

Improved Decision Tree Methodology for the Attributes of Unknown or Uncertain Characteristics – Construction Project Prospective

Vijaya S. Desai, National Institute of Construction Management and
Research (NICMAR), Maharashtra, India

Abstract

Increasing use of computers, leads to accumulation of data of an organization, demanding the need of sophisticated data handling techniques. Many data handling concepts have evolved that support data analysis, and knowledge discovery. Data warehouse and Data mining techniques are playing an important role in the area of data analysis for knowledge discovery. These techniques typically address the four basic applications such as data classification, data clustering, association between data and finding sequential patterns between the data. Various algorithms that address to classification on large data sets have proved to be efficient in classifying the variables of known or certain characteristics. However they are less effective when applied to the analysis of variable of unknown or uncertain characteristics and creating classes by combining multiple correlated variables in real world. A methodology presented in the paper that addresses two major issues of data classification using decision tree, 1) classification of variables of unknown or uncertain characteristics, 2) creating classification by combining multiple correlated variables.

Keywords

User Intervention, Unknown Characteristics, Known Characteristics, Data Warehouse, Data Mining, Decision Tree, Guillotine Cut, Oblique Tree, Entropy, Gain, Uncertainty Coefficient.

Introduction

Use of computers, is leading to accumulation of valuable data giving rise to voluminous data of an organization. This is demanding the need of sophisticated data handling tools at all levels of business organization. Data warehousing technology comprises a set of

new concepts and tools which support the knowledge workers (executive, manager, and analyst) with information material for decision making (Gatziu and Vavouras, 1999).

Data warehouse is a database created by combining data from multiple databases for the purposes of analysis (AHMAD and NUNOO). Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that both understandable and useful to the data owner (David, et al. 2004). Various data mining techniques are available to mine the information from data warehouse. Such information has proved the basis of accurate decision making in the area of retail, banks, fraud detection, customer analysis etc. Decision tree is one of the classification techniques which generates a tree and set of rules, representing the model of different classes, from a given data set.

Literature review

The available algorithms can be broadly classified under two types: 1) that handle residence data analysis and 2) that handle large data analysis. Algorithms that address to residence data analysis include CART, ID3, C4.5, and C5, CHAID, QUEST, OC1, SAS. The algorithms that address to large data set include SLIQ and SPRINT, RainForest, Approximation Method, CLOUDS, BOAT (Pujari, 2001).

In the late 1970s J. Ross Quinlan introduced a decision tree algorithm named ID3, use information gain for predictions. ID3 was later enhanced in the version called C4.5. C4.5 and addressed several important areas: predictors with missing values, predictors with continuous values, and pruning. Classification and Regression Trees (CART) is a data exploration and prediction algorithm developed by Leo Breiman,

Jerome Friedman, Richard Olshen and Charles Stone. CHAID is similar to CART in that it builds a decision tree but it differs in the way that it chooses its splits. In SLIQ a single attribute list is maintained for an attribute. Ontology-Driven Decision Tree (ODT) algorithm describes an algorithm to learn classification rules at multiple levels of abstraction (Zhang et al. 2002). The researchers on the QUEST at IBM by Rakesh Agarwal and Team, proposed SLIQ in sequel. SLIQ is a scalable algorithm, which uses a pre-sorting technique integrated with a breadth-first tree growing strategy for the classification of the disk-resident data. SPRINT is the updated version of SLIQ and is meant for parallel implementation. SLIQ, SPRINT, RAINFOREST methods adopt exact methods. CLOUD (Classification of Large or Out-of-core Data Sets) is a kind of approximate version of the SPRINT method. It also uses the breadth first strategy to build the decision tree. CLOUD uses the gini index for evaluating the split index of the attributes. BOAT (Bootstrap Optimistic Algorithm for Tree Construction) is another approximate algorithm based on sampling.

Basic of decision tree

Decision tree is a classification technique which generates a tree and a set of rules, representing the model of different classes, from a given data set. The set of records available for developing classification is generally divided into two disjoint subsets – a training set and a test set. The former is used for deriving the classifier, while the latter is used to measure the accuracy of the classifiers. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified. The construction of decision tree involves the following three main phases (Pujari 2001).

- **Construction phase:** The initial tree is constructed in this phase based on the entire training data set. It requires the recursively partitioning the training set into two, or more, sub-partitions using splitting criteria, until a stopping criterion is met.
- **Pruning phase:** The tree constructed in the previous phase may not result in the best possible set of rules due to over-fitting. The pruning phase removes some of the lower branches and nodes to improve its performance.
- **Processing the pruned tree** to improve understandability.

The generic algorithm for decision tree construction is stated below (Almullim et al. 2002). Let $S = \{(X_1, c_1), (X_2, c_2), \dots, (X_k, c_k)\}$ be a training sample. Constructing a decision tree from S can be done in a divide-and-conquer fashion as follows:

Step 1: If all the examples in S are labeled with the same class, return a leaf labeled with that class.

Step 2: Choose some test t (according to some criterion) that has two or more mutually exclusive outcomes $\{O_1, O_2, \dots, O_r\}$.

Step 3: Partition S into disjoint subsets S_1, S_2, \dots, S_r , such that S_i consists of those examples having outcome O_i for the test t , for $i = 1, 2, \dots, r$.

Step 4: Call this tree-construction procedure recursively on each of the subsets S_1, S_2, \dots, S_r , and let the decision trees returned by these recursive calls be T_1, T_2, \dots, T_r .

Step 5: Return a decision tree T with a node labeled t as the root and the trees T_1, T_2, \dots, T_r as subtrees below that node.

