ISSN: 2456-9348
Vol (03) _Issue (08)
Impact Factor: 4.520

iJETRM
International Journal of Engineering Technology Research & Management

DATA MINING SYSTEM FOR CLASSIFICATION UNIVERSITY OF COMPUTER STUDENTS' GRADES USING DECISION TREE ALGORITHM

Lai Lai Yee
Faculty of Information Science, University of Computer Studies (Mandalay)
Mandalay, Myanmar
Lai2yee@gmail.com

ABSTRACT

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses or other information repositories. This can be viewed as a result of the natural evolution of information technology. The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. This paper is data mining system for classification system for University students' grade using C4.5. Firstly, the input data are randomly partitioned into two independent data, a training data and a test data. And then two third of the data are allocated to the training data and the remaining one third is allocated to the test data. Final step is C4.5 Algorithm Process, the training data is used to derive C4.5 algorithm. Classification Process, test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable the rules can be applied to the classification of new data.

KEYWORDS: Classification, Data mining, Students' Grade, C4.5

INTRODUCTION

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. A knowledge discovery process include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining system can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used or the applications adapted.Data mining also known as Knowledge-Discovery Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc.

Data mining is iterative and interactive processes that explore and analyze voluminous data in order to discover valid, novel and meaningful patterns, associations or rules, using computationally efficient techniques. It is related to the sub area of statistics called exploratory data analysis, which has similar goals and relied on statistical measures and also closely related to the sub areas of artificial intelligence called knowledge discovery and machine learning.

Data mining has been attracted huge attention in numerous research communities due to its wide applicability in many areas such as retail industry, financial forecast, and decision support and intrusion detection. Data mining methods include associations, clustering, classification and prediction. One of the most important fields of the data-mining domain is the association mining.The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

There are four main objectives in this research paper. The first objective is to study classification theory under data mining system in details. The second objective is to know how decision trees were constructed by C4.5 algorithm. The third objective is to know C4.5 (based ID3) algorithm is applied to the mining of very large real-world databases. The fourth objective is to classify the exam result on the test data set. The last is to show many kinds of student' mark using computerized system.

Classification of Students' Grade

This paper claims that classification of students' data. The goal of this research paper is to classify student data by using decision tree classification method and to assess its accuracy by holdout method.

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Many classification and prediction methods have been proposed by researchers in machine learning, expert systems, statistics and neurobiology.

Classification is a two-step process. These two steps are model construction and model usage. Model construction describes a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction training set. The model is represented as classification rules, decision trees, or mathematical formulae.

Model usage is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over fitting will occur.

Data mining is an interdisciplinary field, the confluence of a set of disciplines including database systems, statistics, machine learning, visualization, and information science. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology.

WORK PLAN

The data set in this paper was obtained from University of Computer Studies , Mandalay. The data set has ten attributes: Name, Roll No, Myanmar, English, Subject 1, Subject 2, Subject 3, Subject 4, Subject 5 and Subject 6. Class label is Exam Result.

Classification and prediction methods can be compared and evaluated according to the following criteria.

(1)**Predictive accuracy**: This refers to the ability of the model to correctly predict the class label of new or previously unseen data.

(2)**Speed**: This refers to the computation costs involved in generating and using the model.

(3)**Robustness:** This is the ability of the model to make correct predictions given noisy data or data with missing values.

(4)**Scalability:** This refers to the ability to construct the model efficiently given large amounts of data.

(5)**Interpretability**: This refers to the level of understanding and insight that is provided by the model.

Classification is a preliminary data analysis step for examining a set of cases to see if they can be grouped based on „similarly" to each other. Data analysis methods vary on the way how they detect patterns. The ultimate reason for doing classification is to increase understanding of the domain or to improve predictions compared to unclassified data.

The decision tree method like the nearest neighbors method, exploits clustering regularities for construction of decision-tree representation. It shows implicitly which variables are more significant with respect to classification decisions. The decision tree learning method requires the data to be expressed in the form of classified examples.

The method of Memory Based Reasoning also called the nearest neighbor method finds the closet part analysis of the present situation and chooses the same solution such systems demonstrate good results in vastly diverse problems. Such systems do not create any models or rules summarizing the previous experience.Genetic algorithms are powerful technique for solution of various combinational or optimization problems. They are more an instrument for scientific research rather than a tool for generic practical data analysis. Nonlinear Regression Methods (NR) is based on searching for a dependence of the target variable in the form of function. This method has better chances of providing reliable solutions in medical diagnostics applications. Support Vector Machines methods are based on Structural Risk minimization principle from statistical learning theory. They generalize better than NR. Maximum Likelihood Estimation (MLE) deals with finding the set of models and parameters that maximizes this probability. This is generally satisfactory only if the probability of the chosen class is representative.

