen size	Camera resolution	Sim	Price (Rs.)
Alcatel OT355D (o_1)	No	No	1.8''	VGA	Dual	3500
Black Berry pearl 3G 9100 (o_2)	Yes	No	2.2''	3.2MP	Single	13300
HTC sensation (o_3)	Yes	Yes	4.3''	8.0MP	Single	29300
Nokia E71 (o_4)	Yes	No	2.4''	3.2MP	Single	13000
Apple iphone 3G (o_5)	Yes	Yes	3.5''	2.0MP	Single	30000

An ordered information system is defined as $OIS = \{I, \{\prec_x : x \in A\}\}$ where, I is a standard information system and \prec_a is an order relation on attribute a . An ordering of values of a particular attribute a naturally induces an ordering of objects:

$$x \prec_{\{a\}} y \Leftrightarrow f_a(x) \prec_a f_a(y) \quad (3)$$

where, $\prec_{\{a\}}$ denotes an order relation on U induced by the attribute a . An object o_i is ranked ahead of object o_j if and only if the value of o_i on the attribute a is ranked ahead of the value of o_j on the attribute a . For example, information system given in Table I becomes order information system on introduction of the following ordering relations.

$$\prec_{3G}: \text{Yes} \prec \text{No}$$

$$\prec_{ts}: \text{Yes} \prec \text{No}$$

$$\prec_{ss}: 4.3'' \prec 3.5'' \prec 2.4'' \prec 2.2'' \prec 1.8''$$

$$\prec_{cr}: 8\text{MP} \prec 3.2\text{MP} \prec 2\text{MP} \prec \text{VGA}$$

$$\prec_s: \text{Dual} \prec \text{Single}$$

$$\prec_{\text{Price}}: 30000 \prec 29300 \prec 13300 \prec 13000 \prec 3500$$

For a subset of attributes $B \subseteq A$, we define:

$$x \prec_B y \Leftrightarrow f_a(x) \prec_a f_a(y) \quad \forall a \in B$$

$$\Leftrightarrow \bigwedge_{a \in B} f_a(x) \prec_a f_a(y) \Leftrightarrow \bigcup_{a \in B} \prec_{\{a\}}$$

It indicates that x is ranked ahead of y if and only if x is ranked ahead of y according to all attributes in B .

The above definition is a straightforward generalization of the standard definition of equivalence relations in rough set theory [4], where the equality relation is used. Knowledge mining based on order relations is a concrete example of applications on generalized rough set model with non equivalence relations [17].

In this paper we use rough sets on fuzzy approximation space to find the attribute values that are α -identical before introducing the order relation. This is because exactly ordering is not possible when the attribute values are almost identical. Also, it generalizes the Pawlak's indiscernibility relation for $\alpha = 1$.

IV. Bayesian Classification

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can help provide us with a better understanding of the high dimensional data. In general, classification is used to predict future data trends. However, classification also predicts categorical labels [20]. In this section, we discuss the fundamental concepts of Bayesian classification that can predict class membership probabilities. Bayesian classification is derived from Bayes' theorem. Different studies on classification algorithm are found in [21].

Let T be a data tuple. In Bayesian terms, T is considered as evidence. As usual, it is described by measurements made on a set of q -attributes. Let H be some hypothesis, such that the data tuple T belongs to some specified class C . For classification, we determine $P(H|T)$, the probability that the hypothesis H holds given the evidence T . We define according to Bayes' theorem as:

$$P(H|T) = \frac{P(T|H)P(H)}{P(T)} \quad (4)$$

Now, we present the definitions, notations and results on Bayesian classification. Let D be a training set of tuples, where each tuple T is represented by q -dimensional attribute vector $T = (t_1, t_2, t_3, \dots, t_q)$, with $t_i = x(a_i)$, $i = 1, 2, 3, \dots, q$; depicting q measurements made on the tuple from q attributes, respectively, $a_1, a_2, a_3, \dots, a_q$.

Suppose that the decision attribute d has m classes $C_1, C_2, C_3, \dots, C_m$. Therefore, $V_d = \{C_1, C_2, C_3, \dots, C_m\}$. Given a tuple T , the classifier will predict that T belongs to class having the highest posterior probability, conditioned on T . That is, the naive Bayesian classifier predicts that tuple T belongs to class C_i if and only if $P(C_i|T) > P(C_j|T)$ for $j \neq i; 1 \leq i, j \leq m$. The class

