Development of Selection Model to Support Higher Education (Undergraduate Programs) using Data Mining Technique

Asst. Prof. Dr. Waraporn Jirapanthong

This research has been funded by
Dhurakij Pundit University
2010

รายงานผลการวิจัย
เรื่อง
การพัฒนาโมเดลเพื่อสนับสนุนการเลือกศึกษาต่อระดับอุดมศึกษา ระดับปริญญาตรีโดยใช้เทคนิคดาต้าไมน์นิ่ง

โดย

ผศ. ดร. วราพร จิระพันธุ์ทอง

รายงานการวิจัยนี้ได้รับทุนอุดหนุนจาก

มหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2553

Acknowledgement

I am grateful to Dhurakijpundij University for the financial support for this research.

Declaration

Some of the material in this report has been previously published in the paper:

- W. Jirapanthong, "Classification Model for Selecting Undergraduate Programs", the 8th International Symposium on Natural Language Processing (SNLP 2009), Bangkok, Thailand, 2009.

Contents

List of Figures

List of Tables

Abstract

This research project focused on the study of the influencing factors to the academic success of undergraduate students. The research project thus applied data mining concepts and techniques to develop a model for supporting the selection of higher education, particularly undergraduate programs. The research proposed to apply the techniques of classification. In particular, the classification model was built by learning process and presented in term of decision tree model. Many versions of models were tuned to best fit the relationships between the data attributes and class labels of the input data. Examples of the use of the model are shown. They illustrated the accuracy of model and some advantages in use.

บทคัดย่อ

งานวิจัยนี้มุ่งเน้นการศึกษาปัจจัยที่มีอิทธิพลต่อความสำเร็จในการเรียนของนักศึกษาปริญญาตรี โดยได้อาศัยหลักการและเทคนิคของเหมืองข้อมูลเพื่อที่จะพัฒนาโมเดลสำหรับการสนับสนุนการเลือกศึกษาต่อขั้นสูงโดยเฉพาะในระดับปริญญาตรี งานวิจัยได้นำเสนอการประยุกต์ใช้เทคนิคของการแบ่งคลาส โมเดลสำหรับการแบ่งคลาสถูกสร้างขึ้นเพื่อใช้สนับสนุนการเลือกศึกษาต่อขั้นสูงโดยเฉพาะในระดับปริญญาตรี โมเดลนี้ได้ถูกสร้างจากการกระบวนการเรียนรู้ข้อมูลและแสดงในรูปแบบของโมเดลต้นไม้สำหรับการตัดสินใจ โมเดลหลายเวอร์ชั่นได้ถูกสร้างและปรับเปลี่ยนเพื่อให้สอดคล้องกับลักษณะของข้อมูลได้ดีที่สุด นอกจากนี้ตัวอย่างของการใช้โมเดลได้ถูกนำเสนอเพื่อแสดงความถูกต้องของโมเดลและผลดีที่ได้จากการใช้โมเดลนี้

Chapter I Introduction

1.1 Research Motivation

Nowadays it is strongly believed that having an adequate knowledge of business was essential to achieving success in life. A large number of high-school students aim to obtain the degree from the universities. A key to succeed in academic life is that the students need to put themselves into the right course regarding their knowledge, potential skills, and interests. Every student is expected to acquire new knowledge irrespective of the course of study chosen.

In addition, the technologies of data warehousing, OLAP and data mining have received a great deal of attention from various researchers in the past decade. Recent successful applications include the following: OLAP application can be used to access to data and support ad hoc analysis, reporting, and estimation. Data mining application can be applied for supporting on education selection.

This research project focuses on the study of the influencing factors to the academic success of undergraduate students. The research project thus applies data mining concept and techniques to develop a model for supporting the selection of higher education, particularly undergraduate programs. The model is built by learning an input data set. The learning algorithm is used to discover the model that best fits the relationships between the attribute set and class labels of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records which it has never seen before. The model is then applied with data set to classify the class labels of students.

1.2 Problems Statement

The problems in the selection of study course in undergraduate programs could be classified into two categories as follows:

1. The students are not supported by utilizing the information and data for analysis, planning and making a decision in course selection.

Even though the data and information about undergraduate courses are largely available and easily accessible, many students still fail in selecting the right course to be fitted their potential skills, knowledge, and interests.

2. Lack of useful knowledge although the existing system has large amount of data.

A large amount of data of information systems in the universities has been accumulated for years and was expand continuously to support daily operations of end-users. The daily process is a static process, so the end-users will not get new knowledge. The end-users lack of useful knowledge to enhance planning management.

1.3 Research Objectives

The objectives of this research project are as follows:

1. To increase the degree of effectiveness and decrease time of inquiry for prospective students.
2. To improve the success in the student selection system for higher education.
3. To improve and develop a data repository and information system in an academic institute.
4. To develop the system to supporting a decision making for high-school schools who are applying for undergraduate programs.

1.4 Scope of Work

1. This research project applies data mining techniques to develop a model for supporting the selection of higher education, particularly undergraduate programs.
2. This research project focuses on the improvement in academic quality in public and private universities.
3. This research has applied an *input data set* which are concerned the personal profiles of students who are studying in undergraduate level.

The remainder of this report is organized in five chapters as described below:

Chapter 2 describes the background of education in Thailand and presents an overview of Data Warehousing, OLAP technology, and Data Mining.

Chapter 3 describes a learning process including the structure of input data set applied in the learning process and presents the model that is built from the learning process.

Chapter 4 contains a description of the experiments that we have developed to demonstrate the work and analyses the experimental results of using the model.

Chapter 5 discusses the conclusions and directions for future work.

Chapter II Literature Review

This research project has made a study of fundamental structure of education in Thailand and related concepts of data warehousing, OLAP technology, and data mining. This chapter describes two sections: education in Thailand in Section 2.1; and data warehousing in Section 2.2.

2.1. Education in Thailand

2.1.1. Introduction to Education Structure and Administration [Anonymous 2002., Anonymous. 2004.]

The structure of school education in Thailand is based on a 6+3+3 system: six years of primary school, three years of lower secondary school and another three years of upper secondary school. The language of instruction is Thai, but English is taught as a second language in most secondary schools. In 1995, the government made English language study compulsory beginning at the primary school level.

Many primary and secondary schools were damaged or destroyed as a result of the tsunami disaster late last year. In the aftermath, 212 deaths and 108 missing persons were reported. Higher education is offered at universities, institutes of technology, vocational and technical colleges, teachers colleges and professional colleges (e.g. nursing schools).

The school year runs from mid-May to the end of March for primary and secondary schools, and from June to March for higher education. The Ministry of Education supervises all aspects of education from pre-school through upper secondary school and some higher education programs (e.g. teacher training and technical and vocational education). Private primary and secondary schools are managed under the Private Education Commission. The Office of the Higher Education Commission is

responsible for the administration and management of both public and private colleges and universities. The Department of Vocational Education is responsible for vocational education and training. Programs in this sector are designed to meet the needs of the job market and are offered at both the secondary and postsecondary levels.