The splitting attributes, is selected based on the influence the undependable attribute over the dependable attribute which is carried out by finding out the splitting indices. A popular practice is to measure the expected amount of information provided by

the test based on information theory. Given a sample S , the average amount of *information needed (entropy)* to find the class of a case in S is estimated by the function,

$$\text{info}(S) = - \sum_{i=1}^k \frac{|S^i|}{|S|} \times \log_2 \frac{|S^i|}{|S|}$$

where $S^i \subseteq S$ is the set of examples S of class i and k is the number of classes. If the subset S is further partitioned than suppose t is a test that partitions S into S_1, S_2, \dots, S_r ; then the *weighted average entropy* over these subsets is computed by,

$$\sum_{i=1}^r \frac{|S_i|}{|S|} \times \text{info}(S_i).$$

The information gain represents the difference between the information needed to identify an element of test t and the information needed to identify an element of test t after the value of attribute X is obtained. The *information gain* due to a split on the attribute is computed as,

$$\text{gain}(t) = \text{info}(S) - \sum_{i=1}^r \frac{|S_i|}{|S|} \times \text{info}(S_i)$$

To select the most informative test, the information gain for all the available test attributes is computed and the test with the maximum information gain is then selected. Although the information gain test selection criterion has been experimentally shown to lead to good decision trees in many cases, it was found to be biased in favor of tests that induce finer partitions. As an extreme example, consider the (meaningless) tests defined on attributes like *Activity Name* and *Project Name*. These tests would partition the training sample into a large number of subsets, each containing just one example. Because these subsets do not have a mixture of examples, their entropy is just 0, and so

the information gain of using these trivial tests is maximal. This bias in the gain criterion can be rectified by dividing the information gain of a test by the entropy of the test outcomes themselves, which measures the extent of splitting done by the test

$$\text{split}(t) = - \sum_{i=1}^r \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|},$$

Giving the *gain-ratio* measure

$$\text{gain-ratio}(t) = \frac{\text{gain}(t)}{\text{split}(t)}.$$

Objective of the study

Two major issues of concern in all of these algorithms are analysis of variables of unknown/uncertain characteristics and classification based on combining multiple variables. The algorithms that handle large data sets have proved to be efficient in classifying the variables of known or certain characteristics. For example in a retail shop a product has a unique characteristic once it is defined. That is a product such as ‘Washing machine’ of a make and model cannot change with customer. However analysis of variables of unknown or uncertain characteristics, these algorithms are found to be less effective. Example: work to be done by a labor depends on type of soil (hard, soft etc. at various locations, temperature). Here the soil has different characteristics and therefore cannot be uniquely defined.

Second issue is, splitting on single attribute may not correspond too well with the actual distribution of records in the decision space. This is called guillotine cut phenomenon (Pujari 2001). There can be variables having strong correlation and therefore have more accurate meaning in real world. The more accurate business meaning

can be therefore derived by combining the multiple variables. Dan Vance and Anca Ralescu have presented a methodology, to show how it is possible for a binary class problem to have a univariate decision tree that uses all attribute at once and create oblique line(s). Egmonts Treigut has presented a methodology to determine the correlation between attributes using *Group Method of Data Handling* developed by A.G.Ivakhnenko (Treiguts 2002). The method allows analyzing the correlation of attributes and its influence to the value of a class. The method however is experimented on small data set and needs to be experimented on large data domain that contains much more records than the order of function of approximation (Treiguts 2002).

In short the available algorithms are less efficient when applied to the analysis of variables of “*unknown/uncertain characteristics*” and **does not support combination of multiple variables** which have greater decision meaning in real world. The pruning techniques used sometimes may ignore those variables which may have more influence in real world. Selection of best test, scalability, overfitting, deciding the threshold to remove the attributes is based on statistical methods, ignoring the variables that may have greater meaning.

The objective of the study is, 1) to develop a methodology that will, facilitate the analysis of variables of unknown characteristics and enable to combine multiple variables for classification, 2) to apply the methodology on the practical data,

Methodology

The meaning of certain and uncertain characteristics of attributes are needed to understand to better apply the methodology.

Known characteristics (Certain): Variable of known characteristics can be defined as an object having certain and predefined characteristics. **Example:** The characteristics of a person whom a loan is offered are, his EMI, date of payment, interest rate, term of payment. These characteristics remain same and therefore said to be certain or known. The defaulting behavior of person to pay the loan is easy to analyze based on these certain and predefined characteristics.

Unknown characteristics (Uncertain): Variable of Unknown characteristics can be defined as an object whose characteristics are not known or uncertain. **Example 1:** Minimum and maximum temperature on a day, say 31st October can be different at different places on the same day. **Example 2:** Characteristics of soil can be different at different places of the world.

The methodology proposes a “*User Intervention*” approach at different stages of decision tree induction. Here an user is defined as a “person having sound knowledge and experience of the domain area on which analysis is to be carried out”. *User Intervention* is proposed at: 1) Selecting the dependent and independent attributes that would participate in tree construction 2) Selection of attributes to be combined for classification with the help of a proposed mathematical approach for combining multiple attributes., and 3) Defining the threshold.

Selecting the attributes: Steps for selecting the attributes are

The user will select the independent and dependable attribute. This will ensure that irrelevant attributes are not included in classification. The information gain for the selected attribute is calculated. The attributes will be sorted in the descending order of information gain. The attribute having highest information gain will be selected as root

node. The relevance for every attribute is carried out by computing uncertainty coefficient. The average uncertainty coefficient is considered as the specified threshold. Only the top most relevant attribute whose relevance exceeds the specified threshold are considered for classification. This will ignore the irrelevant attributes.

An Approach to combine multiple variables

The user can intervene and select the attributes to be combined. If the selected attributes are numeric then the median of the selected attributes are calculated. Each attribute then will have a left side and right side. Number of combinations of classes thus can be calculated as follows. Let the user select two numeric attributes, A1 and A2. As per the Step 2 and Step 3, let A1 has L1, R1 and A2 has L2, R2 sides. The possible combinations of classes are presented below.

L1 —————> L2

R1 —————> R2

Number of combinations at this stage are = 2

L1 —————> L2

R1 —————> R2

Number of combinations at this stage are = 2 + 1

L1 —————> L2

R1 —————> R2

Number of combinations at this stage are = 2 + 1+1 = 4.

If n is the number of selected attributes then the above takes a form = $n+n(n-1)$. The classes formed are; 1) $C1 = (A1 \leq m1 \text{ and } A2 \geq m2)$, 2) $C2 = (A1 > m1 \text{ and } A2 < m2)$, 3)

$C3 = (A1 \leq m1 \text{ and } A2 \leq m2)$, and 4) $C4 = (A1 > m1 \text{ and } A2 > m2)$. These classes can be added to the single node horizontally (Figure 1).