Decision tree is a classifier in the form of a tree structure, where each node is either a leaf node indicates the values of the target attributes (class) of examples or a decision node specifies test to be carried out on a single attribute value, with one branch and sub-tree for each possible outcome of the test.

A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance. Decision trees are powerful and popular tools for classification and prediction.

Decision tree generation consists of two phases. These two are tree construction and tree pruning.
Build the decision tree from the training set (conventional ID3). Convert the resulting tree into an equivalent set of rules. The number of rules is equivalent to the number of possible paths from the root to a leaf node. Prune each rule by removing any preconditions that results in improving its accuracy, according to a validation set. Sort the pruned rules in descending order according to their accuracy, and consider them in this sequence when classifying subsequent instances.Using training data to derive a classifier and then to estimate the accuracy of the classifier can result in misleading overoptimistic estimates due to overspecialization of the learning algorithm to the data. Holdout and cross-validation are two common techniques for assessing classifier accuracy, based on randomly sampled partitions of the given data.

Holdout method estimates the classifier performances in this system. The given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two thirds of the data are allocated to the training set and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set (Figure). The estimate is pessimistic since only a portion of the initial data is used to derive the classifier. Random subsampling is a variation of the holdout method in which the holdout is repeated k time. The overall accuracy estimate is taken as the average of the accuracies obtained from each iteration.

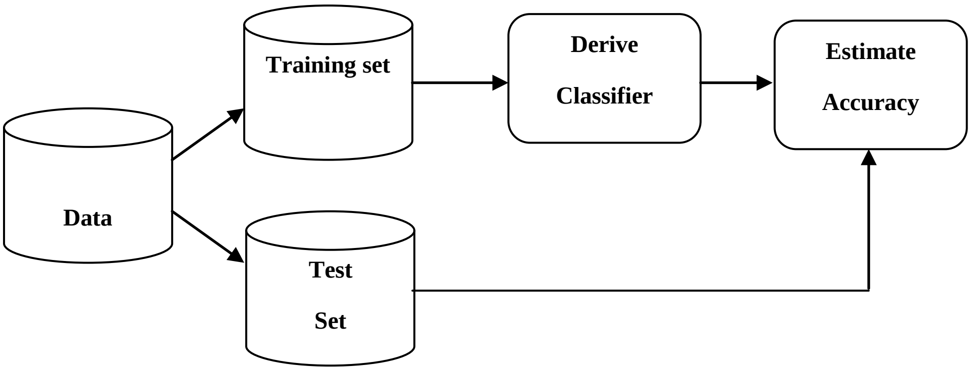


FIGURE: Estimating Classifier Performances with Holdout Method

PROPOSED SYSTEM

This paper shows that classification of toxicology data. The aim of this research is to classify toxicology data by using decision tree classification method and to assess its accuracy by holdout method.

There are two processes in Fig 3.2. Input data are randomly partitioned into two independent data, a training data and a test data. Typically two third of the data are allocated to the training data and the remaining one third is allocated to the test data. C4.5 Algorithm Process, the training data is used to derive C4.5 algorithm. Classification Process, test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable the rules can be applied to the classification of new data.

