IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IX (Mar-Apr. 2014), PP 74-81
www.iosrjournals.org

A Multi-Relational Decision Tree Learning (MRDTL) Approach: A Survey

Patel Rinkal[1] , Rajanikanth Aluvalu[2]
M.Tech Student Computer Engineering Department, R.K University Rajkot[1]
Computer Engineering Department, R.K University Rajkot[2]

Abstract: *Data Mining is the process of extracting useful knowledge from large set of data. There are number of data mining techniques available to find hidden knowledge from huge set of Data. Among these techniques classification is one of the techniques to predict the class label for unknown data based on previously known class labeled dataset. Several classification techniques like decision tree induction, Naivy Bayes model, rough set approach, fuzzy set theory and neural network are used for pattern extraction. Now a day's most of the real world data stored in relational database but the decision tree induction method is used to find knowledge from flat data relations only, but can't discover pattern from relational database. So to extract multi-relational pattern from relational tables we use MRDTL approach. In real world Missing value problem are common in many data mining application. This paper provides survey of multi-relational decision tree learning algorithm to discover hidden multi-relational pattern from relational data sets and also includes some simple technique to deal with missing value.*

***Keyword**: Data Mining, Multi-relational Data Mining Framework, Multi-relational Decision Tree Learning (MRDTL), Relational Database, Missing Value*

I. Introduction

In today's world structured data is stored in relational databases. Many important classification approaches, such as neural networks and SVM technique, can only be applied to data represented by single flat data relations. And it is very difficult to convert a multi-relational database into a single flat relation without losing important information. The developments of high throughput data achievement, digital storage, and communications technologies have made it possible to collect very large amounts of data in many scientific and commercial domains. Most of this data are stored in multiple relations. So, the task of learning from relational database has begun to receive important attention [1].

A. What is Data Mining?

Data mining is the process of extracting and finding patterns from huge data sets by combining methods from statistics and artificial intelligence. Data mining consist of different techniques like classification, clustering, prediction, outlier analysis etc. for finding hidden knowledge. Classification [16] predicts categorical class label and construct model based on training set data which contain class label. This model represented as classification rule. Classification [12] is supervised learning.

Data mining [12] has a variety of fields which provides the different tools and the techniques for handling the large database. Through this technique we will obtain the new, valuable non-trivial and existing information. KDD and Data Mining sometimes are used imprecisely. A more recent convention mentioned in [18] (Blockeel, 1998) establishes that the process of knowledge discovery actually contains of three subtasks. The first task is to adjust the original format of the data to fit the input format of the data mining algorithm (called *Preprocessing of data*). Once the data are formatted, one or more algorithms must be applied to find out patterns, regularities or general laws from the data, this is the phase called *Data Mining*. Once the results of the data mining process are obtained they may require to be translated to a more understandable format. This last stage is known as post-processing of the results.

Data Mining Techniques are,

- Classification Technique
- Clustering Technique
- Association rule mining
- Prediction
- Outlier Analysis
- Characterization and Description[13]

B. What is Classification?

"It defines as mining patterns that can classify future data into known classes". Classification [12] is the most common learning task in data mining many methods. Decision trees, neural networks, support vector machines, Bayesian networks[16], etc.The algorithm is trained on some part of the data called training data and the accuracy tested on independent data (or use cross-validation)called test set data ,Optimization is related to many classification methods.

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Like CLS and ID3, C4.5 generates classifiers.[16]

Different Classification Methods

- Decision Tree Induction
- Naive Bayesian Classification
- Rule Based Classification
- Classification by Backpropagation
- Classification using Frequent pattern
- Other Classification Methods[13,16]

Application of Classification Techniques

- Target marketing
- detection of fraud
- Prediction of Performance
- Manufacturing
- Medical diagnosis[13,16]

C. Decision Tree Induction

Decision tree [13] is a predictive model that generates a tree from the given training samples. The tree construction process is heuristically guided by selecting the most relevant attribute at each step, whose aim is to minimizing the number of tests needed for classification.

Decision tree induction [14] is the process of learning a decision tree from class labeled training tupples.it is a flow chart like structure where each internal node of a tree denote a test on attribute, each branch represent outcome of the test, and each leaf node holds a class label, the top most node in the tree is called root node.