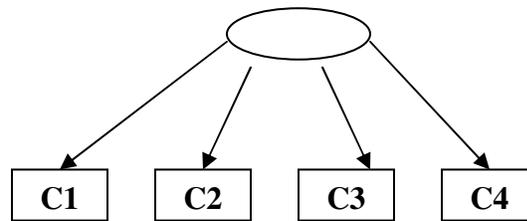


Figure 1: Classes

Defining threshold

The relevance analysis approach is adopted to define the threshold. The uncertainty

coefficient ($UC(X) = \frac{gain(X, T)}{Info(T)}$) for each attribute is computed. The average

uncertainty coefficient is considered as specified threshold. Only the top most relevant attribute whose relevance is greater than the specified threshold are considered for classification. However, some attributes may have uncertainty coefficient very near but less than the specified threshold and may have influence on the dependable variable. The method ignores such attributes. The user intervention at this stage can help in selecting the attributes. This will eventually add to the construction of more accurate decision tree.

Results

The variables of unknown characteristics defined in the study are commonly found in construction projects. Potential areas where the methodology can be used are, analysis Equipment out put analysis, Labour productivity analysis, Delays control: Pattern searching – e.g. “the activity that has a pattern of 50% probability of delay.”

Analysis of labour productivity has been selected for the application of the proposed methodology. The labors of different types, work on various construction projects road,

building, bridges etc. The labors also work on various activities like brick work, concrete, reinforcement etc., which are of different nature. Based on these assumptions following parameters that may influence the labor productivity are identified.

1. Project type: building, road, bridge, jetty, power, plant, railway, etc.
2. Activity type: brick work, shuttering, plaster, concrete, fabrication, earthwork, excavation, formwork, foundation, scaffolding, slab etc.
3. Surrounding area of the project: metro, rural, urban
4. Location of the project: Karnataka, Maharashtra (here only the states have been considered).
5. Minimum and maximum temperatures during the day while the labor was working
6. Age groups of the labor: 18-25, 26-35, 36-50 and above 50
7. Height and depth of the place
8. Physical Mental over burden
9. Wages paid to the labor
10. Hours per day

Although the above listed parameters seem to have influence on the labor productivity, all of them may not influence in reality. Sometimes the multiple parameters together may influence the productivity. For example, Minimum and maximum temperatures during the same day can differ from locations to locations. The temperatures can be extreme in the places like Delhi on a given day and can be moderate in the places like Pune on the same day. Therefore the influence of such correlated parameters need to be calculated by combining them.

Data collection

Labor work data of 27 projects from various locations have been collected. A data collection form (ANNEXURE I) was designed and distributed to the site engineers of the respective projects and requested for filling the labor work records. The type of information collected was mainly related to project, activities and labours. Projects data include project type, locations, min/max temperatures, climatic conditions, height and depth of the place, physical stress, site management, work hours, overtime, and wages paid. Activity data include activity type, duration of the activity, number of skilled/unskilled labours used, sources of labours (local or outsourced), and total man days. Labour data include, age group, total work done by the labor on an activity, and labour productivity.

Data Standardization

Data was stored in a normalized relational database structure using Ms Access. The data was standardized, summarized, cleaned and organized into multidimensional model (Figure 2). Activities were classified as Activity Type (Table) 1. Total of 329 records of labour work data for an activity of a project were collected. **Masonry Brick Work** activity type recorded to be having maximum labour records of 80 and therefore was selected for the study as data set S (ANNEXURE II). The data warehouse model for the data set S is presented in Table 2.

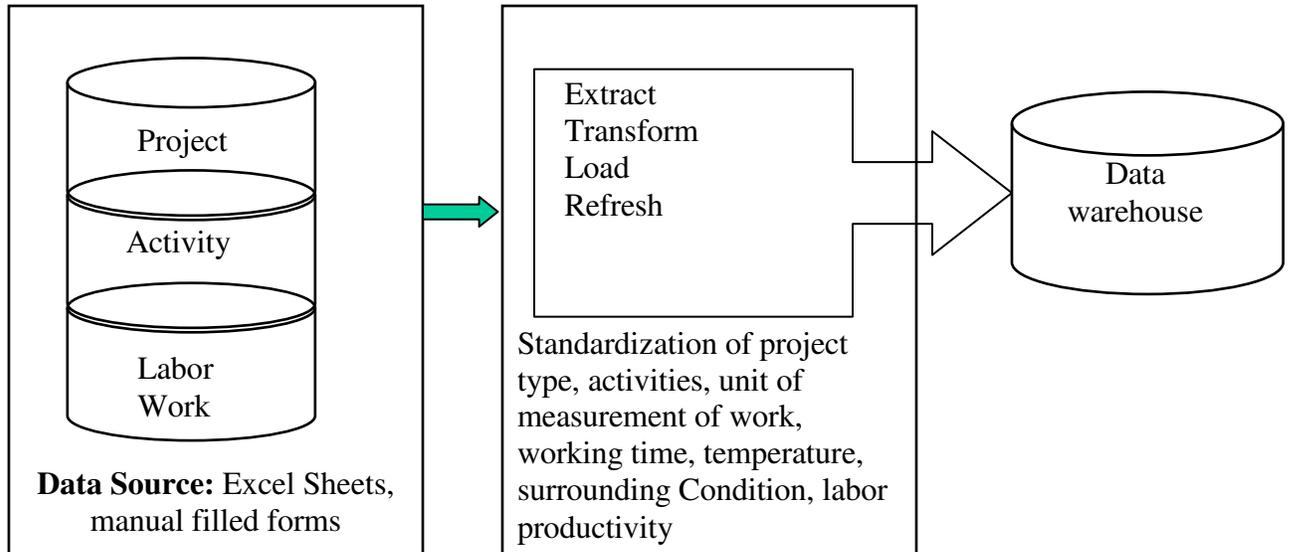


Figure 2: Data Warehouse Model

Table 1: Activity Types

Type of Activity	No. of Records
Masonry Brick Work	24
Reinforcement	23
Shuttering	22
Plaster	12
Concrete	6
Painting	3
Fabrication	2
Earthwork	1
Excavation	1

Table 2: Data Warehouse Model

Produ ctivity	Activity Type											
	Masonry Brick Work			Reinforcement			Shuttering			Plaster		
	Surrounding Area			Surrounding Area			Surrounding Area			Surrounding Area		
	Metro	Rural	Urban	Metro	Rural	Urban	Metro	Rural	Urban	Metro	Rural	Urban
>=1.44	4	15	6									
<1.44	3	17	35									

Probability Calculation

There are two class labels; 1) \geq average productivity, and 2) $<$ average productivity. The average productivity for the selected data set is 1.44 cum per day. The probability for each split class is thus calculated based on the number of outcomes of splitting attribute in each class label. For example: the sub sets of surrounding area such as metro, rural and urban have 7, 32, and 41 outcomes respectively. Out of total 7 outcomes of metro, 4 outcomes belong to class label ≥ 1.44 and 3 outcomes belong to the class label < 1.44 . Thus the probability for class labels ≥ 1.44 is calculates as $(100*4)/7 = 57.16 \%$ and the probability of class label, $< 1.44 = (100*3)/7 = 42.86$.