2.1.2. Higher Education [Champoaunjam 2005]

There are a total of 780 public and private institutions in Thailand offering courses and programs in higher education. The Office of the Higher Education Commission sets educational standards, approves curriculum, and is the main institutional and professional accrediting body.

In 1999, approximately 26 percent of the18-21 year-old age group was enrolled in higher education studies. This leaves about 3,300,000 of the population from this age group not enrolled in an institution of higher education. However, higher education enrollments increased over 19 percent between 1998 and 1999, a trend boosted in part by the governments introduction of a student loan program.

In 2003, there were 4,170 foreign students enrolled in 49 Thai higher education institutions. China sent the most students that year (1,186), followed by Myanmar (359 students), India (329 students), Vietnam (304 students), Laos (226 students), United States (203), Japan (161 students), Taiwan (159 students), Cambodia (128 students) and Bangladesh (122 students) consecutively.

There were approximately 3,223 government funded Thai students in 1999 enrolled in overseas higher education programs.

The number of private universities has been increasing in recent years to help meet the growing demand for higher education. These institutions charge higher tuition fees than their public counterparts. In 1999, there were a total of 49 private universities enrolling 199,464 students. Private universities come under the authority of the Private Higher Education Institutions Division of the Office of the Higher Education Commission, which must approve and accredit new institutions. The largest private university is Bangkok University with 22,135 students.

The Rajamangala Institute of Technology offers bachelor programs in technology and technical fields. There are 50 of these institutes based around the country. The

Rajabhat Institutes, formerly teacher training colleges, and located in provinces throughout Thailand, provide training in practical fields such as tourism management and business administration. The Asian Institute of Technology (AIT), formerly the SEATO Graduate School of Engineering, is the only university that is not supervised by the Office of the Higher Education Commission. The AIT is a largely autonomous institution established under its own charter. Only 20 percent of its funding is provided by the Thai government. AIT enrolls students throughout the Asia-Pacific region and boasts an international faculty.

Thai institutions of higher education have also collaborated with foreign universities to set up specialized international programs. The Sasin Graduate Institute of Business Administration at Chaulalongkorn University, for example, offers an English language MBA program in collaboration with the Wharton School and the Kellogg Graduate School of Management.

2.1.3. Admission to Higher Education [Anonymous 2002., Anonymous. 2004.]

Admission to an institution of higher education requires the Certificate of Secondary Education (Matayom VI). Most public universities also require applicants to take the Joint Higher Education Entrance Examination (JHEEE), which is held each year in April and administered by the Office of the Higher Education Commission. Students who have successfully completed the Certificate of Vocational Education are also eligible to take the JHEEE.

Applicants are required to complete up to seven sections on the exam depending on the desired faculty. Science faculties for instance require mathematics, physics, chemistry, biology and English. Social Sciences faculties require social studies, Thai, English, other foreign languages and mathematics. Students may apply to as many as five faculties at one or more universities of their choice.

The JHEE is a highly competitive exam. Only about 30 percent of those who take the examination succeed in securing a place at a public university. In April 1994,

134,654 students took the JHEEE. Out of that number 22,000 were admitted to public universities and 17,000 were admitted to private universities.

Some institutions hold their own entrance exams while the country's two open universities, Ramkhamhaeng University and Sukhothai Thammathirat Open University, do not require applicants to take and entrance exam. Private institutions have their own admissions process, which includes a joint entrance examination similar to the JHEEE.

2.1.4. University Higher Education [Office of Higher Education Commission. 2006.]

The programs and degrees of universities in Thailand can be described as follows:

Stage I: The Bachelor's degree requires four years of full-time study in most fields. However, undergraduate programs in pharmacy, and graphic art requires five years of study. Bachelor's degree programs in medicine, dentistry and veterinary science require six years leading respectively to the Doctor of Medicine, Doctor of Dental Surgery and Doctor of Veterinary Medicine.

All bachelor's degree programs are comprised of the following: 30 credits of general education including humanities, social sciences, science, and mathematics: 15 credits in the specialized field and three credits of free electives. The four years bachelor's degree at all Thai universities are between 120 and 150 credits. A minimum grade point average of 2.0 is required for graduation.

Stage II: The Master's degree requires between one and two (usually two) years of full-time study with a minimum of 36 semester credits beyond the bachelor's degree. Some programs require an entrance exam for admission. Both coursework and a thesis (or comprehensive final examination) are required.

Stage III: The Doctorate requires between two and five years of study beyond the master's level. Admission to a doctoral program requires completion of a master's program with a cumulative GPA of 3.5 and an entrance examination. Programs require coursework, research and the defense of a dissertation.

2.2. Data Warehousing

Data warehouse is defined as "a subject–oriented, integrate, time-variant, and nonvolatile collection of data in support of management's decision making process" [Inmon, W.H. 2005]. Data warehouse integrates information from internal and external data sources of an organization into a center large database system.

Data warehouse uses multidimensional data store to support and implement querying and analysis for purpose of business decision making and enterprise management. The construction of data warehouses involves Extracting, Transforming, and Loading (ETL) processes, includes Extraction of data from many heterogeneous data sources, Transformation of extracted data into data warehouse structure, and Loading of transformed data into warehouse [Han, J. and Kamber, M. 2006].

2.2.1. Data Warehouse Architecture

Figure 2.1 illustrates multi-tiered data warehouse architecture, which consists of six parts as follows:

1. Data source is raw data from operational databases which record details of business transactions and other external sources

2. Back–end process or ETL process is data warehouse constructing process including data extraction, cleaning, transformation, load, and refresh utilities for populating warehouse. A brief detail of each process is as follow:

   2.1. Data extraction: The process that selected and extracted target data from multiple, heterogeneous, and external sources and gather it to the staging area. Extraction in data warehouse can be divided into 2 categories as follows:

   - Full extraction: the data is extracted completely from data sources.
   - Incremental extraction: only the data that has changed since a well defined event back in history will be extracted.

2.2. Data cleaning: The process that eliminates inconsistencies in the multiple data sources, detect and correct inaccurate record/attribute to enhance data quality. Example of data cleaning includes removing missing value or null value handling.

2.3. Data transformation: The process of transforming source data structure into warehouse data structure. Converting data include:
- Add new calculating values for an attribute which is not contain in source data.
- Split one attribute into multiple attributes.
- Merge several attributes from different source into one attribute, which the same object.
- Create surrogate key as primary key when the key from source data cannot used to primary key.
- Aggregating data according to defined user requirement instead of loading all detail of data into warehouse.

2.4. Loading: The process of loading data into the warehouse may still be required including sorting, consolidating, computing views, checking integrity, and building indices.