For a given tupple, X, for which the associated class label unknown, the attribute value of the tupple is tested against the decision tree .a path, is traced from the root to a leaf node, which holds the class prediction for that tupple. Decision tree can easily convert into classification rule.

The training set contains class labeled data and this data is used to classify new test set data One type into multi-relational classification is Inductive Logic Programming called ILP.[11]The ILP classification approaches targeted at searching hypotheses in particular format that can predict class labels of given data, depend on background knowledge. They achieve good classification accuracy in multi-relational classification. However, most ILP approaches are not scalable with respect to the large number of relations and attributes in the database [2].

The ILP classification approaches aim at finding hypotheses of certain format that can predict class labels of examples, based on background knowledge. They achieve good classification accuracy in multi-relational classification. However, most ILP approaches [11] are not scalable with respect to the number of relations and attributes in the database. [2]

Multi-relational data mining aim for discovering attractive information directly from multiple tables without joining data of multiple tables into a single table explicitly. It has been successfully applied in many application areas, such as marketing, sales, economics, fraud discovery, and natural sciences. Relational data mining methods used for searching patterns that occupy multiple tables from a relational database. One of method to apply traditional data mining (which assume that the data stored in a single table) is propositional, which converts multiple relational data into a single flat data relation, using joins and aggregations operation. However, it could direct to the generation of a huge, objectionable "universal relation" (including all of the attributes) [1].

The most apparent difference between a standard rule and a multi-relational rule [9] is that in case of a multi-relational rule, attribute names in conditions are annotated using the name of the relational table to which the attribute is related. However, the table name detail does not need to be presented to the user.

The induction of decision trees has been getting a lot of attention in the area of KDD [13] over the past few years. This reputation has been largely due to the efficiency with which the decision trees can be induced from large datasets, as well as to the well-designed and spontaneous representation of the knowledge that is discovered. However, traditional decision tree approaches have suffer from one drawback. Because of their propositional nature, they can't be used to analyze relational databases which containing multiple tables. This type of databases can be used to describe objects with some internal structure, which differentiate one object from the other one. The main ideas behind this description of objects in terms of occurrence of a substructure are simply not available in (attribute-value) decision trees [3].

Multi-relational Data Mining (MRDM) aims to extract useful patterns across multiple tables in a relational database. In general, a relational database consists of multiple interconnected tables, which are connected by means of foreign key relations. Since their first release in 1970s, relational databases have been routinely used to collect and organize many real-world data from financial transactions, medical records, to health informatics observation [4].

Decision tree induction [14] from a large relational database needs a framework with the following characteristics: [8]

1. Both attribute-value and structural information are included in the analysis.
2. The search space is significantly pruned by using the data model's constraints. This means that we are considering only the structural information that is intended by the design of the database.
3. The negation and complementary concept of objects are representable.Decision trees recursively divide the data set up into complementary sets of objects. It is necessary that both the positive split, as well as the complement of that, can represent efficiently.
4. Efficiency is achieved by a group of primitives that can be used to review both attribute-value and structural information.
5. The framework can be implemented by dedicated client/server based architecture.

II. Multi-relational Classification Methods

The earliest and most widely used approach for multi-relational classification is Inductive Logical Programming (ILP) [11]. The ILP is a classification approaches used for finding hypotheses of certain format that can predict the class labels of object, based on the background knowledge. in addition to ILP, probabilistic models are also very popular in multi-relational classification approaches. They differ from the Inductive Logic Programming approaches by specifying a probability distribution over a fixed set of random variables in the form of attributes-value pair. Recently Several new methods are planned.. Tupple ID propagation is a technique for performing virtual join, which largely improves efficiency of multi-relational classification. Instead of physically joining relations, they are virtually joined by attaching the IDs of target tuples to tuples in non-target relations [1].

The MRC algorithm [4] enables one to classify relational objects by applying conventional data mining methods, while there is no need to flatten multiple relations to a universal one.

A. Tuple ID Propagation

Suppose the primary key of the target relation is an attribute of integers, which represent the IDS of the target tuples. We use the ID of each target tuple to represent that tuple. We propagate the tuple IDs from Loan relation to Account relation. Or to say, for Each tuple *t* in the Account relation, we store the IDs of the target tuples associated with *t* by natural join.[17]

In fact tuple ID propagation is a way to do virtual join. Instead of physically joining the two relations, we virtually join them by attaching the tuple IDs of the target relation to the tuples of another relation. [2]

B. Multi-relational Data Mining Framework

The attribute-value pattern, which includes many common Data Mining algorithms, they only allow the analysis of simple objects. It requires that each object can be described by a fixed set of attributes and each of which can only have a single (unstructured) value [3].