C_i for which $P(C_i|T)$ is maximum is called the maximum posteriori hypothesis. By Bayes theorem

$$P(C_i|T) = \frac{P(T|C_i)P(C_i)}{P(T)} \quad (5)$$

As $P(T)$ is constant for all classes, only $P(T|C_i)P(C_i)$ need to be maximized. If the class prior probabilities are not known, then it is assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \dots = P(C_m)$. Therefore, we would maximize $P(T|C_i)$. Otherwise the class prior probability $P(C_i) = |C_i|/|D|$ is to be estimated, where $|C_i|$ is the cardinality of the C_i and $P(T|C_i)P(C_i)$ is to be maximized. Given dataset with many attributes it is observed that computing $P(T|C_i)$ is computationally expensive. Thus the naïve assumption of class conditional independence is made to reduce computations. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple. Therefore,

$$P(T|C_i) = \prod_{k=1}^q P(t_k|C_i) \quad (6)$$

where t_k refers to the value of attribute a_k for tuple T . We define $P(t_k|C_i) = n_c/n$, where n_c, n is defined as the number of times that the attribute value t_k was seen with the label C_i and the number of times C_i is seen in the decision attribute d respectively. It is also observed that sometimes we might not see a particular value t_k with a particular label C_i . This results in zero probability as $n_c = 0$. Again this zero probability will dominate the classification of future instances as Bayes classifier multiplies the $P(t_k|C_i)$ together. In order to overcome this problem we need to hallucinate some counts to generalize beyond our training material by using m -estimate as discussed by Mitchell in machine learning [22]. Therefore, we define $P(t_k|C_i)$ as

$$P(t_k|C_i) = \frac{n_c + mp}{n + m} \quad (7)$$

where p is the prior estimate of the probability and m is the equivalent sample size (constant). In the absence of other information, assume a uniform prior $p = \frac{1}{k}$, where $k = |V_{a_k}|$.

V. Proposed Prediction Model

In this section, we propose our association rule prediction model that consists of pre process and post process as shown in Fig. 1. In pre process, we process the data after data cleaning by using rough set on fuzzy

approximation space and ordering rules. Based on the classification obtained in pre process, Bayesian classification is used in post process to predict the missing association of attribute values. The main advantage of this model is that, it works for both literature and numerical data.