2.5. Refreshing: The process of propagating updates on source data to correspondingly update base data and derived data stored in the warehouse [Chaudhuri, S. and Dayal, U. 1997].

3. Data Storage is a warehouse database server that is commonly relational database system. Back-End Tools and utility are used to feed data into warehouse from data source.

4. Metadata Repository is a database designed to stores and manages data definitions and metadata from source to target. Metadata repository should contain a description of the structure of the data warehouse, operational metadata, the algorithms used for summarization, the mapping from the operational environment to the

data warehouse, data related to system performance, and business metadata.

5. Online Analytical Processing (OLAP) Engine or OLAP Server, which a special-purpose server that directly implements multidimensional data and operation. An OLAP sever that is almost ROLAP (Relational OLAP) and MOLAP (Multidimensional OLAP) (see more details in Section 2.3.)

6. Front-End Tools is front-end client, which include query/report, OLAP tools for multidimensional analysis and data mining tools for find previously unknown useful knowledge

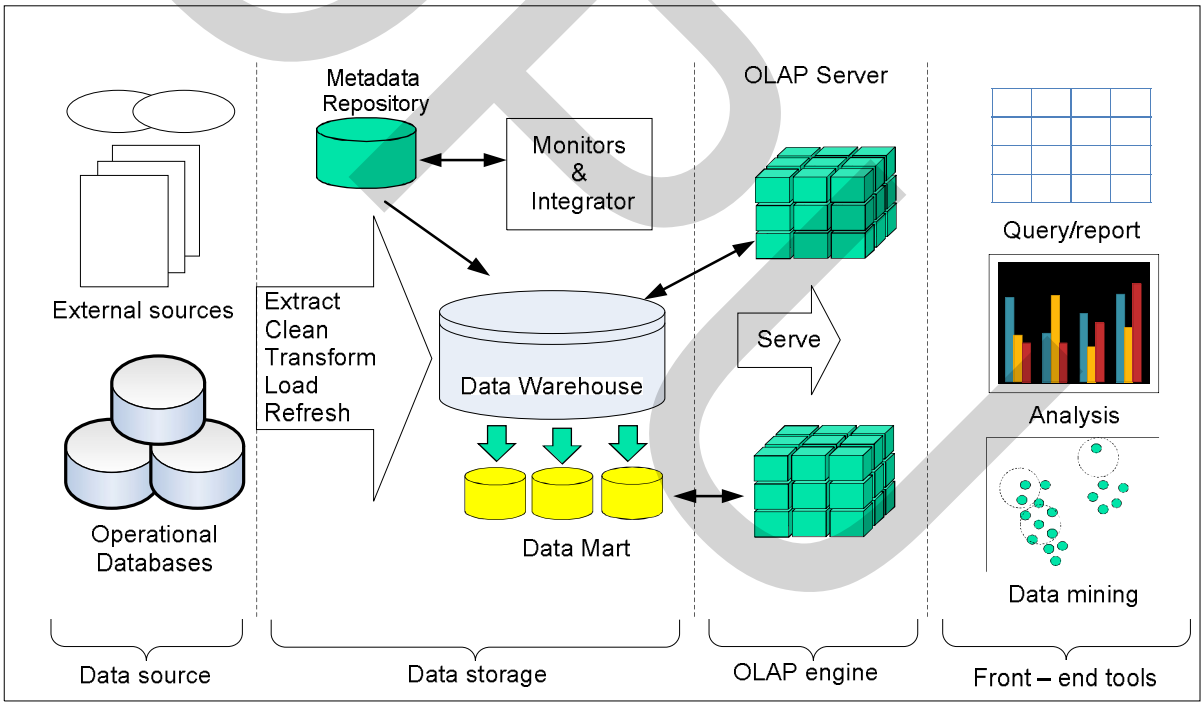


Figure 2.1. Data warehousing architecture [Han, J. and Kamber, M. 2006].

2.2.2. Dependent and Independent Data Mart

Data mart [Inmon, W.H. 2005] is small a data warehouse. The scope is confined to specific selected subjects. Data mart can be categorized as dependent data marts and independent data marts.

A dependent data mart is sourced directly from enterprise data warehouse. Figure 2.2 (a) shows a dependent data mart requires forethought and investment. It requires multiple users to pool their information needs for the creation of the data warehouse.

An independent data mart is built directly from the legacy applications. Figure 2.2 (b) shows an independent data mart. An independent data mart can be integrate need of data and create data standardize to decrease access time. The independent data mart represents a subset of the entire Decision Support System requirement for an organization. An independent data mart is a relatively inexpensive thing to build, and it is data source for OLAP and data mining.

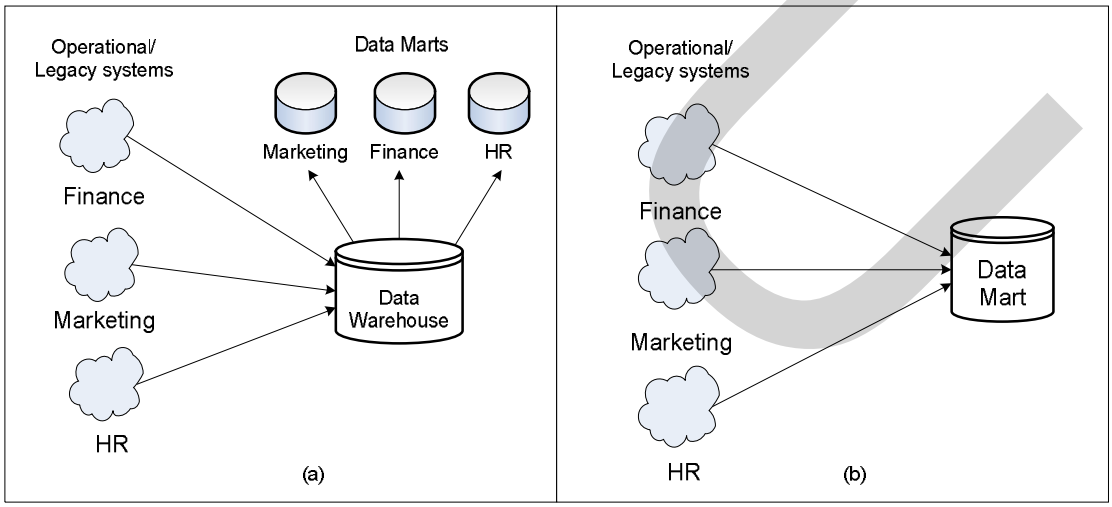


Figure 2.2. (a) dependent data mart (b) an independent data mart

2.2.3. Multidimensional Data Model

Data Warehouse and OLAP technology are based on a multidimensional data model [Han, J. and Kamber, M. 2006]. This model views data in the form of a data cube which data is modeled and viewed in multiple dimensions. Dimension table and fact table are basic component of multidimensional data model and are defined as follows.