Multi-relational data mining framework is based on the search for interesting patterns in the relational database, where multi-relational patterns can be viewed as "pieces of substructure occurred in the structure of the objects of interest"

We say that multi-relational object is covered by a multi-relational pattern if and only if the substructure described by the multi-relational pattern, in terms of attribute-value conditions and structural conditions, occurs at least once in the multi-relational object.Multi-relational patterns also can be viewed as subsets of the objects from the database having some property. The most interesting subsets are chosen according to some measure (i.e. information gain for Classification task), which guides the search in the space of all patterns [8].

Multi Relational Data Mining [5] as a term was initially used by in a way to describe a narrative approach for relational learning and knowledge discovery from relational databases and data consisting of complex objects. In multi-relational data mining framework, the data model consists of many tables; each one describing features of particular objects, only one view of the objects is central to the analysis. [14]

The user can analyze object by selecting one of the table known as target table. The point of importance is that each record in the target table will refer to a single object in the database. Once the target table has been selected, a particular descriptive attribute from that table can be selected for classification, this attribute is known as the target attribute within the target table. [7]
Relational Databases

A relational database consists of a set of tables denoted as T = {T1, T2 ...TN}, and a set of relations between pairs of tables. In each table a row is used to describe one record. A column is used to show values of some attribute for the records in the table.
Ex. An attribute A from table X is denoted by T.A [8].

Definition1: The domain of the attribute T.A is denoted as DOM (T.A) and is defined as the set of all different values that the records from table T have in the column of attribute A. Associations between tables are defined through primary and foreign key relations.

Definition2: A primary key attribute of table T, denoted as T.ID, has a unique value for each row in the table.

Definition3: A foreign key attribute in table Y referencing table T, denoted as Y.T_ID, takes values from DOM (T.ID).

C. Selection Graphs

To describe the constraints related to a multi-relational pattern[5], we introduce the concept of *selection graphs*: Selection graphs can be represented graphically and also called directed graphs.

Definition: A selection graph S is a directed graph represented as S (P, Q), where P is a set of Nodes (t, C, s), t is a table in the data model and C is a, possibly empty or set of conditions on attributes in t of type t.a operator c; the operator is one of the normal selection operators like, =, > etc. s is a flag has a values open and closed.

Q is a set of Edges (*r*, *s*, *a*, *e*) called *selection edges*, where r and *s* are selection nodes and *a* is an association between r.*t* and *s.t* in the data model. *e* is a flag with possible values *present* and *absent*. The selection graph contains at least one node *n0* that belongs to the target table *t0*.

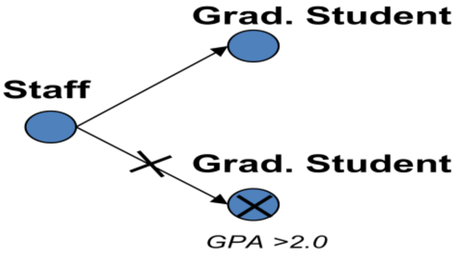

Figure1: Selection Graph

D. Multi-relational Patterns

For finding interesting knowledge from relational databases we want to consider not only attribute-value descriptions, as is common traditional algorithms, but the structural information also taken into consideration, which is available through the relations between tables. We will refer to descriptions of certain features of multi-relational objects as multi-relational patterns. We can look at multi-relational patterns as small pieces of substructure which we want to meet in the structure of the objects that we are considering.

For example [5], if we are considering a database of chemical compounds, a multi-relational pattern might describe a subgroup of the compounds by listing some of the elements and bonds between them.

*Definition1***:** The *Support* of a multi-relational pattern inside a relational database is the numbers of objects within this database are covered by the multi-relational pattern.

*Definition2***:** Patterns with a large support, generally above some predefined threshold, will be referred to as *frequent* [3].