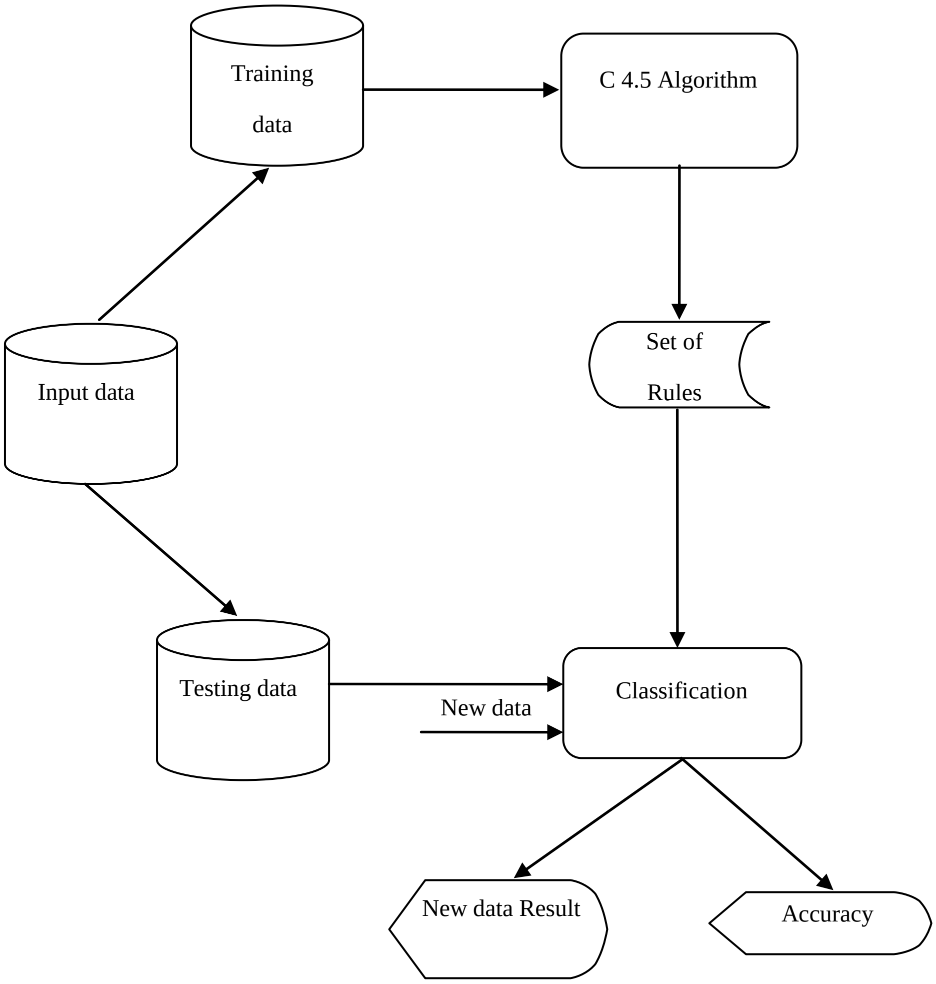


Figure: System Flow Diagram for Students' Grade

EXPERIMENTAL RESULTS

All marks in five groups: A,B,C,D,E.

Name	Description
A	(81-100) Marks
B	(61-80) Marks
C	(41-60) Marks
D	(21-40) Marks
E	(0-20) Marks

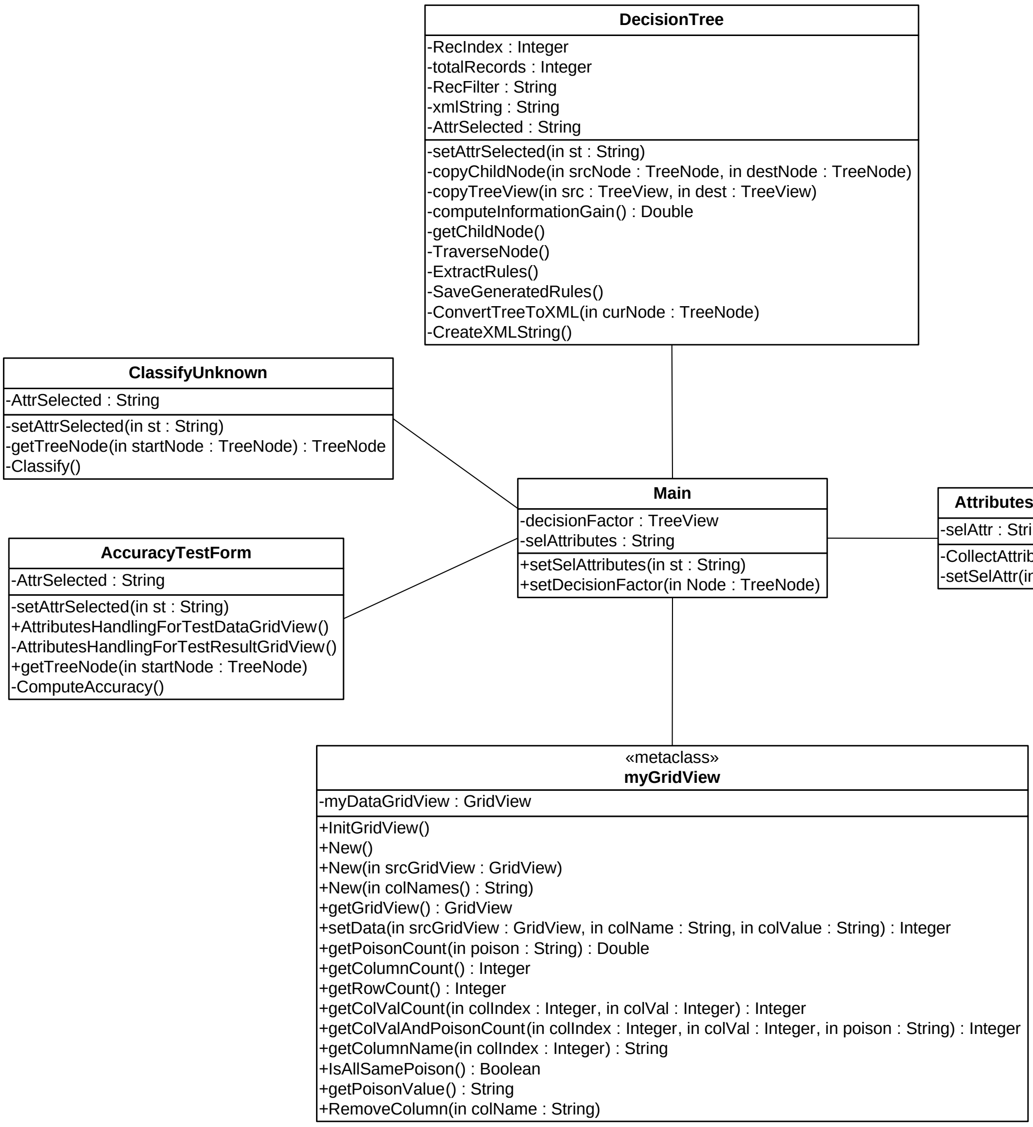
Exam Result has three Class Label.