- Dimension table

Dimension table is the perspective or entity that related to fact table. The dimension tables contain the textual description of the business. The examples of dimension are taxpayer, location, and time.

- Fact table

Fact table is used to analyze and summarize to gain a better understanding of the business. The fact contains numerical data that measures the business operations and foreign key that link to primary keys of related dimension tables.

The schemas for data warehouse modeling can be divided into 3 categorized**:**

1. Star Schema

A single large central fact table connected to a set of dimension tables with its foreign keys. The dimension tables are not normalized. The advantage of star schema is simple design, maintenance, and efficient queries because of fewer joins of a fact table with dimension tables. Figure 2.3 is an illustration of a star schema with the central sales fact table that contains keys to each of four dimensions, namely, location, store, calendar, and product, and along with two measures: sales amount and unit sold quantity.

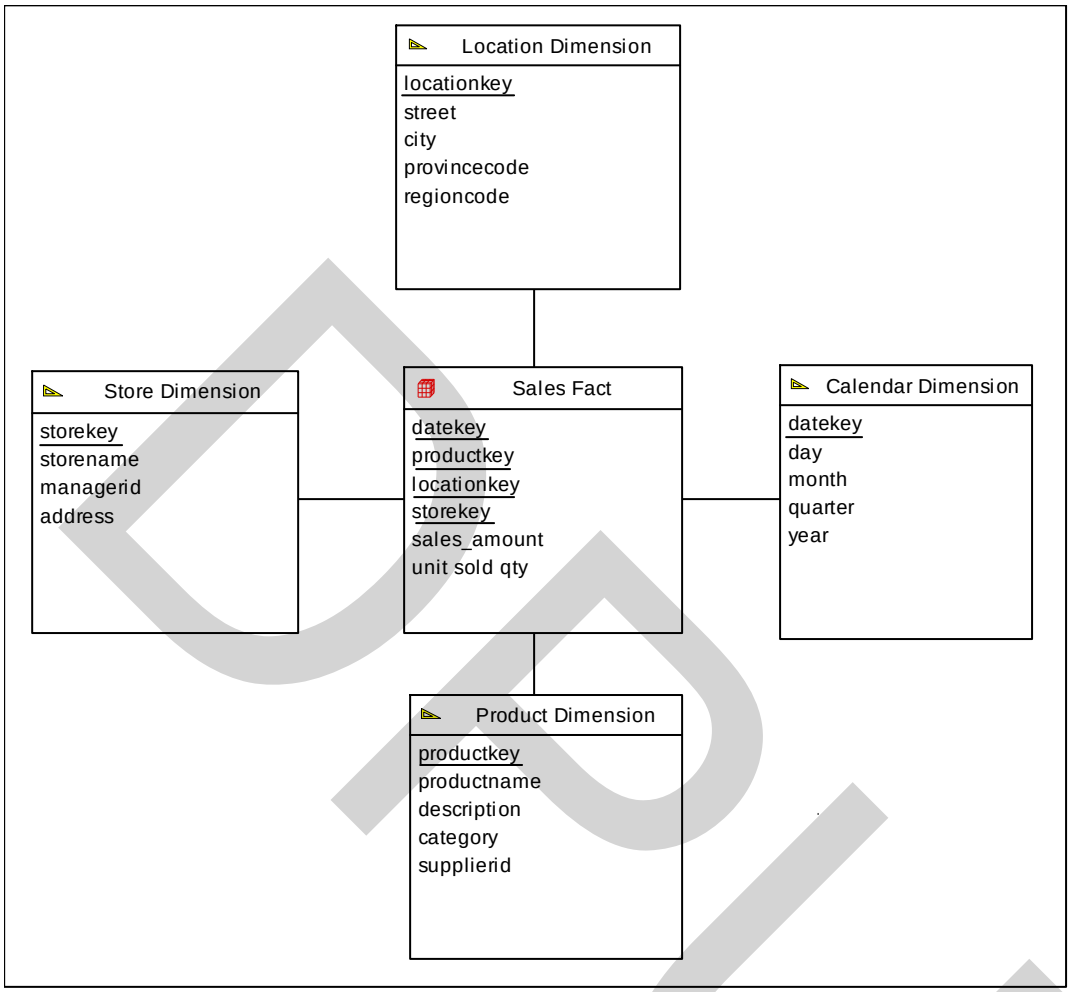


Figure 2.3. Star schema

2. Snowflakes Schema

The snowflakes schema is a variant of the star schema model. Snowflakes schema provides a refinement of star schema where some dimension tables are in normalized form. It divided into multiple tables or sub-dimension tables joined to a primary dimension table. The snowflakes schema increases the complexity and cost of queries because of more number of joins. The advantage of snowflake schema is to reduce redundancies. Such a dimension table is easy to maintain and save storage space. Figure 2.4 illustrates snowflakes schema with the main difference between the star and snowflake is in the dimension tables. The

single dimension table for location in the star schema is normalized in the snowflake schema, resulting in new location and province dimension where province code in location is linked to province. Similarly, the single dimension table for product in the star schema can be normalized into three new tables: product, supplier, and category.