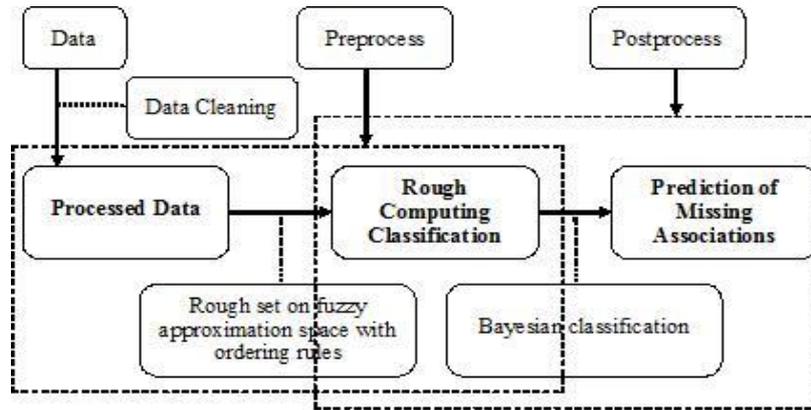


Fig. 1: Proposed prediction model

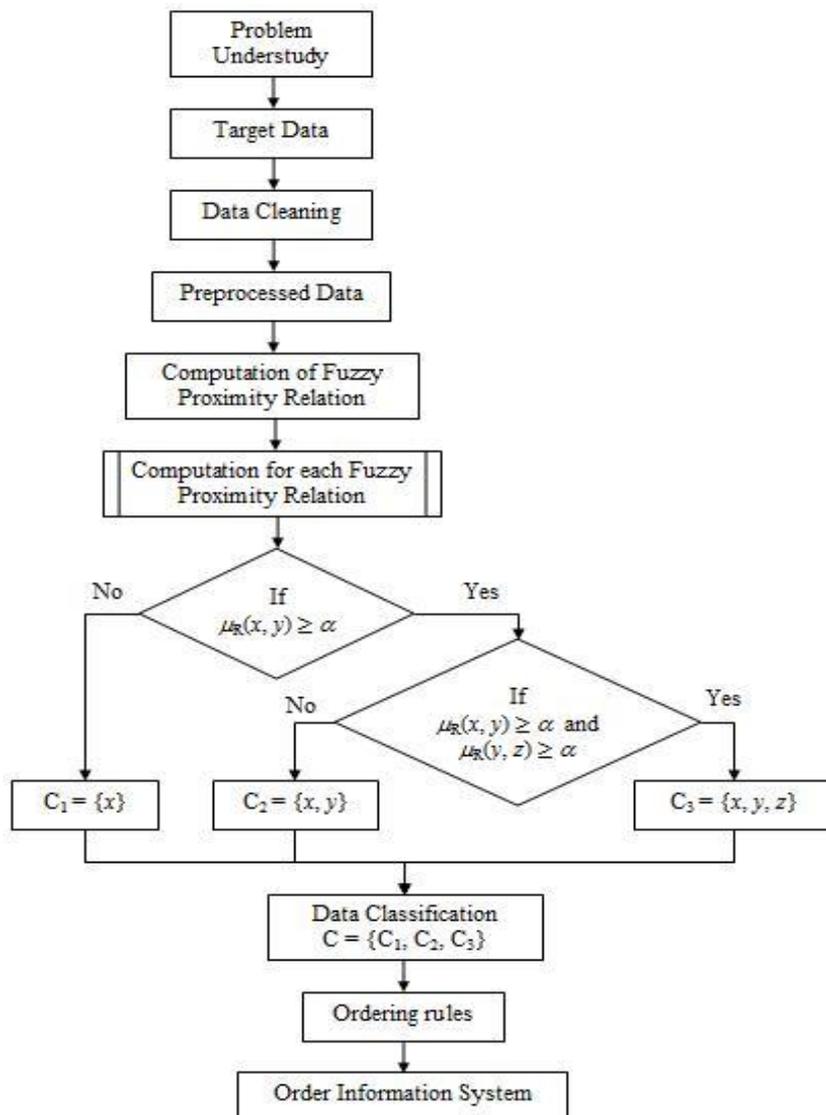


Fig. 2: Preprocessing architecture design

The fundamental step of any model is the identification of right problem. Incorporation of prior knowledge is always associated with the problem definition however; the potential validity or usefulness of an individual data element or pattern of data element may change dramatically from organization to organization because of the acquisition of knowledge and reasoning that may be involved in vagueness and incompleteness. It is very difficult for human beings to predict missing associations that is hidden in the high dimensional data. Therefore, the most important challenge is to predict data pattern and unseen associations from the accumulated high dimensional data. Hence, it is essential to deal with the incomplete and vague information in classification, data analysis, and concept formulation. To this end we use rough set on fuzzy approximation space with ordering rules in preprocess to mine suitable classification. In preprocess as shown in Fig. 1 we use rough set on fuzzy approximation spaces with ordering rules for processing data, and data classification after removal of noise and missing data. Based on the classification obtained in preprocess, we use Bayesian classification to predict decision for missing or unseen associations.

5.1 Preprocessing Architecture Design

In this section, we present our preprocess architecture design that consists of problem undergone, target data, data cleaning, fuzzy proximity relation, data classification, and ordering rules as shown in Fig. 2. Problem definition and incorporation of prior knowledge are the fundamental steps of any model. Then structuring the objectives and the associated attributes a target dataset is created on which data mining is to be performed. Before further analysis a sequence of data cleaning tasks such as removing noise, consistency check, and data completeness is done to ensure that the data are as accurate as possible. Finally for each attribute, we compute the α -equivalence classes based on the almost indiscernibility relation as discussed in section 2. The fuzzy proximity relation identifies the almost indiscernibility among the objects. This result induces the α -equivalence classes. We obtain categorical classes on imposing order relation on this classification.

VI. An Empirical Study on Marketing Strategies

In this section, we demonstrate the proposed model by considering a real life problem for extracting information. We consider the case study in which we study the different cosmetic company's business strategies in a country. In Table 2 given below, we consider a few parameters for business strategies to get maximum sales; their possible range of values and a fuzzy proximity relation which characterizes the relationship between parameters. We define a fuzzy proximity relation $R(x_i, x_j)$ in order to identify the

almost indiscernibility among the objects x_i and x_j , where

$$R(x_i, x_j) = 1 - \frac{|V_{x_i} - V_{x_j}|}{2(V_{x_i} + V_{x_j})} \quad (8)$$

The membership function has been adjusted in such a manner that their values should lie in $[0, 1]$ and these functions must also be symmetric. The requirement in the numerator necessitates a major of 2 in the denominator. The companies having high expenditure in marketing, advertisement, distribution, miscellaneous, and research and development is the ideal case. But such a blend of cases is rare in practice. So, a company may not excel in all the parameters in order to get maximum sales. However, out of these parameters, some parameters may have greater influence on others. But, the attribute values on these parameters obtained are almost indiscernible and hence can be classified by using rough set on fuzzy approximation space [9] and ordering rules.

The membership function has been adjusted such that its value should lie in $[0, 1]$ and also the function must be symmetric. The companies are judged by the sales output that is produced. The amount of sales is judged by the different parameters of the companies. These parameters form the attribute set for our analysis. Here the marketing expenditure means, all expenditure incurred for corporate promotion, which includes event marketing, sales promotion, direct marketing etc. which comes to around 6%. The advertising expenditure includes promotional activities using various medium like television, newspaper, internet etc. which comes around 36%. The miscellaneous expenditure is mainly incurred through activities like corporate social responsibility and it leads to maximum of 28%. The distribution cost includes expense on logistic, supply chain etc. and it comes around 24%. The investment made on new product development and other research activities are taken on research and development activities and it takes around 6% and the last one, the sales which basically deals with the sales that a company can produce after investing the expenditure in different fields mentioned above. The company can observe the profit by subtracting the value of the total expenditure from the value of the total sales. The data collected is considered to be the representative figure and tabulated below in Table 2.