III. Multi-relational Decision Tree Learning

The set of patterns derived from a relational database is potentially much bigger than the set of patterns which can be derived from a single table. Therefore, a lot of attention will have to be given to reducing the search space and to efficiently evaluating potentially interesting patterns. We will show how a lot of information which is stored in the data model can be used to prune the search space, and thus make the set of candidate patterns manageable. The candidate patterns that are valid according to the data model, and that will have to be evaluated, are sent to an efficient server which validates patterns against the data. [3]

Multi-relational data mining algorithms search for and successively filter out attractive patterns and select appropriate ones based on some impurity measure (e.g. information gain).

CrossMine uses a sequential covering algorithm, which repeatedly constructs clauses and removes positive examples covered by each clause. To construct a clause, it repeatedly searches for the best literal and appends it to the current clause. During the search process, CrossMine limits the search space to relations related to the target relation or related to relations used in the clause. In this way the strong semantic links can be identified and the search process is controlled in promising directions. [17]

Multi-relational decision tree learning [6] algorithm generate a decision tree whose nodes symbolize as multi-relational patterns i.e., selection graphs. The framework of MRDTL uses structured query language (SQL) to gather the information needed for generating classifiers (e.g., decision trees) from multi-relational data. Based on this framework, Leiva developed a multi-relational decision tree learning algorithm (MRDTL) [7].

MRDTL is based on the algorithm described by (Knobbe), and the logical decision tree induction algorithm called TILED [18] proposed by (H. Blockeel). TILDE uses first order logical clauses to represent nodes in the tree. MRDTL deals with records in relational databases, similarly to the TILDE approach. MRDTL adds selection graphs as nodes to the decision tree through a process of successive refinement until some termination condition is reached [1].

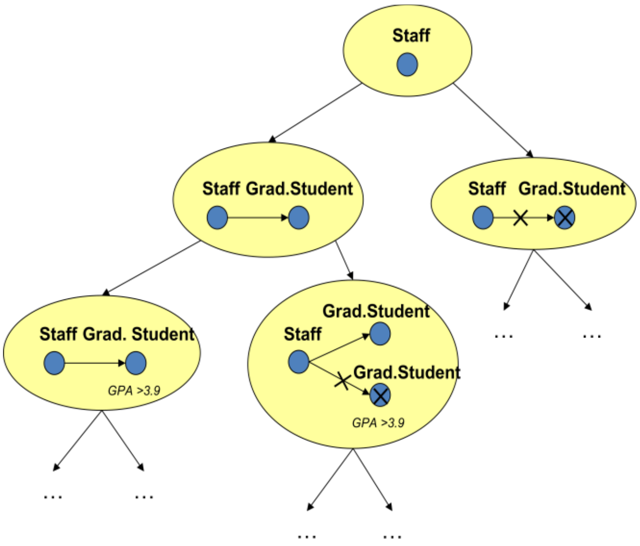

Figure2: Multi-relational Decision Tree

Multi-relational decision tree learning algorithm constructs a decision tree whose nodes are Multirelational patterns i.e., selection graphs. MRDTL-2 algorithm was improvement over MRDTL proposed by (Leiva) and the logical decision tree induction algorithm called TILDE [18] algorithm proposed by ((H. Blockeel)). TILDE uses first order logic rules to represent nodes in the tree, when data are represented in first order logic. MRDTL deals with records in relational databases, similarly to the TILDE's approach. Basically, MRDTL adds selection graphs as the nodes to the decision tree through a process of successive refinement until some termination criterion is reached [5].

Top-down induction of decision trees is basically a Divide and Conquer strategy. The algorithm starts with a single node at the root of the tree which represents the set of all objects in the relational database. By analyzing all possible refinements of the empty selection graph, and investigating their quality by applying some interestingness measure, we determine the optimal refinement. These optimal refinements, together with its complement, are used to generate the patterns associated with the left and the right branch of the tree. Based on the stopping condition it may go to the optimal refinement and its complement do not give reason for further splitting. Whenever the optimal refinement does provide a good split criteria, a left and right branch are introduced and the procedure is applied to each of these recursively.