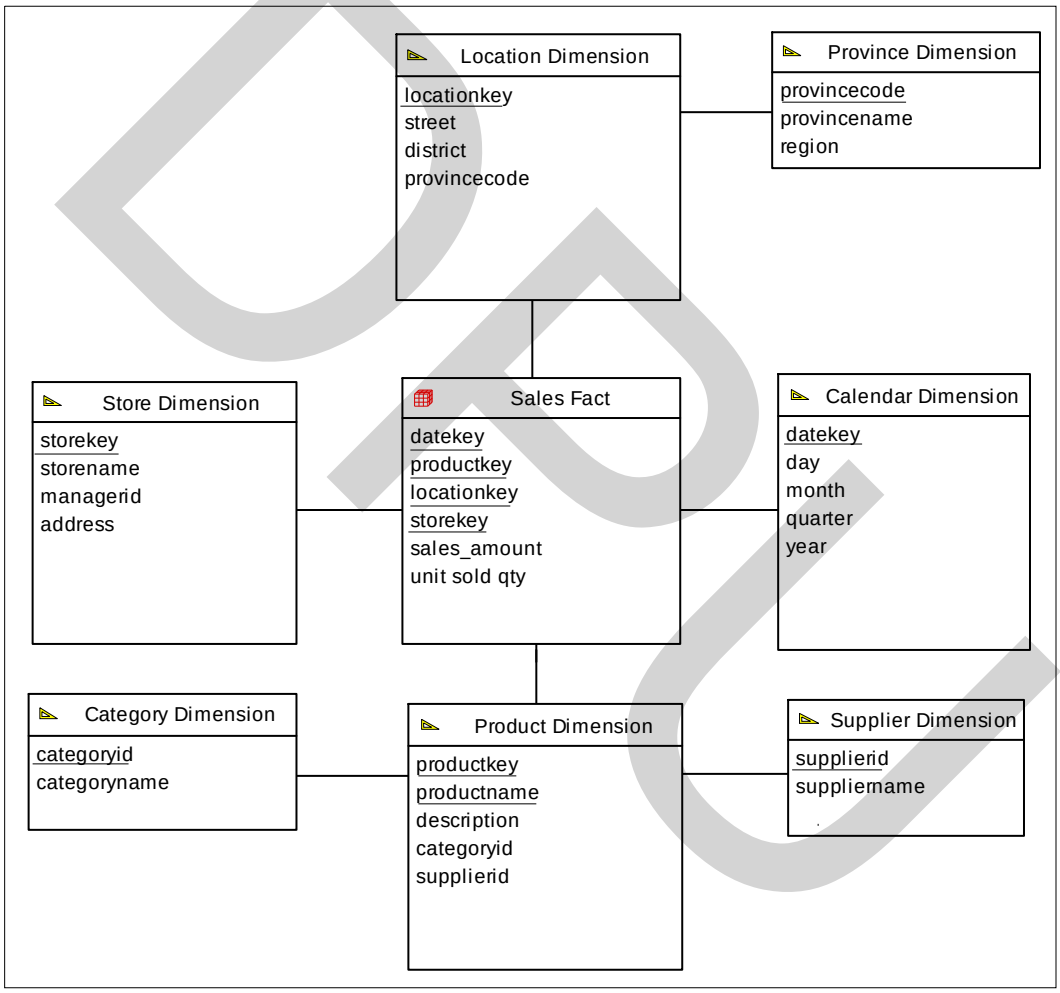


Figure 2.4. Snowflakes schema

- Fact Constellations Schema

Multiple fact table share some dimension tables. This kind of schema can be a collection of stars. The fact constellations schema facilitate to access data and decreases storage space for dimension tables. Figure 2.5 is an illustration a fact constellations with specifies two fact tables, sales and purchases. The sale fact table definition is identical to that of star schema. The purchases fact table has three dimensions, namely, location, calendar, and product and along with two measures: quantity and purchases.

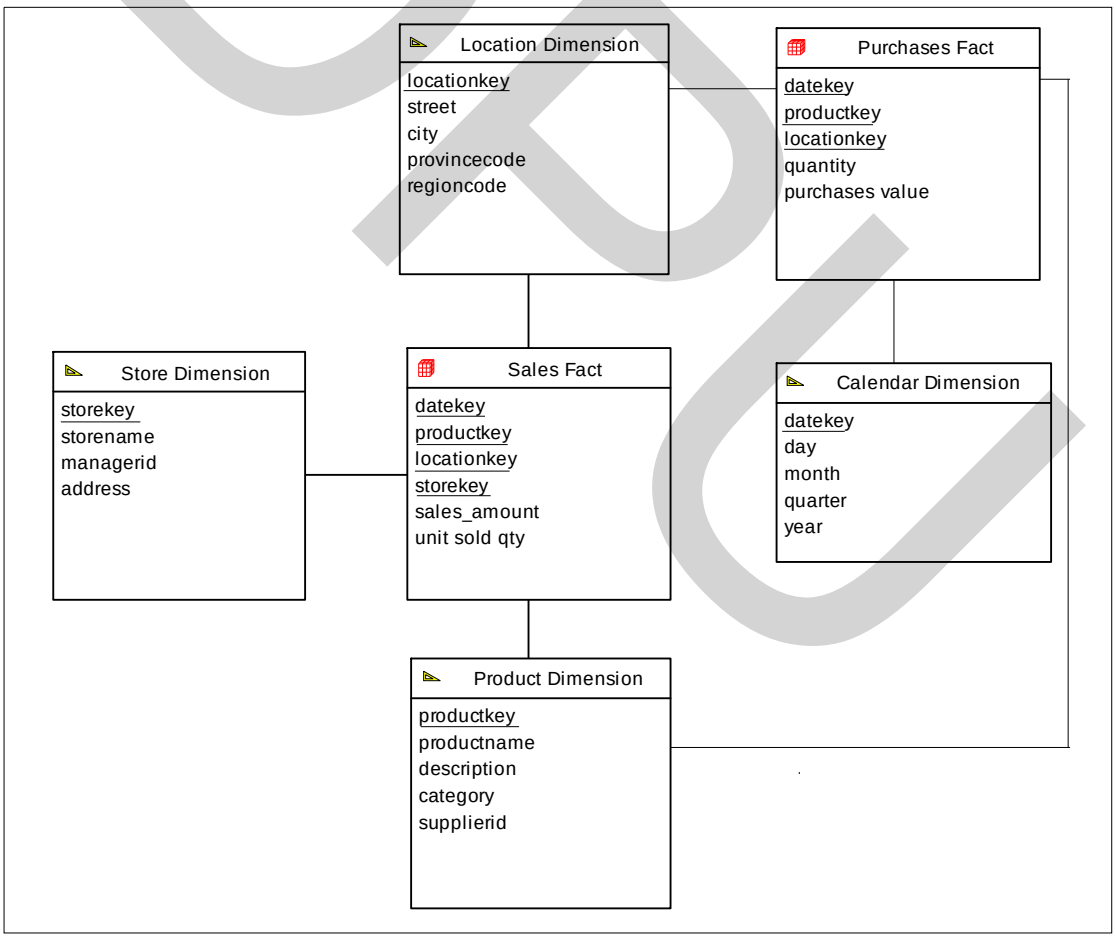


Figure 2.5. Fact constellations schema

2.3. Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) sometimes called multidimensional DBMS processing is data warehouse technology. A data is organized into multiple dimensions, and dimension contains multiple levels of abstraction defined by concept hierarchies. OLAP provide an information system with the structure that allows an organization to have very flexible access to data, to slice and dice data in many ways, and to dynamically explore the relationship between summary and detail data. Five basic OLAP operations are described as follows:

1. Roll- up is the ability to start at a detail number and to increase to the high level of aggregation that summarize into a successively coarse of summarization according to a defined concept hierarchy. Figure 2.6 depicts roll up operation perform on sale amount by claiming up to a concept hierarchy of location dimension from level of province to level of region.