Table 2: Notation representation table

Parameter	Attribute	Possible range	Membership Function
Expenditure on marketing	Mkt.	[1 – 150]	$1 - \frac{ V_{x_i} - V_{x_j} }{2(V_{x_i} + V_{x_j})}$
Expenditure on advertisement	Advt.	[1 – 900]	$1 - \frac{ V_{x_i} - V_{x_j} }{2(V_{x_i} + V_{x_j})}$
Expenditure on distribution	Dist.	[1 – 600]	$1 - \frac{ V_{x_i} - V_{x_j} }{2(V_{x_i} + V_{x_j})}$
Expenditure on miscellaneous	Misc.	[1 – 700]	$1 - \frac{ V_{x_i} - V_{x_j} }{2(V_{x_i} + V_{x_j})}$
Expenditure on research and development	R&D	[1 – 150]	$1 - \frac{ V_{x_i} - V_{x_j} }{2(V_{x_i} + V_{x_j})}$
Sales	Sales	[1– 12000]	$1 - \frac{ V_{x_i} - V_{x_j} }{2(V_{x_i} + V_{x_j})}$

In the Table 3 we present the data obtained from ten different companies. However, we keep their identity confidential due to various official reasons. Here we use the notation x_i , $i = 1, 2, 3, \dots, 10$ for different

companies for the purpose of our study to demonstrate the proposed prediction model. It is to be noted that, in the information table all non-ratio figures shown in the Table 3 are ten million INR.

Table 3: Sample information system

Comp.	Mkt.	Advt.	Dist.	Misc.	R&D	Sales
x_1	18.276	162.236	30.236	72.146	9.156	1220.586
x_2	2.076	5.393	6.793	8.290	0.383	215.767
x_3	0.496	1.330	0.433	2.733	0.393	42.593
x_4	0.940	0.060	0.666	5.890	1.243	166.41
x_5	27.333	38.660	16.496	24.343	1.523	561.697
x_6	7.033	866.916	508.676	637.530	38.963	11449.56
x_7	4.323	4.173	1.753	3.176	0.003	60.89
x_8	38.516	40.046	3.126	8.026	0.056	303.57
x_9	0.466	0.460	0.993	3.803	0.053	62.836
x_{10}	0.603	0.036	0.393	0.613	0.016	20.523

6.1 Preprocess of Empirical Study

In this section, we discuss in detail the subsequent steps of the preprocess architecture design for the empirical study taken under consideration. A target dataset for analysis as shown in Table 3 is considered. We have designed fuzzy proximity relations based on the attributes and computed the almost similarity between them. The fuzzy proximity relation identifies

the almost indiscernibility among the objects. This result induces the equivalence classes. We obtain categorical classes on imposing order relation on this classification. The fuzzy proximity relations R_i , $i = 1, 2, 3, 4, 5, 6$ corresponding to the attributes Mkt., Advt., Dist., Misc., R&D, and Sales is given in Tables 4, 5, 6, 7, 8, 9, respectively.

Table 4: Fuzzy proximity relation for attribute mkt.

R_1	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000	0.602	0.526	0.549	0.901	0.778	0.691	0.822	0.525	0.532
x_2	0.602	1.000	0.693	0.812	0.571	0.728	0.824	0.551	0.683	0.725
x_3	0.526	0.693	1.000	0.845	0.518	0.566	0.603	0.513	0.984	0.951
x_4	0.549	0.812	0.845	1.000	0.533	0.618	0.679	0.524	0.831	0.891
x_5	0.901	0.571	0.518	0.533	1.000	0.705	0.637	0.915	0.517	0.522
x_6	0.778	0.728	0.566	0.618	0.705	1.000	0.881	0.654	0.562	0.579
x_7	0.691	0.824	0.603	0.679	0.637	0.881	1.000	0.601	0.597	0.622
x_8	0.822	0.551	0.513	0.524	0.915	0.654	0.601	1.000	0.512	0.515
x_9	0.525	0.683	0.984	0.831	0.517	0.562	0.597	0.512	1.000	0.936
x_{10}	0.532	0.725	0.951	0.891	0.522	0.579	0.622	0.515	0.936	1.000

Table 5: Fuzzy proximity relation for attribute advt.

R_2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000	0.532	0.508	0.500	0.692	0.658	0.525	0.698	0.503	0.500
x_2	0.532	1.000	0.698	0.511	0.622	0.506	0.936	0.619	0.579	0.507
x_3	0.508	0.698	1.000	0.543	0.533	0.502	0.742	0.532	0.757	0.526
x_4	0.500	0.511	0.543	1.000	0.502	0.500	0.514	0.501	0.615	0.875
x_5	0.692	0.622	0.533	0.502	1.000	0.543	0.597	0.991	0.512	0.501
x_6	0.658	0.506	0.502	0.500	0.543	1.000	0.505	0.544	0.501	0.500
x_7	0.525	0.936	0.742	0.514	0.597	0.505	1.000	0.594	0.599	0.509
x_8	0.698	0.619	0.532	0.501	0.991	0.544	0.594	1.000	0.511	0.501
x_9	0.503	0.579	0.757	0.615	0.512	0.501	0.599	0.511	1.000	0.573
x_{10}	0.500	0.507	0.526	0.875	0.501	0.500	0.509	0.501	0.573	1.000

Table 6: Fuzzy proximity relation for attribute dist.