Step 1: Selecting the attribute at root node

The independent attributes such as Location, Age_Group, Min_Temperature, Max_Temperature, Climatic_Condition, Physical_Mental_Overburden, Surrounding_Area , and the dependent variable Labour_Productivity are selected. The task is to predict the influence of independent attributes over the dependable attribute. The dependable attribute Labour_Productivity is numerical. The average of Labour_Productivity is calculated to 1.44 cum per day. Let the S be the data set of Masonry Brick Work and has 80 outcomes in the entire data base. The class label here is Labour_Productivity ≥ 1.44 or Labour_Productivity < 1.44 . The number of outcomes in data set S for Labour_Productivity ≥ 1.44 is 25 and the number of outcomes in data set S for is Labour_Productivity < 1.44 is 55. The entropy of S is calculated as,

$$= -\frac{25}{80} \log_2 \frac{25}{80} - \frac{55}{80} \log_2 \frac{55}{80} = 0.896$$

The data set S has sub sets $S1, S2, S3, S4$ for Maximum and Minimum Temperature, Surrounding Area, Physical Mental Overburden, Climatic Condition, Age Group,

Location respectively. The sub sets $S1$, $S2$, $S3$, $S4$, $S5$, and $S6$ have sub sets as presented in Table 3. Min_Temperature and Max_Temperature are two different attributes.

Combination of these attributes may have more influence on Labour_Productivity and therefore sub set $S1$ represents the combined set of Maximum and Minimum Temperature.

Table 3: Sub sets of $S1$, $S2$, $S3$, $S4$, $S5$, $S6$

S1	S11	S12	S13	S14	
	Min_Temperature ≤ 19 and Max_Temperature ≥ 39	Min_Temperature > 19 and Max_Temperature < 39	Min_Temperature ≤ 19 and Max_Temperature ≤ 39	Min_Temperature > 19 and Max_Temperature > 39	
S2	S21 metro	S22 rural	S23 urban		
S3	S31 more	S32 medium	S33 less		
S4	S41 normal	S42 good	S43 extreme		
S5	S51 18-25	S52 26-36	S53 35-50	S54 Above 50	
S6	S61 AP	S62 Haryana	S63 Karnataka	S64 Maharashtra	S65 Jammu Kashmeer

Combination of Min_Temperature and Max_Temperature to form a single test split

The split classes formed by combining Minimum and Maximum Temperature are, 1) Min_Temperature ≤ 19 and Max_Temperature ≥ 39 – $S11$, 2) Min_Temperature > 19 and Max_Temperature < 39 – $S12$, 3) Min_Temperature ≤ 19 and Max_Temperature ≤ 39 – $S13$, 4) Min_Temperature > 19 and Max_Temperature > 39 – $S14$. Entropy for $S1$ { $S11$, $S12$, $S13$, $S14$ } is calculated as shown in Table 4.

Table 4: Weighted average entropy & gain for $S1$ (Max. and Min. Temperature)

Entropy Calculation for sub set $S11$, $S12$, $S13$, $S14$				
Sets	Total	outcomes of Productivity Class	Entropy	for weighted

	outcomes	≥ 1.44 (average)	< 1.44 (average)		average entropy
Min_Temperature ≤ 19 and Max_Temperature ≥ 39	38	9	29	0.790	0.375
Min_Temperature > 19 and Max_Temperature < 39	29	13	16	0.992	0.360
Min_Temperature ≤ 19 and Max_Temperature ≤ 39	12	3	9	0.811	0.122
Min_Temperature > 19 and Max_Temperature > 39	1	0	1	0.000	0.000
weighted average entropy					0.857
gain					0.040

$$\text{Where weighted average entropy} = \frac{38}{80} \times 0.790 + \frac{29}{80} \times 0.992 + \frac{12}{80} \times 0.811 + \frac{1}{80} \times 0 = 0.857$$

$$\text{Gain for S1} = \text{entropy (S)} - \text{weighted average entropy (S1)} = 0.896 - 0.857 = 0.040$$

Same way the weighted average entropy and gain for S2, S3, S4, S5 and S6 are computed

(Table 5, 6, 7, 8 and 9)

Table 5: Weighted average entropy and gain for S2 (Surrounding Area)

Entropy Calculation for sub set S21, S22, S23					
Sets	Total outcomes	outcomes of Productivity Class		Entropy	for weighted average entropy
		≥ 1.44 (average)	< 1.44 (average)		
metro	7	4	3	0.985	0.086
rural	32	15	17	0.997	0.399
urban	41	6	35	0.601	0.308
weighted average entropy					0.793
gain					0.103

Table 6: Weighted average entropy and gain for S3 (Physical Mental Overburden)

Entropy Calculation for sub set S31, S32, S33					
Sets	Total outcomes	outcomes of Productivity Class		Entropy	for weighted average entropy
		≥ 1.44 (average)	< 1.44 (average)		
more	20	4	16	0.722	0.180
medium	48	15	33	0.896	0.538
less	12	6	6	1.000	0.150
weighted average entropy					0.868
gain					0.028

Table 7: Weighted average entropy and gain for S4 (Climatic Condition)

Entropy Calculation for sub set S41, S42, S43					
Sets	Total outcomes	outcomes of Productivity Class		Entropy	for weighted average entropy
		≥ 1.44 (average)	< 1.44 (average)		
normal	60	16	44	0.837	0.627
good	8	4	4	1.000	0.100
extreme	12	5	7	0.980	0.147
weighted average entropy					0.874
gain					0.022