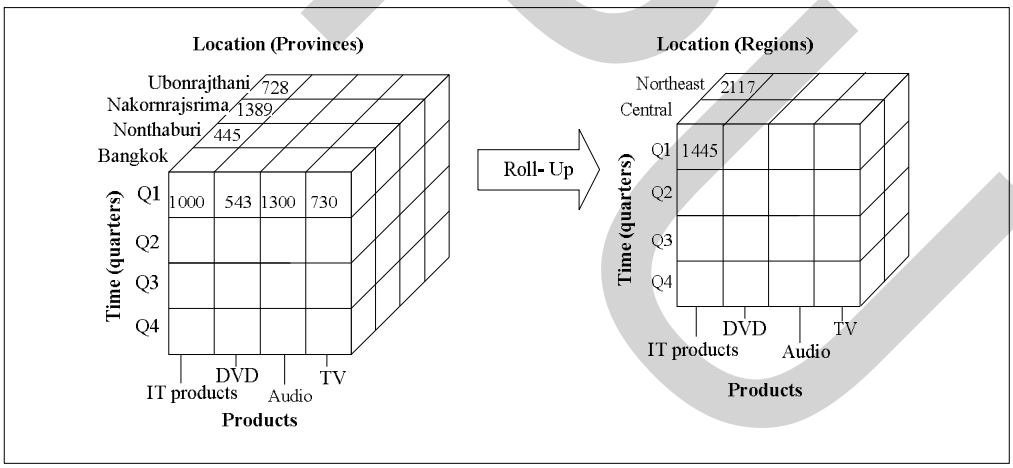


Figure 2.6. Roll-up operations

2. Drill-down is the ability to start at a summary number and to break that summary into a successively finer set of summarization according to a defined concept hierarchy. Figure 2.7 shows a simple drill-down operation performed on sales amount in 1st quarter by stepping down a concept hierarchy of time dimension from level of quarter to level of month.

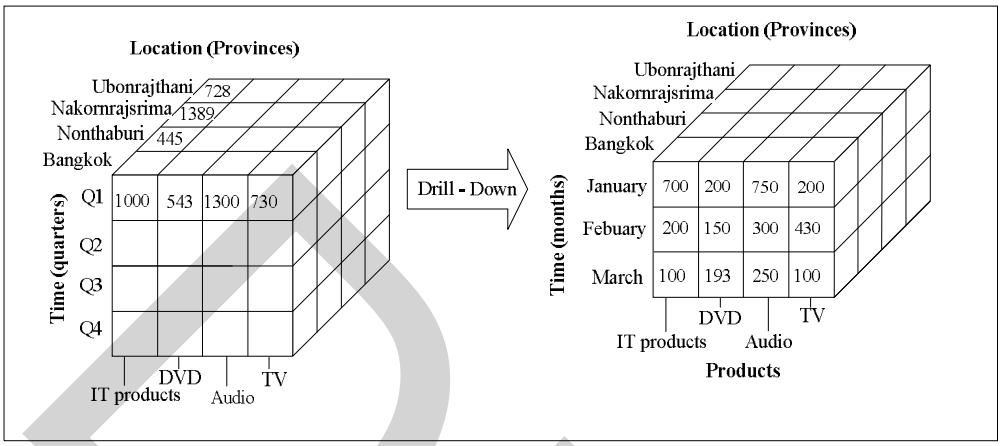


Figure 2.7. Drill-down operations

3. Slice operation performs a selection on one dimension of a given cube, resulting in a subcube. A slice operation is show in Figure 2.8 where the sales are selected from the central cube for the dimension time using the criterion time = "Q1".

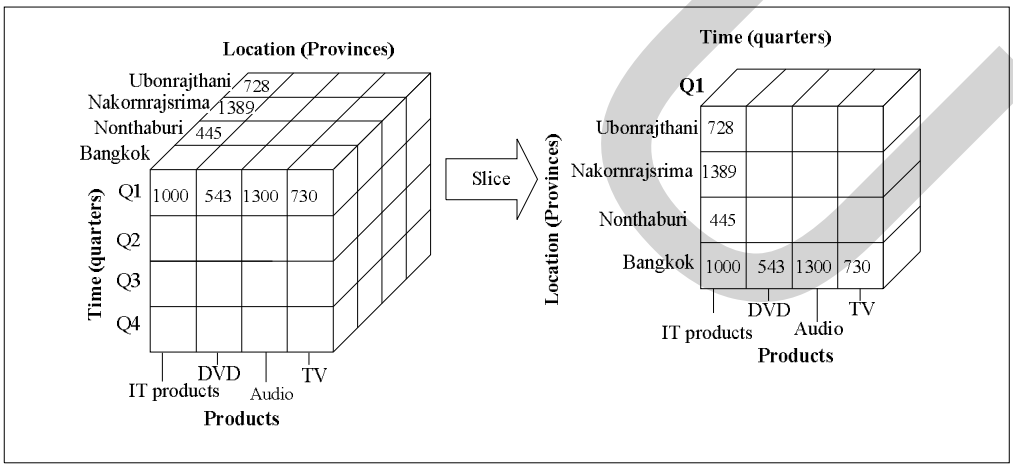


Figure 2.8. Slice operations

4. Dice operation defines a subcube by performing a selection on two or more dimensions. Figure 2.9 illustrates a dice operation on the central cube based on the following selection criteria that involves three dimensions: the dimension time using criterion time = "Q1" and the dimension location using criterion location = "Bangkok" and the dimension products using criterion product = "TV".

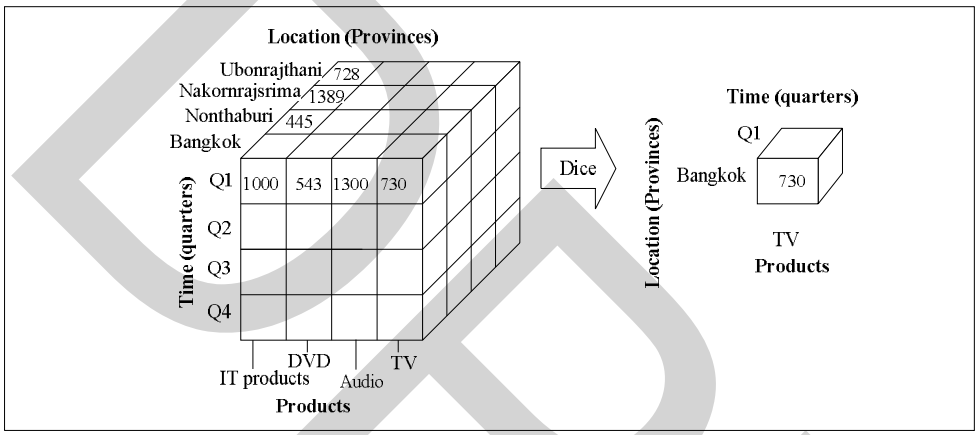


Figure 2.9. Dice operations

5. Pivot (Rotate) is an operation that rotates the data axes in order to provide an alternative presentation of the data. Figure 2.10 show a pivot operation where the products and provinces axes in a 2-D sliced and rotated.