R_3	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000	0.683	0.514	0.522	0.853	0.556	0.555	0.594	0.532	0.513
x_2	0.683	1.000	0.560	0.589	0.792	0.513	0.705	0.815	0.628	0.555
x_3	0.514	0.560	1.000	0.894	0.526	0.501	0.698	0.622	0.804	0.976
x_4	0.522	0.589	0.894	1.000	0.539	0.501	0.775	0.676	0.901	0.871
x_5	0.853	0.792	0.526	0.539	1.000	0.531	0.596	0.660	0.557	0.523
x_6	0.556	0.513	0.501	0.501	0.531	1.000	0.503	0.506	0.502	0.501
x_7	0.555	0.705	0.698	0.775	0.596	0.503	1.000	0.859	0.862	0.683
x_8	0.594	0.815	0.622	0.676	0.660	0.506	0.859	1.000	0.741	0.612
x_9	0.532	0.628	0.804	0.901	0.557	0.502	0.862	0.741	1.000	0.784
x_{10}	0.513	0.555	0.976	0.871	0.523	0.501	0.683	0.612	0.784	1.000

Table 7: Fuzzy proximity relation for attribute misc.

R_4	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000	0.603	0.536	0.575	0.752	0.602	0.542	0.600	0.550	0.508
x_2	0.603	1.000	0.748	0.915	0.754	0.513	0.777	0.992	0.814	0.569
x_3	0.536	0.748	1.000	0.817	0.601	0.504	0.963	0.754	0.918	0.683
x_4	0.575	0.915	0.817	1.000	0.695	0.509	0.850	0.923	0.892	0.594
x_5	0.752	0.754	0.601	0.695	1.000	0.537	0.615	0.748	0.635	0.525
x_6	0.602	0.513	0.504	0.509	0.537	1.000	0.505	0.512	0.506	0.501
x_7	0.542	0.777	0.963	0.850	0.615	0.505	1.000	0.784	0.955	0.662
x_8	0.600	0.992	0.754	0.923	0.748	0.512	0.784	1.000	0.821	0.571
x_9	0.550	0.814	0.918	0.892	0.635	0.506	0.955	0.821	1.000	0.639
x_{10}	0.508	0.569	0.683	0.594	0.525	0.501	0.662	0.571	0.639	1.000

Table 8: Fuzzy proximity relation for attribute r&d.

R_5	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000	0.540	0.541	0.620	0.643	0.690	0.500	0.506	0.506	0.502
x_2	0.540	1.000	0.994	0.736	0.701	0.510	0.508	0.628	0.622	0.540
x_3	0.541	0.994	1.000	0.740	0.705	0.510	0.508	0.625	0.619	0.539
x_4	0.620	0.736	0.740	1.000	0.949	0.531	0.502	0.543	0.541	0.513
x_5	0.643	0.701	0.705	0.949	1.000	0.538	0.502	0.535	0.534	0.510
x_6	0.690	0.510	0.510	0.531	0.538	1.000	0.500	0.501	0.501	0.500
x_7	0.500	0.508	0.508	0.502	0.502	0.500	1.000	0.551	0.554	0.658
x_8	0.506	0.628	0.625	0.543	0.535	0.501	0.551	1.000	0.986	0.722
x_9	0.506	0.622	0.619	0.541	0.534	0.501	0.554	0.986	1.000	0.732
x_{10}	0.502	0.540	0.539	0.513	0.510	0.500	0.658	0.722	0.732	1.000

Table 9: Fuzzy proximity relation for attribute sales.

R_6	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.000	0.650	0.533	0.619	0.815	0.596	0.547	0.699	0.548	0.516
x_2	0.650	1.000	0.644	0.995	0.777	0.518	0.720	0.915	0.725	0.586
x_3	0.533	0.644	1.000	0.703	0.570	0.503	0.911	0.623	0.903	0.825
x_4	0.619	0.995	0.703	1.000	0.728	0.514	0.767	0.854	0.774	0.609
x_5	0.815	0.777	0.570	0.728	1.000	0.546	0.597	0.850	0.600	0.535
x_6	0.596	0.518	0.503	0.514	0.546	1.000	0.505	0.525	0.505	0.501
x_7	0.547	0.720	0.911	0.767	0.597	0.505	1.000	0.667	0.992	0.752
x_8	0.699	0.915	0.623	0.854	0.850	0.525	0.667	1.000	0.671	0.563
x_9	0.548	0.725	0.903	0.774	0.600	0.505	0.992	0.671	1.000	0.746
x_{10}	0.516	0.586	0.825	0.609	0.535	0.501	0.752	0.563	0.746	1.000