Table 8: Weighted average entropy and gain for S5 (Age Group)

Entropy Calculation for S51, S52, S53, S54					
Sets	Total outcomes	outcomes of Productivity Class		Entropy	for weighted average entropy
		≥ 1.44 (average)	< 1.44 (average)		
S1 - 18-25	24	7	17	0.871	0.261
S2 - 26-25	24	8	16	0.918	0.275
S3 - 35-50	23	8	15	0.932	0.268
S4 - Above 50	9	2	7	0.764	0.086
weighted average entropy					0.891
gain					0.005

Table 9: Weighted average entropy and gain for S6 (Location)

Entropy Calculation for sub set S61, S62, S63, S64, S65					
Sets	Total outcomes	outcomes of Productivity Class		Entropy	for weighted average entropy
		≥ 1.44 (average)	< 1.44 (average)		
AP	4	0	4	0.000	0.000
Haryana	3	0	3	0.000	0.000
Karnataka	39	16	23	0.977	0.476
Maharashtra	32	8	24	0.811	0.325
Jammu Kashmir	2	1	1	1.000	0.025
weighted average entropy					0.826
gain					0.070

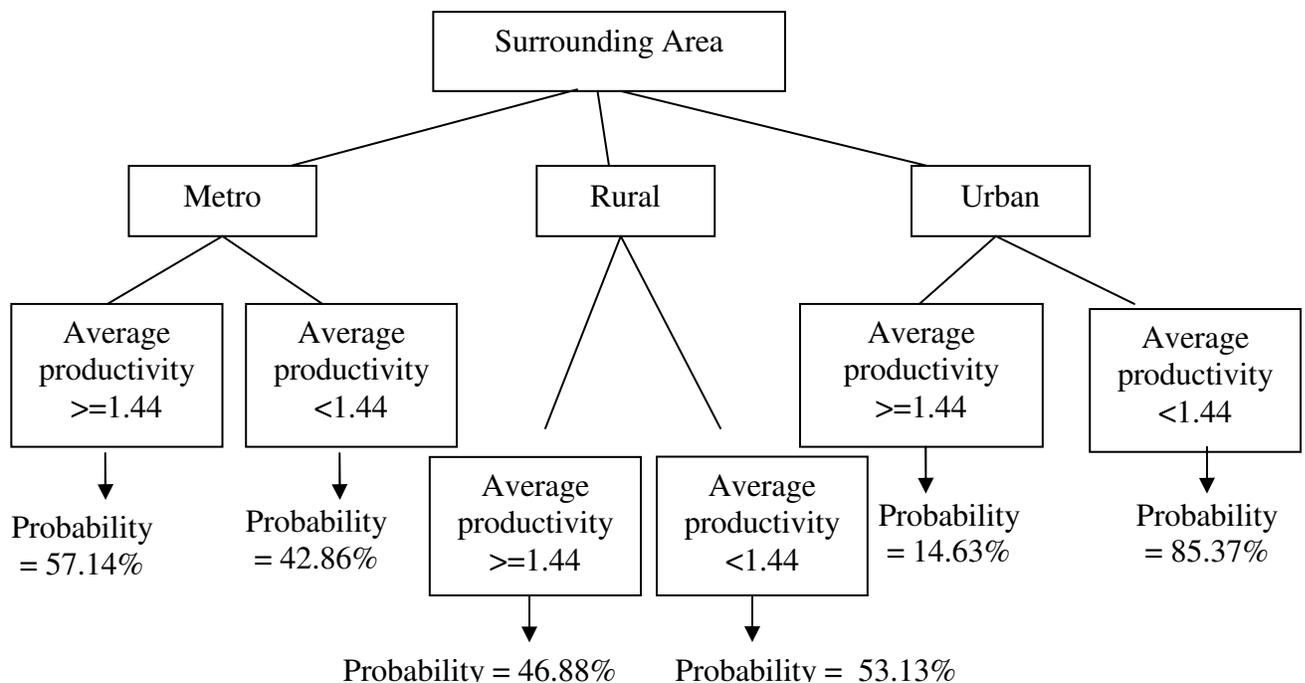
Comparison of the Gains

The gains sorted in descending order and uncertainty coefficient is calculated (Table 10).

Table 10: Comparison of Gain

Ranking	Data Set	Attribute	Gain	Entropy	Uncertainty Coefficient (gain/entropy)
1	S2	Surrounding Area	0.103	0.793	0.12
2	S6	Location	0.070	0.826	0.08
3	S1	Maximum and Minimum Temperature	0.040	0.857	0.05
4	S3	Physical Mental Overburden	0.028	0.868	0.03
5	S4	Climatic Condition	0.022	0.874	0.02
6	S5	Age Group	0.005	0.891	0.01

Data set S2 of Surrounding Area has maximum gain of 0.103 and ranked as 1 and other data sets such as S6, S1, S3, S4 and S5 are ranked as 2,3,4,5, and 6 respectively. At this stage the attribute, surrounding Area is selected to be placed at the root node. The tree constructed at this stage is presented in Figure 3(a).

**Figure 3: Decision tree (a)**

Data set S6 has second highest gain and is selected as branches to the root node. Gain for AP and Haryana (Table 9) is zero and are therefore dropped. Remaining three sets,

Karnataka, Maharashtra, Jammu Kashmir are selected for construction of tree further. A split test that has zero outcomes is automatically dropped. The Metro and Rural have, Karnataka and Maharashtra branches respectively; and Urban has, Karnataka and Maharashtra branches. The tree constructed at this stage is presented in Figure 3(b).