MRDTL Algorithm

```
TREE INDUCTION (D, G)
Input Database D, selection graph G
Output The root of the tree, T
P: = optimal refinement (G)
if stopping criteria (G)
return leaf
else
Tleft: = TREE INDUCTION (D, P (G))
Tright: = TREE INDUCTION (D, P (G))
return node (Tleft, Tright, P) [5, 6]
```

Algorithms for implementing MRDTL

(1) FOIL-First Order Induction of Logical Decision tree
(2) TILDE-Top-Down Induction of Logical Decision Tree
(3) CrossMine[17]
(4) MRC-Multi-relational Classification Algorithm[4]
(5) MRDTL Software[6]

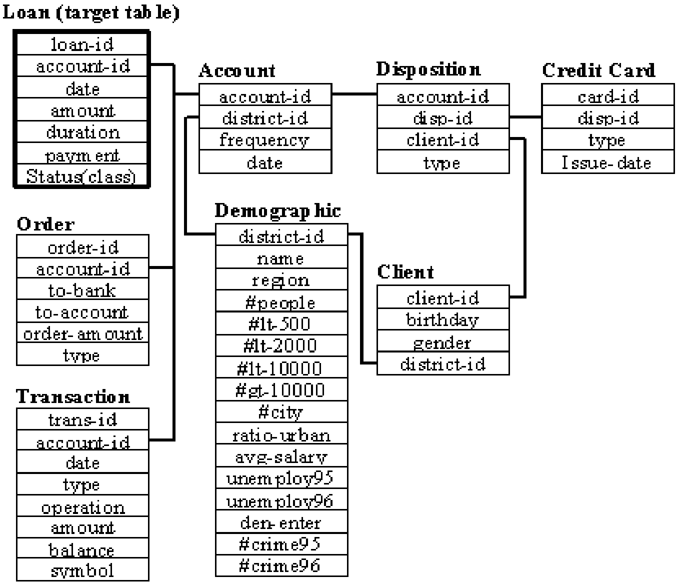

Figure3: A simple database from PKDD 99

A. *Refinements of Selection Graph*

once the target table and the target attribute have been selected (i.e. the kind of objects central to the analysis have been completely defined) a number of possible *refinements* can be made to the initial node that represents *T0* and successive nodes in order to find a hypothesis to be steady with the data in the training database.

Steps for Refinement [6]

(1) Add positive condition. This refinement stage will simply add a condition to a selection node in Graph without changing the structure of *G*.

(2) Add negative condition. Suppose a node which is refined does not correspond to the target table, this refinement will introduce a new absent edge from the parent of the selection node.

(3) Add present edge and open node. This refinement will begin an association in the data model as a present edge together with its corresponding table and add these to *Graph*.

(4) Add absent edge and closed node. This refinement will begin an association in the data model as an absent edge together with its corresponding table and add these to *Graph [7]*.

B. What Missing-Value?

The issue before us is whether we have complete data from all research participants on all variables (at all possible time points, if it is a repeated-measures design). If any data on any variable from any participant is not present, the researcher is dealing with missing or incomplete data. For the purposes of the rest of this chapter, we use the term *missing* to indicate that state of affairs. [15] Legitimate missing data can be dealt with in different ways. One common way of dealing with this sort of data could be using analyses that do not require (or can deal effectively with) incomplete data.

Handling missing values

The current implementation of MRDTL provides a simple method for dealing with missing attribute values [10] in the data. We construct Naive Bayes model [15] for each attribute in a table based on the other attributes (excluding the class attribute). Missing attribute [9] values are 'filled in' with the most possible value predicted by the Naive Bayes predictor for the corresponding attribute. The issue with missingness is that nearly all classic and modern statistical techniques assume (or require) complete data and most common statistical packages default to the least desirable options for dealing with missing data: deletion of the case from the analysis.[15]

(1) Discard instances: discarding instances with missing values is an approach often taken by researchers wanting to evaluate the performance of a learning method. For such an evaluation, this strategy is appropriate when the features are missing completely at random.

(2) Acquire missing values: a missing value may be accessible by incurring a cost, such as the cost of performing a analytical test or the cost of getting consumer data from a third party.[15]

(3) Imputation: imputation is a class of methods by which evaluation of the missing value or of its distribution is used to generate predictions from a given model. In particular, either a missing value is replaced with an estimated value or alternatively the distribution of possible missing values is estimated and analogous model predictions are combined probabilistically.[10]

(4) Reduced-feature Models: Imputation is required when the model being applied to employs an attribute whose value is missing in the test instance. An alternative approach is to apply a different model—one that include only attributes that are known for the test instance.[10,15]

C. Shortcomings of MRDTL

(1) *Slow Running Time***:** MRDTL based on the multi-relational data mining framework employs selection graphs in order to query the databases and for getting the information required for generating the classifier.