Now on considering the almost similarity of 85% *i.e.*, $\alpha \geq 0.85$ it is observed from Table 4 that $R_1(x_1, x_1) = 1$; $R_1(x_1, x_5) = 0.901$; $R_1(x_2, x_2) = 1$; $R_1(x_3, x_3) = 1$; $R_1(x_3, x_9) = 0.984$; $R_1(x_3, x_{10}) = 0.951$; $R_1(x_4, x_4) = 1$; $R_1(x_4, x_{10}) = 0.891$; $R_1(x_5, x_5) = 1$; $R_1(x_5, x_8) = 0.915$; $R_1(x_6, x_6) = 1$; $R_1(x_6, x_7) = 0.881$. Thus, the companies x_1, x_5, x_8 are α -identical. Similarly, x_6, x_7 are α -identical; x_3, x_4, x_9, x_{10} are α -identical and x_2 is α -identical. Therefore, we get

$$U/R_1^\alpha = \{ \{x_1, x_5, x_8\}, \{x_6, x_7\}, \{x_3, x_4, x_9, x_{10}\}, \{x_2\} \}$$

Therefore, the values of the attribute expenditure on marketing are classified into four categories namely very low, low, average and high and hence can be ordered.

Similarly, the different equivalence classes obtained from Table 5, 6, 7, 8, 9 corresponding to the attributes Advt., Dist., Misc., R&D, and Sales are given below.

$$U/R_2^\alpha = \{ \{x_1\}, \{x_2, x_7\}, \{x_3\}, \{x_4, x_{10}\}, \{x_5, x_8\}, \{x_6\}, \{x_9\} \}$$

$$U/R_3^\alpha = \{ \{x_1, x_5\}, \{x_2\}, \{x_3, x_4, x_7, x_8, x_9, x_{10}\}, \{x_6\} \}$$

$$U/R_4^\alpha = \{ \{x_1\}, \{x_2, x_3, x_4, x_7, x_8, x_9\}, \{x_5\}, \{x_6\}, \{x_{10}\} \}$$

$$U/R_5^\alpha = \{ \{x_1\}, \{x_2, x_3\}, \{x_4, x_5\}, \{x_6\}, \{x_7\}, \{x_8, x_9\}, \{x_{10}\} \}$$

$$U/R_6^\alpha = \{ \{x_1\}, \{x_2, x_4, x_5, x_8\}, \{x_3, x_7, x_9\}, \{x_{10}\}, \{x_6\} \}$$

From the above classification, it is clear that the values of the attribute expenditure on advertisement are classified into seven categories namely poor, very low, low, average, high, very high and outstanding. The values of the attribute expenditure on distribution are classified into four categories namely low, average, high, and very high. The values of the attribute expenditure on miscellaneous are classified into five categories namely very low, low, average, high and very high. The values of the attribute expenditure on research and development are classified into seven categories namely poor, very low, low, average, high, very high and outstanding. Finally, the values of the attribute sales are classified into five categories namely very low, low, average, high and very high. Therefore the ordered information system of the business strategies of different cosmetic companies of Table 3 is given below in Table 10.

Table 10: Order information system

Comp.	Mkt.	Advt.	Dist.	Misc.	R&D	Sales
x_1	High	Very high	High	High	Very high	High
x_2	Low	Avg.	Avg.	Low	Avg.	Avg.
x_3	Very low	Low	Low	Low	Avg.	Low
x_4	Very low	Poor	Low	Low	High	Avg.
x_5	High	High	High	Avg.	High	Avg.
x_6	Avg.	Outstanding	Very high	Very high	Outstanding	Very high
x_7	Avg.	Avg.	Low	Low	Poor	Low
x_8	High	High	Low	Low	Low	Avg.
x_9	Very low	Very low	Low	Low	Low	Low
x_{10}	Very low	Poor	Low	Very low	Very low	Very low

$\prec_{Mkt.}$: High \prec Average \prec Low \prec Very Low

$\prec_{Advt.}$: Outstanding \prec Very High \prec High \prec Average \prec Low \prec Very Low \prec Poor

$\prec_{Dist.}$: Very High \prec High \prec Average \prec Low

$\prec_{Misc.}$: Very High \prec High \prec Average \prec Low \prec Very Low

$\prec_{R\&D.}$: Outstanding \prec Very High \prec High \prec Average \prec Low \prec Very Low \prec Poor