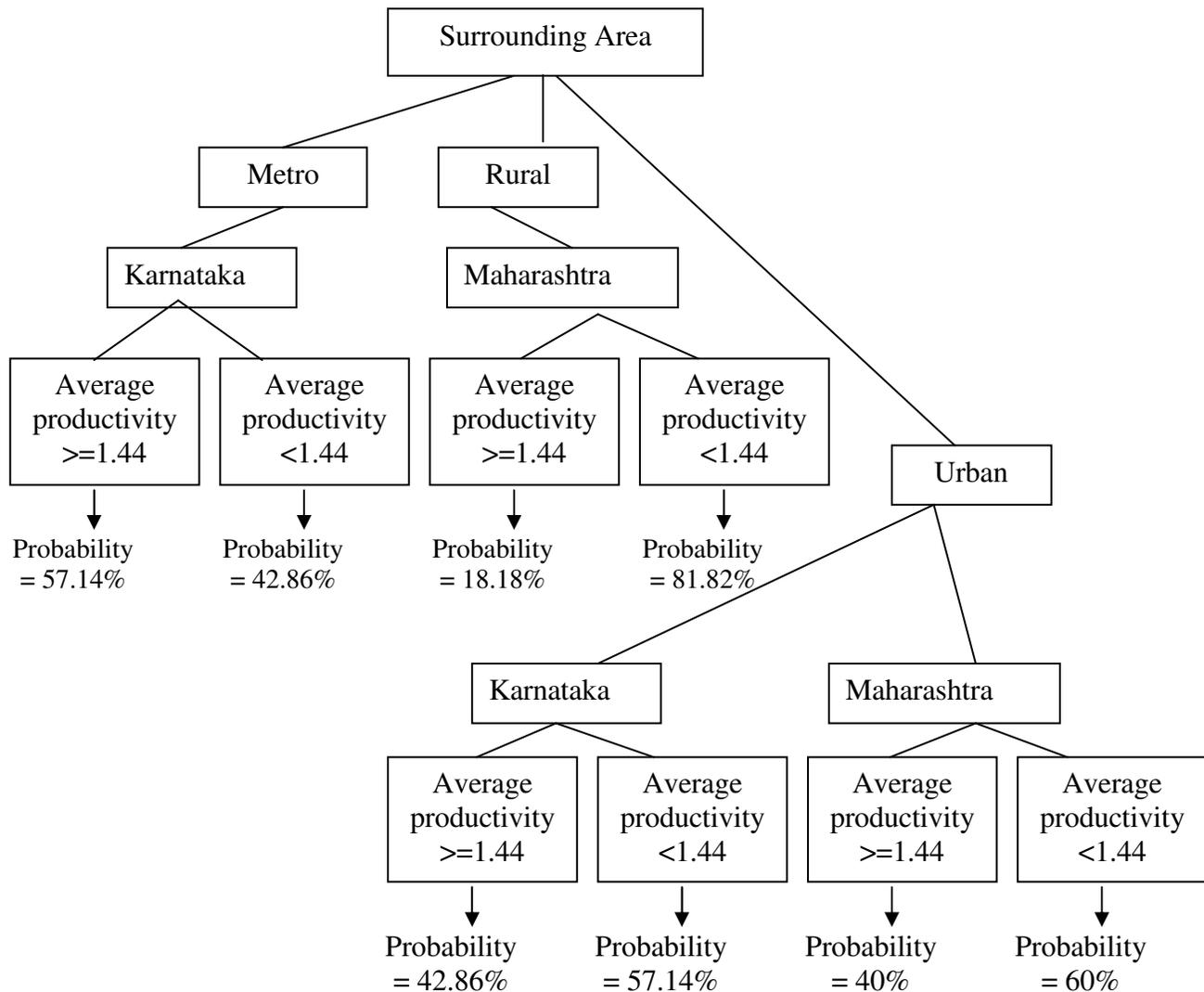


Figure 3: Decision tree (b)

The data set S1 has third highest gain (Table 10). The last row of the table shows the number of outcomes under each class values T1, T2, T3 for location for surrounding

area (Table 11). The reliability analysis is carried out by computing the uncertainty coefficient (UC) (Table 10). The average of uncertainty coefficient is 0.05. The data sets, S3, S4, S5 (Physical Mental Overburden, Climatic Condition, Age Group) are having UC ≤ 0.05 and therefore ignored. The final decision tree is the one presented in Figure 3 (c1 and c2).

Table 11: Data classification for activity type – Masonry Brick Work

Activity Type																		
Masonry Brick Work (80)																		
Surrounding Area																		
Metro						Rural						Urban						
Location						Location						Location						
Maharashtra			Karnataka			Maharashtra			Karnataka			Maharashtra			Karnataka			
Produ ctivity	Maximum and Minimum Temperature																	
	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
≥ 1.44	0	0	0	0	4	0	4	0	0	0	0	0	4	0	0	0	9	3
< 1.44	0	0	0	0	3	0	14	0	3	4	0	0	3	3	0	0	10	6

T1 = Min_Temperature ≤ 19 and Max_Temperature ≥ 39 (S11),

T2 = Min_Temperature > 19 and Max_Temperature < 39 (S12),

T3 = Min_Temperature ≤ 19 and Max_Temperature ≤ 39 (S13)