(2) *Unable to Handle Missing Values Attribute:* In many multi-relational databases used in many real-world applications of data mining, a considerable portion of the data has one or more missing value attribute.[5]

IV. Conclusion

Multi-relational classification is a very major area of research because of the attractiveness of relational database. in addition to classification, there are many other important tasks in relational databases, such as object matching, schema matching, and data cleaning .Unfortunately most existing approaches of multi-relational learning are not scalable when the number of relations and the complexity of the database schema are more. The main advantage of using standard decision tree algorithms is the increase in expressiveness. The main advantage of using the ILP approach towards decision trees is the increase in efficiency achieved by exploiting the domain knowledge present in the data model of the database. One of the main outstanding challenges is to expand this framework such that the selection graphs may also contain cycles.

References

[1] Jing-Feng Guo, Jing Li and Wei-Feng Bian,” An Efficient Relational Decision Tree Classification Algorithm”, College of Information Science and Engineering, Anshan University, China..

[2] Xiaoxin Yin, Jiawei Han and Jiong Yang,” Efficient Multi-relational Classification by Tuple ID Propagation”, Department of Computer Science University of Illinois at Urbana-Champaign.

[3] Arno J. Knobbe, Arno Siebes, Daniël van derWallen,” Multi-Relational Decision Tree Induction”, 3821 AE Amersfoort the Netherlands.

[4] Hongyu Guo, Herna L. Viktor,” Multirelational Classification: A Multiple View Approach”, School of Information Technology and Engineering, University of Ottawa, Ontario, Canada, Jan 19, 2008.

[5] Anna Atramentov,” Multi-relational decision tree algorithm -implementation and experiments”, Graduate College Iowa State University, 2003.

[6] Anna Atramentov, Hector Leiva, and Vasant Honavar,” Experiments with MRDTL -- A Multi-relational Decision Tree Learning Algorithm”, Artificial Intelligence Laboratory Department of Computer Science and Graduate Program in Bioinformatics Iowa State University Ames, IA 50011, USA

[7] Vaibhav Tripathy,” A Comparative Study Of Multi-Relational Decision Tree Learning Algorithm”, International Journal of Scientific and Technology Research Volume 2, Issue 8, August 2013.

[8] Anna Atramentov, Hector Leiva, and Vasant Honavar, ” A Multi-relational Decision Tree Learning Algorithm - Implementation and Experiments” , Artificial Intelligence Research Laboratory Computer Science Department, Iowa State University Ames, IA 50011-1040, USA.

[9] Mahmut Uludag,” Multi-Relational Rule Discovery Progress Report” Eastern Mediterranean University November 2002.

[10] Maytal Saar-Tsechansky, Foster Provost,” Handling Missing Values when Applying Classification Models”, Journal of Machine Learning Research 8 (2007) 1217-1250.

[11] Dzeroski, S. Inductive Logic Programming and Knowledge Discovery in Databases, Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

[12] B.N. Lakshmi,G.H. Raghunandhan,"A Conceptual Overview of Data Mining”, Proceedings of the National Conference on Innovations in Emerging Technology-2011.

[13] Jiawei Han, Micheline kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006, pp 360-361.

[14] Rodrigo Coelho Barros, M´arcio Porto Basgalupp, Andr´e C. P. L. F. de Carvalho, and Alex A. Freitas "A Survey of Evolutionary Algorithms for Decision-Tree Induction"IEEE transaction .

[15] http://www.sagepub.in/upm-data/45664_6.pdf

[16] Raj Kumar, Dr. Rajesh Verma,"Classification Algorithms for Data Mining: A Survey”, International Journal of Innovations in Engineering and Technology.

[17] Xiaoxin Yin, Jiawei Han, Jiong Yang and Philip,"CrossMine: Efficient Classification across Multiple Database Relations”, University of Illinois at Urbana-Champaign, Urbana, IL 61801, US.

[18] Hendrik Blockeel, Luc De Raedt,"Top-down induction of first-order logical decision trees", Katholieke Universities Leuven, 1998.