$\prec_{Sales.}$: Very High \prec High \prec Average \prec Low \prec Very Low

6.2 Postprocess of Empirical Study

Bayesian classification analysis can do the data classification. However data are already classified in preprocess. The objective of this process is to use Bayesian classification to predict the unseen association rule from the order information system and hence to get better knowledge affecting the decision making. In order to show post processing analysis, we consider an unseen association of attribute values $T = \{Mkt. = High, Advt. = Average, Dist. = Low, Misc. = Average, R\&D = High\}$ to predict the decision 'sales'. Let us take $T = \{t_1, t_2, t_3, t_4, t_5\}$, where t_1 is Mkt. = High; t_2

is Advt. = Average; t_3 is Dist. = Low; t_4 is Misc. = Average; and t_5 is R&D = High. From the preprocessing, it is clear that the decision 'sales' has 5 classes say C_1 = Very high, C_2 = High, C_3 =Average, C_4 = Low , and C_5 = Very low. But, $P(t_i | C_1) = 0$ for $i = 1, 2, 3, 4, 5$ with $P(C_1) = 1/10$. Thus, it dominates the classification of future instances as Bayes classifier multiplies $P(t_i | C_1)$ together. Therefore, by using m -estimate we get:

$$P(t_1 | C_1) = \frac{0 + 4(\frac{1}{4})}{1 + 4} = \frac{1}{5}$$

$$P(t_2 | C_1) = \frac{0 + 7(\frac{1}{7})}{1 + 7} = \frac{1}{8}$$

$$P(t_3 | C_1) = \frac{0 + 4(\frac{1}{4})}{1 + 4} = \frac{1}{5}$$

$$P(t_4 | C_1) = \frac{0 + 5(\frac{1}{5})}{1 + 5} = \frac{1}{6}$$

$$P(t_5 | C_1) = \frac{0 + 7(\frac{1}{7})}{1 + 7} = \frac{1}{8}$$

and

$$P(C_1) = \frac{1 + 5(\frac{1}{5})}{10 + 5} = \frac{2}{15}$$

Therefore, by using Bayesian classification we get

$$P(C_1 | T) = \prod_{i=1}^5 P(t_i | C_1) P(C_1) = \frac{1}{5} \times \frac{1}{8} \times \frac{1}{5} \times \frac{1}{6} \times \frac{1}{8} \times \frac{2}{15} = 0.0000138$$

Similarly, we get $P(C_2 | T) = 0.0000277$; $P(C_3 | T) = 0.000516$; $P(C_4 | T) = 0.0000544$; and $P(C_5 | T) = 0.0000173$. From the above computations it is clear that $P(C_3 | T)$ is maximum. Therefore, it is clear that the above unseen association of attribute values belongs to the decision class C_3 = Average. Keeping view to the length of the paper, some of the unseen association of attribute values and its corresponding decision are presented in Table 11.

Table 11: Prediction of unseen associations

Unseen associations	Mkt.	Advt.	Dist.	Misc.	R&D	Sales
1	Very low	High	Low	Low	High	Avg.
2	Very low	Avg.	Low	Low	Avg.	Low
3	Very low	High	Low	Low	Avg.	Avg.
4	High	Very high	Very high	High	Outstanding	High
5	High	Avg.	Low	Low	Avg.	Avg.
6	High	High	Low	Low	High	Avg.
7	Very low	Avg.	Low	Low	Low	Low
8	Very low	Avg.	Low	Low	High	Avg.
9	Avg.	Avg.	Low	Low	Low	Low
10	High	Poor	Low	Low	High	Avg.
11	Avg.	Avg.	Low	Low	Avg.	Low

VII. Conclusion

Prediction of unseen or missing associations of attribute values is a challenging task in the study of high dimensional database. In general, Bayesian classification is used to predict the unseen association rule. However, it is not directly applicable in case of information system containing almost indiscernible attribute values. In order to overcome this problem the proposed prediction model uses both rough computing with ordering rules and Bayesian classification. The model identifies the almost indiscernibility between the attribute values in the preprocess phase whereas

Bayesian classification is used in the post process to predict the decision. This helps the decision maker a priori prediction of sales. We have taken a real life example of 10 cosmetic company's database according to different attributes and shown how analysis can be performed by considering the proposed model. We believe that, rough computing with ordering rules together with Bayesian classification can be used to find furthermore information regardless of the type of associations based soft computing. We also believe that the proposed model is a useful method for decision makers and mining knowledge.