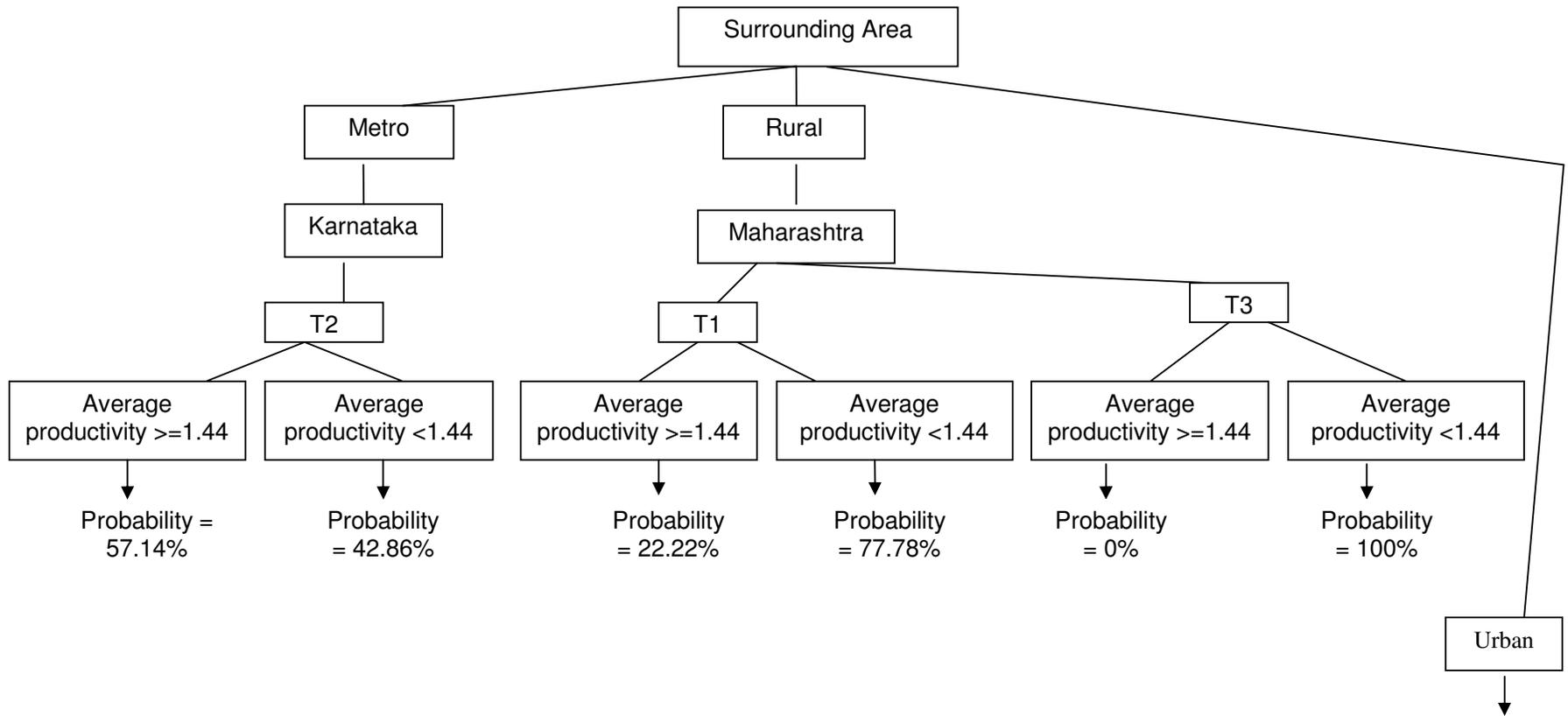


Figure 3: Decision tree (c1)

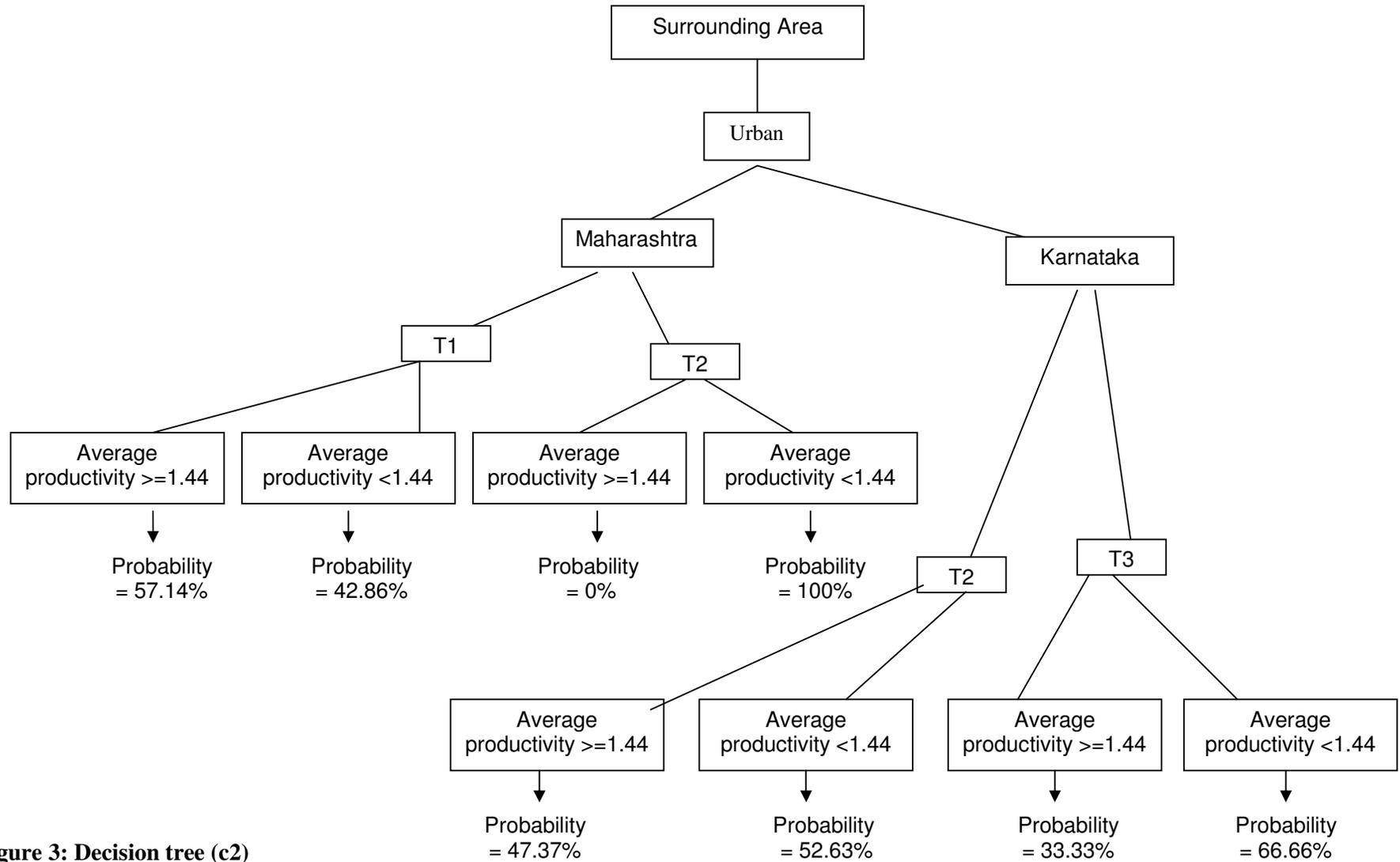


Figure 3: Decision tree (c2)