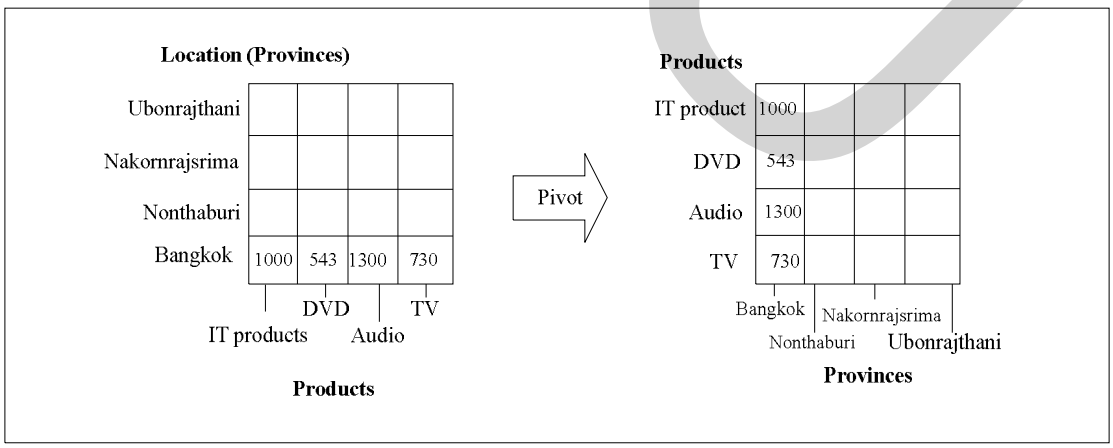


Figure 2.10. Pivot operations

2.4. Data Mining

The types of data mining techniques can be categorized differently. However, the most well-know data mining techniques fall into three methods which are association rule mining, classification and prediction and clustering. Details of each method are as follows:

1) Association Rule Mining

Association Rule Mining (ARM) is a method for discovering interesting relation between variables in a transaction databases. ARM is a famous technique for market basket analysis, introduced by [Agrwal, R. 1993]. There are many application areas including web usage, intrusion detection, and bioinformatics. Popular ARM algorithms are Apriori and FP-Grow.

2) Classification

This technique is a supervised learning technique that classifies data item into pre-defined class label. This appropriate technique builds model that predict future data trend. There are several algorithms for data classification such as Decision Tree, CART (Classification and Regression Tree) and Back Propagation neural network.

3) Clustering

The clustering technique or unsupervised learning technique is a division of data item into similar group without training of class labels. Clustering algorithms have been used in a large variety of applications, including image segmentation, construction the prototype of classifiers, understanding genomics data, market segmentation, etc. There are several clustering algorithms such as K-means, hierarchical agglomerative clustering and Self-Organizing Map.

2.4.1. Classification Techniques

Classification is one of the most useful techniques in data mining to building classification models from an input data set. Techniques for supporting building classification models are such as decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and na?ve Bayes classifiers. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. The key objective of the learning algorithm is to build models with good generalization capability.

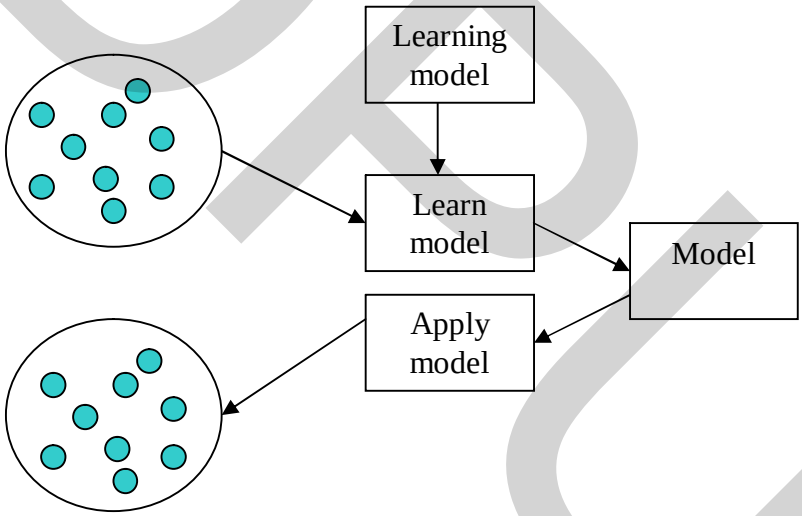


Figure 2.11. General approach for building a classification model

Classification techniques such as artificial neural networks (ANN) and rule-based classifiers are increasingly being used. Decision tree classifiers have, otherwise, been used as widely due to its simplicity, flexibility, and computational efficiency [Friedl et. Al., 1997]. In this section, decision tree classifier applied in this research project is described below.

a) Decision Tree Classifiers

In the usual approach to classification, a common set of features is used jointly in a single decision step. An alternative approach is to use a multi-stage or sequential hierarchical decision scheme. The basic idea involved in any multistage approach is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained in this way would resemble the intended desired solution. Hierarchical classifiers are a special type of multistage classifier that allows rejection of class labels at intermediate stages.

Classification trees offer an effective implementation of such hierarchical classifiers. Indeed, classification tree have become increasingly important due to their conceptual simplicity and computational efficiency. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifiers new data. Decision tree classifiers can perform automatic feature selection and complexity reduction, and their tree structure provides easily understandable and interpretable information regarding the predictive or generalization ability of the classification.

To construct a classification tree by heuristic approach, it is assumed that a data set consisting of feature vectors and their corresponding class labels are available. The features are identified based on problem specific knowledge. The decision tree is then constructed by recursively partitioning a data set into purer, more homogenous subsets on the basis of a sett of tests applied to one or more attribute values at each branch or node in the tree. This procedures involves three steps:

(i) splitting nodes;

(ii) determining which nodes are terminal nodes; and

(iii) assigning class label to terminal nodes.

The assignment of class labels to terminal nodes is straightforward. Labels are assigned based on a majority vote or a weighted vote when it is assumed that certain classes are more likely than others.

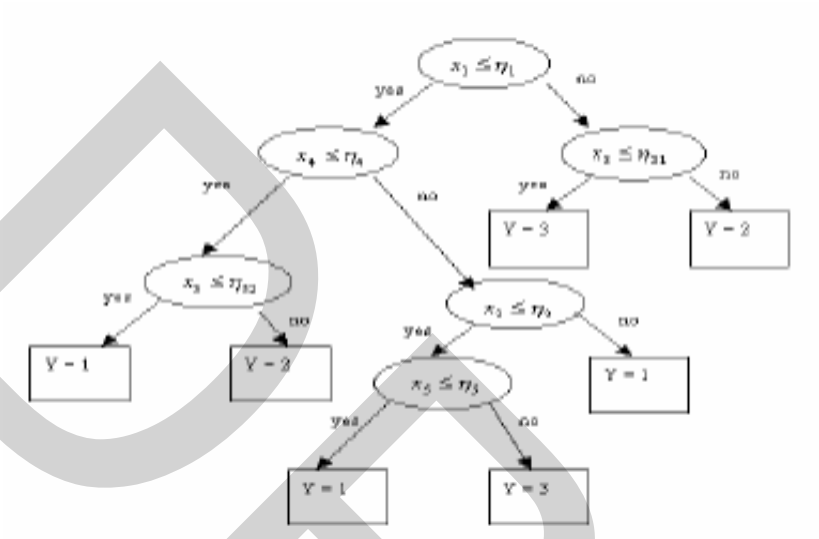


Figure 2.12. A classification tree for a five feature space and tree classes. The X_i's are the feature values, the η_i's are the thresholds, and y is the class label.