References

- [1] Fayaad U M, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: An overview. In: *Advances in Knowledge Discovery and Data Mining* (Fayaad U M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Eds), American Association for Artificial Intelligence (AAAI) Press, California, 1996. 1~34.
- [2] Rocha L M. TalkMine: A soft computing approach to adaptive knowledge recommendation. In: *Soft Computing Agents: New Trends for Designing Autonomous Systems* (V Loia, S Sessa Eds). Series on Studies in Fuziness and Soft Computing, Physica-Verlag, Springer, 2001. 89~116.
- [3] Zadeh L A. Fuzzy sets. *Information and Control*, 1965, 8: 338-353.
- [4] Pawlak Z. Rough sets. *International Journal of Computer and Information Sciences*, 1982, 11: 341-356.
- [5] Pawlak Z, Skowron A. Rudiments of rough sets. *Information Sciences*, Elsevier, 2007, 177 (1): 3-27.
- [6] Pawlak Z, Skowron A. Rough sets: some extensions. *Information Sciences*, Elsevier, 2007, 177 (1): 28-40.
- [7] Pawlak Z, Skowron A. Rough sets and Boolean reasoning. *Information Sciences*, Elsevier, 2007, 177 (1): 41-73.
- [8] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. *International Journal of General System*, 1990, 17: 191-209.
- [9] Acharjya D P, Tripathy B K. Rough sets on fuzzy approximation spaces and applications to distributed knowledge systems. *International Journal of Artificial Intelligence and Soft Computing*, 2008, 1 (1): 1-14.
- [10] Acharjya D P, Tripathy B K. Rough sets on intuitionistic fuzzy approximation spaces and knowledge representation. *International Journal of Artificial Intelligence and Computational Research*, 2009, 1 (1): 29-36.
- [11] Dong Ya Li, Bao Qing Hu. A kind of dynamic rough sets. *Proceedings of the fourth International Conference on Fuzzy Systems and Knowledge Discovery*, 2007. 79~85.
- [12] Slowinski R, Vanderpooten D. A generalized definition of rough approximations based on similarity. *IEEE Trans. on Knowledge and Data Engineering*, 2000, 12 (2): 331-336.
- [13] Rubin D B. Inference and missing data. *Biometrika*, 1976, 63: 581-592.
- [14] Little R J A, Rubin D B. *Statistical analysis with missing data*, Second Edition, Wiley-Interscience, NJ, USA, 2002.
- [15] Yao Y Y. Information tables with neighborhood semantics. In: *Data Mining and Knowledge Discovery-Theory, Tools, and Technology* (Dasarathy B V. Ed.), Society for Optical Engineering, Bellingham, Washington, 2000, 2: 108~116.
- [16] Tripathy B K, Acharjya D P, Cynthia V. A framework for intelligent medical diagnosis using rough set with formal concept analysis. *International Journal of Artificial Intelligence and Applications*, 2011, 2 (2): 45-66.
- [17] Acharjya D P, Ezhilarsi L. A knowledge mining model for ranking institutions using rough computing with ordering rules and formal concept analysis. *International Journal of Computer Science Issues*, 2011, 8 (2): 417-425.
- [18] Acharjya D P. Comparative study of rough sets on fuzzy approximation spaces and intuitionistic fuzzy approximation spaces. *International Journal of Computational and Applied Mathematics*, 2009, 4 (2): 95-106.
- [19] Yao Y Y, Sai Ying. Mining ordering rules using rough set theory. *Bulletin of International Rough Set Society*, 2001, 5: 99-106.
- [20] Han Jiawei, Kamber Micheline. *Data Mining and Concepts and Techniques*. Elsevier, New York, 2006.
- [21] Lin J H, Haug P J. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, Elsevier, New York, 2008, 41: 1-14.
- [22] Mitchell Tom M. *Machine learning*. McGraw Hill, New York, 1997.

Authors' Profiles

D P Acharjya received his M. Tech. degree in computer science from Utkal University, India in 2002; M. Phil. from Berhampur University, India; and M. Sc. from NIT, Rourkela, India. He has been awarded with Gold Medal in M. Sc. He received his Ph. D in computer science from Berhampur University, India. Currently, he is an Associate Professor in the school of computing sciences and engineering, VIT University, Vellore, India. He has authored many national and international journal papers and four books; *Fundamental Approach to Discrete Mathematics*, *Computer Based on Mathematics*, *Theory of Computation*; *Rough Set in Knowledge Representation and Granular Computing* to his credit. He is associated with many professional bodies CSI, ISTE, IMS, AMTI, ISIAM, OITS, IACSIT, CSTA, IEEE and IAENG. He was founder secretary of OITS Rourkela chapter. His current research interests include rough sets, formal concept analysis, knowledge representation, data mining, granular computing and business intelligence.

Debasrita Roy received her B. Tech. degree in computer science from West Bengal University of Technology, West Bengal, India in 2011. She is a M. Tech (CSE) final year student of VIT University, Vellore, India. She has published two papers in International Conference. She has keen interest in teaching and applied research. Her research interest includes rough computing, granular computing, and soft computing.

Md. A Rahaman received his B. Tech. degree in computer science from West Bengal University of Technology, West Bengal, India in 2011. He is a M. Tech (CSE) final year student of VIT University, Vellore, India. He has keen interest in teaching and applied research. He has published two papers in International Conference. His research interest includes rough computing, and soft computing.

How to cite this paper: D. P. Acharjya, Debasrita Roy, Md. A. Rahaman, "Prediction of Missing Associations Using Rough Computing and Bayesian Classification", *International Journal of Intelligent Systems and Applications(IJISA)*, vol.4, no.11, pp.1-13, 2012. DOI: 10.5815/ijisa.2012.11.01