Interpretation of the decision tree data analysis

The meaning derived from decision tree presented in Figure 3 (a) is 1) Chances of productivity becoming more than the average productivity are more in metro (57.14%), and less in Rural (46.88). The same is low in urban (14.63%). The tree further grows by adding the location to the surrounding area (Figure 2 b). The interpretation is, 1) Chances of productivity getting more then the average productivity are more in **metro Karnataka** (57.14%) than the **urban Karnataka** (42.86%). 2) Chances of productivity getting more then the average productivity are very low in **rural Maharashtra** (18.18%) then in **urban Maharashtra** (40%). The tree further grows by adding the class of combined attributes, Maximum Temperature and Minimum Temperature and is interpreted as (Figure 3 c1, c2), 1) Chances of productivity getting more then the average productivity are higher in **metros of Karnataka** (57.14%) during **normal** (T2) temperature then the **urban Karnataka** (47.37%). 2) Chances of productivity getting more then the average productivity are lower in **rural Maharashtra** (22.22%) during **extreme** (T1) temperature then the **urban Maharashtra** (57.14%). Every branch presents a pattern as to how labor productivity gets influenced by various other parameters.

Conclusions and further research

The approach remains same irrespective of the type of activity. The proposed approach can be applied to any activity and the patterns can be predicted. The methodology facilitates analysis of attributes which are of unknown characteristics more efficiently than the available methods. It can be applied directly on real data, reducing the dependency on training data set. It allows User's intervention, at variable selection; threshold definition improves the

performance of the decision tree and, facilitates combining multiple variables for classification

There can be an issue related to combining the categorical attributes which can be taken for further study. The software can be developed by adopting the proposed method.

REFERENCES

- Ahmad I. and Nunoo C., Data Warehousing in the Construction Industry: Organizing and Processing Data for Decision-Making, *Data Warehousing in Construction, Department of Civil and Environmental Engineering*, Florida International University, Miami, Florida, USA.
- Ahmad Irtishad and Azhar Alman, (2005), Data Warehousing in Construction, Advancing Engineering, Management and Technology, *Third International Conference in the 21st Century (CITC-III)*, Athens.
- Almullim Hussein and Kaneda Shigeo and Akiba Yasuhiro, (2002), Development and Applications of Decision Trees, *Expert Systems*, Vol. 1, pp.53-77.
- Caldas H. Carlos H. and Soibelman Lucio, (2002), Automated Classification Methods: Supporting the Implementation of Pull Techniques for Information Flow Management, *Proceedings IGLC-10*, Gramado, Brazil.
- Dan Vance and Ralescu Anca, The Hyperplane Algorithm, A Decision Tree Using Oblique Lines, *ECECS Department University of Cincinnati*.
- Gatzu Stella and Vavouras Athanasios, (1999), Data Warehousing: Concepts and Mechanisms, *Informatique*, Vol 1, pp.8-11.
- Hand David, Mannila Heikki, Smyth Padhraic, (2004), Principles of Data Mining, *MIT Press*, Cambridge, MA, USA, pp. 141-157, ISBN-81-203-2454-9.
- Leo Breiman, Jerome Friedman, Charles J. Stone, R. A. Olshen, (1998), Classification and regression trees, *Boca Raton, Fla., Chapman & Hall/CRC*.
- Pujari K. Arun, (2001), *Data Mining Techniques*, University Press (India) Pvt. Ltd., Hederabad, pp 7-68, ISBN 81 7371 380 4.
- Treiguts Egmonts, (2002), Constructing New Attributes for Algorithms of Decision Trees Induction, Information Technology and Management Science, *Scientific Proceedings of Riga Technical University*.

Zhang Jun, Silvescu Adrian, and Honavar Vasant, (2002), Ontology-Driven Induction of Decision Trees at Multiple Levels of Abstraction, *SARA 2002*, LNAI 2371, pp.316-323.

ANNEXURE I: Data Collection Form with sample data

NAME OF PROJECT	:ABC									
LOCATION	:Maharashtra									
FACILITIES PROVIDED	:Proper conveyance,Accomodation,driking water facility,Medical facility etc.									
TEMPERATURE VARIATION	:Max-42 degree centigrade,Min-15 degree centigrade, Ave-28degree centigrade (for maximum working period)									
CLIMATIC CONDITION(SEVIER,NORMAL.GOOD)	:good									
HEIGHT/DEPTH OF WORK	:Max hieght-13 mt, Max depth-7mt									
PHYSICAL/MENTAL OVERBURDEN(MORE/MEDIUM/LESS)	:more									
SITE MANAGEMENT(LABORS HANDLING EFFICIENCY), 100% IS MAX	:80% As local labors are dominant ,tendency towards work is good									
WORKING HRS/DAY	:12 Hrs									
EXTRA O.T RATE FOR SKILLED/UNSKILLED LABORS	:4 hrs per day									
% OF LOCAL LABORS	:30%									
SURROUNDING AREA (RURAL/URBAN/METRO)	:Rural & suburban									
SPECIALITY IN WORK(eg.SHUTTERING,BARBENDING etc.)	: Reinforcement, Shuttering,Concreting,fabrication									
WEDGES PAID AS PER RULE FOR THAT AREA	:1)Skilled--225.0 Rs 2)Unskilled--100.0 Rs									
Activity: Brick work masonry for wall										
AGE GROUP	Activity Duration	Number of labours per day		Source of Labours		Type of Labour	unit of work	total work done	hours of work per day	wages paid per day
		Skilled	Unskilled	Local	Outside					
18-25	1 day	2	2	2	2	helpers,mason	cum	5.5cum	11	Skilled-225, Unskilled-100
26-35	1 day	2	2	2	2	helpers,mason	cum	6 cum	11	Skilled-225, Unskilled-100
35-50	1 day	2	2	2	2	helpers,mason	cum	6 cum	11	Skilled-225, Unskilled-100
ABOVE 50	1 day	2	2	2	2	helpers,mason	cum	4.5 cum	11	Skilled-225, Unskilled-100

ANNEXURE II: Number of outcomes for activity type: Masonry Brick Work

Case ID	Location	Age_Group	Labour Productivity	Min Temperature	Max Temperature	Climatic Condition	Physical_Mental Overburden	Sorrounding Area
1	Karnataka	18-25	0.300	20	35	normal	medium	metro
2	Karnataka	18-25	0.500	18	36	normal	less	urban
3	Maharashtra	18-25	1.375	15	42	good	more	rural