Name	Description
Credit	Grade A,B
Pass	Grade B,C
Fail	Grade D,E

Training Data set for Student Exam Result

ID	STUDENT NAME	ROLL NO	MYANMAR	ENGLISH	SUB 1	SUB 2	SUB 3	SUB 4	SUB 5	SUB 6	EXAM RESULT
1	Mg Mg	3CS-3		B	B	A	B	A	A	A	Credit
2	Ma Ma	2CS-4		C	B	C	D	B	C	C	Fail
3	Su Su	2CT-2		B	C	B	B	C	B	B	Pass
4	Ag Ag	1CST-5	B	A	E	D	C	C	E	E	Fail
5	Aye Aye	1CST-1	A	B	B	A	A	A	A	A	Credit
6	Mu Mu	2CS-1		A	B	C	D	C	B	B	Fail
7	Swe Swe	4CS-10		C	C	E	B	B	C	C	Fail

Class Diagram of the System


DecisionTree
-RecIndex : Integer
-totalRecords : Integer
-RecFilter : String
-xmlString : String
-AttrSelected : String
-setAttrSelected(in st : String)
-copyChildNode(in srcNode : TreeNode, in destNode : TreeNode)
-copyTreeView(in src : TreeView, in dest : TreeView)
-computeInformationGain() : Double
-getChildNode()
-TraverseNode()
-ExtractRules()
-SaveGeneratedRules()
-ConvertTreeToXML(in curNode : TreeNode)
-CreateXMLString()
ClassifyUnknown
-AttrSelected : String
-setAttrSelected(in st : String)
-getTreeNode(in startNode : TreeNode) : TreeNode
-Classify()
Main
-decisionFactor : TreeView
-selAttributes : String
+setSelAttributes(in st : String)
+setDecisionFactor(in Node : TreeNode)
AttributesSelection
-selAttr : String
-CollectAttributes()
-setSelAttr(in st : String)
AccuracyTestForm
-AttrSelected : String
-setAttrSelected(in st : String)
+AttributesHandlingForTestDataGridView()
-AttributesHandlingForTestResultGridView()
+getTreeNode(in startNode : TreeNode)
-ComputeAccuracy()
«metaclass»
myGridView
-myDataGridView : GridView
+InitGridView()
+New()
+New(in srcGridView : GridView)
+New(in colNames() : String)
+getGridView() : GridView
+setData(in srcGridView : GridView, in colName : String, in colValue : String) : Integer
+getPoisonCount(in poison : String) : Double
+getColumnCount() : Integer
+getRowCount() : Integer
+getColValCount(in colIndex : Integer, in colVal : Integer) : Integer
+getColValAndPoisonCount(in colIndex : Integer, in colVal : Integer, in poison : String) : Integer
+getColumnName(in colIndex : Integer) : String
+IsAllSamePoison() : Boolean
+getPoisonValue() : String
+RemoveColumn(in colName : String)

Decision Tree Algorithm

Input : Training dataset, D.

Output : Decision tree.

Method :

1 Compute *Information gain* for each column of D.

2 Select the max information gain column as *root node* of the tree T.

```
Select the max information gain column as class node of the tree T.4
colCount = Total columns of D exclude ID fields.
getChildNode (colCount, "Yes", D, rootName, rootNode, classNode);
getChildNode (colCount, "No", D, rootName, rootNode, classNode);
```

Procedure : getChildNode(colCount, filter, dataset, rootName, rootNode, classNode)

```
if ( colCount > 0 && rootName is not Empty )
{
    Copy structure of dataset to childDataset and leave the column act as a root.
    rowCount = childDataset.setData( D, rootName, filter );
    if ( rowCount > 0 )
    {
        Compute Information gain for each column of childDataset.
        Select the max information gain column as root node of the tree T.
        Select the max information gain column as class node of the tree T.
        colCount = Total columns of childDataset.
        getChildNode (colCount, "Yes", childDataset, rootName, rootNode, classNode);
        getChildNode (colCount, "No", childDataset, rootName, rootNode, classNode);
    }
}
```

Generate Decision Rules Algorithm

Input : Decision Tree, T.

Output : Decision rules.

Method :

```
TraverseNode ( currentNode, Expression, rules )
    if ( expression == "Yes" AndAlso currentNode.Nodes[0].count <= 0)
        rules += currentNode.Text + "THEN"
    else if ( expression == "Yes" AndAlso currentNode.Nodes[0].count > 0)
        rules += currentNode.Text + "AND"
    else if ( expression == "No")
        rules += "ELSE IF" + currentNode.Text + "AND"
    else if ( expression == Poison Name)
```

```
        rules += currentNode.Text
    else
        rules += “IF” + currentNode.Text + “=”
    For each node in currentNode.Nodes
    {
        TraverseNode ( currentNode, Expression, rules )
    }
    Display rules.
```

ACKNOWLEDGEMENT

We thank the staff and our colleagues from the Rural Heath Unit of Jose Abad Santos, Davao Occidental, Philippines, headed by the Municipal Health Officer, Dr. Amparo A. Lachica, who provided insight and expertise that greatly assisted the research. We thank the Graduate School of Government and Management, University of Southeastern Philippines for assistance and for comments that greatly improved the manuscript. We are expressing our gratitude to our families for being an inspiration. Above all, to God.

CONCLUSION

With the recent emergence of the field of data mining, there is a great need for algorithms for building classifiers that can handle very large databases. This paper implements the user to view the knowledge of data mining. This paper has described the feature subset selection problem in supervised learning, which involves identifying the relevant or useful features in a dataset and giving only that subset to the learning algorithm. For the evaluation, the user used hold-out method as accuracy estimation technique. It shows how a decision tree is constructed by C4.5 algorithm. This paper is an attempt to use Data Mining techniques to analyze students academic data and to enhance the quality of higher educational system. This paper use decision tree induction algorithms for students exam grade result. The higher managements can use such as classification model to improve students’ performance according to the extracted knowledge. This paper demonstrates efficiency and effectiveness in dealing with toxicology for classification This paper does not compare with other classification methods such as Naïve Bayesian Classifier (NBC), Bayesian Belief Network (BBN).The future work will extend decision tree induction (C4.5 algorithm) to work on the other data sets. We can plan to test the toxics dataset by using other classifiers such as Naïve Bayesian and K-Nearest Neighbor.

REFERENCES

[1] J. Han and M. Kamber. “Data Mining Concepts and Techniques”. Morgan Kaufmann, 2001.
[2] Pang-Ning, Tan Michael Steinbach and Vipin Kumar“Introduction to Data Mining”
[3] Tom M. Mitchell “Machine Learning”, McGraw Hill, New York, 1997.
[4] http:// en. Wekipedia. Org.
[5] http://www.mindtools.com/pages/article/newTED-04.htm
[6] http://www.csse.monash.edu.au/~dld/MML.htm
[7] http://www.vanguardsw.com/DpHelp4/dph0007
[8] http:// www.decsiontrees.net
[9]University of Computer Studies , Mandalay
[10] http://www.onlamp.com/pub/a/python/2006/02/09/ai-decision-trees.html
[11] http:// eruditionhome.com/datamining
[12] L.Breiman,J.Friedman,R.A.Olshen and C.J.Stone, ” Classification and regression trees”
[13] Visual Studio, 2005