A tree is composed of a root node (containing all the data), a set of internal nodes (splits), and a set of terminal nodes (leaves). Each node is a decision tree has only one parent node and two or more descendent node (as shown in the figure). A data set is classified by moving down the tree and sequentially subdividing it according to the decision framework defined by the tree until leaf is reached.

(i) Pruning Decision Tree

Decision tree classifiers divide the available training data into subsets, each representing a single class. The result of this procedure is often a very large and complex tree. In most cases, fitting a decision tree until all leaves contain data for a single class may overfit to the noise in the training data, as the training samples may not be representative of the population. If the training data contain errors, then overfitting the tree to the data in this manner can lead to poor performance

on unseen cases. To minimize this problem, the original tree must be pruned to reduce classification errors when data outside of the training set are to be classified.

A decision tree is not usually simplified by deleting the whole tree in favour of a leaf. Instead, parts of the tree that do not contribute to classification accuracy on unseen cases, thus producing less complex and more comprehensible tree, are removed [Quilan, R. 1993, Mingers 1989, and Brieman et.al. 1984].

(ii) Boosting

In recent years, a number of works proposing the use of combining the predictions of multiple classifiers to produce a single classifier have been reported. The resulting classifier, referred to as an ensemble, is generally found to be more accurate than any of the individual classifiers making up the ensemble. For example, ensembles of natural network [Giacinto et. Al. 1997] and integration of the results of different type of classifiers [Wilkinson et. al. 1995] are found to be effective in improving classification accuracy. Much of this research is focused on improving the classification accuracy, as accuracy is the primary concern in all applications of learning. Only a few authors have reported the use of boosting, another technique to improve the performance of any learning algorithm (e.g. Fridl, et. al.). the basic difference between the use of ensembles of classifiers and boosting is that boosting uses the same learning algorithm that consistently generates multiple classifiers in an iterative manner.

Boosting is a general method for improving the performance of any learning algorithm. Boosting can be used to reduce the error of any weak learning algorithm that consistently generate classifiers on various distributions over the training data, and then combining the classification produced by the weak learner into a single composite classifier.

b) C4.5 algorithm

C4.5 is an algorithm used to generate a decision tree developed by [Quinlan, R. 1993]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1,s_2,...$ of already classified samples. Each sample $s_i = x_1,x_2,...$ is a vector where $x_1,x_2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1,c_2,...$ where $c_1,c_2,...$ represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sublists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

In pseudocode, the algorithm is:

i. Check for base cases

ii. For each attribute a

    (a) Find the normalized information gain from splitting on *a*

iii. Let a_best be the attribute with the highest normalized information gain

iv. Create a decision node that splits on a_best

v. Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node

2.4.2. Clustering Techniques

Clustering is one of the most useful techniques in data mining for partitioning large data set into groups according to their similarity and focusing on a particular set of future analysis. A definition of optimal clustering is a partitioning that minimizes distances within clusters and maximizes distances between clusters. The optimal clustering is shown in Figure 2.13. In this section, Euclidean distance and two clustering algorithms: Kohonen's Self Organizing Map and K-means are described below.

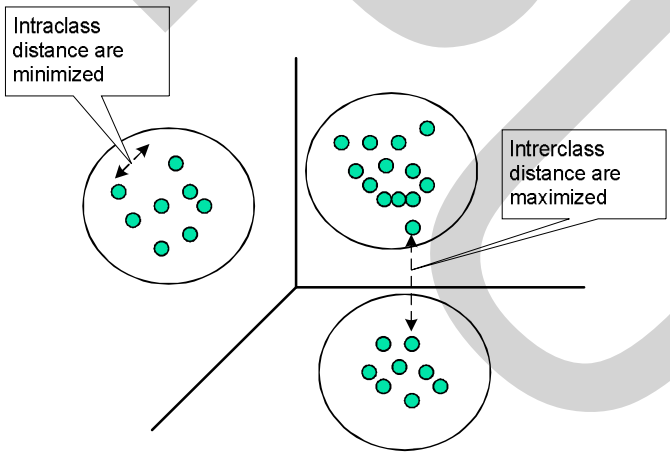


Figure 2.13. Optimal clustering

a) Euclidean Distance [Deza, E. and Deza, M. 2006]

Euclidean distance is the most popular distance measure to calculate the dissimilarity (similarity) between two data objects from same space is to clustering procedures. Euclidean distance is defined as the equation (2.1).

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2} \quad (2.1)$$

Where:

i = (x_{i1}, xi_2,…,x_{in}) are two n- dimensional data objects.

j = (x_{j1},x_{j2},…,x_{jn}) are two n- dimensional data objects.

b) Kohonen's Self Organizing Map Algorithm

The Kohonen's Self Organizing Map (SOM) algorithm, widely also known as Kohonen network is a popular algorithm for unsupervised learning. The SOM algorithm based on neural network structure in two layers [Kohonen T. 2001].The first layer represents the input data, the other one show output map. The goal of SOM is to represent all point in a high dimensional source space by points in a low dimensional target space. Figure 2.14 illustrates architecture of SOM.

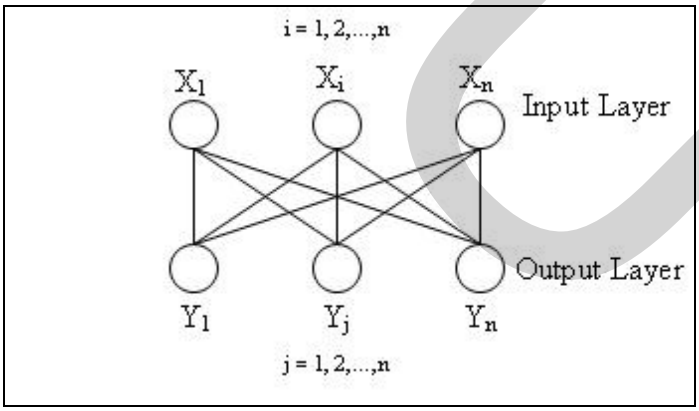


Figure 2.14. A typical Self-Organization Maps architecture

The concept of Kohonen's Self-Organizing Maps Neural Network is iterative for each data to find weight according number of clusters as follows:

1. Initialize the weights vector of all the output neurons.

$$\left|X - W_m\right| = \min_{j=1...M} \left|X - W_j\right| \quad (2.2)$$

2. Determine the output wining neuron *m* by searching for the shortest normalized Euclidean distance between the input vector and the weight vector of each output neuron, by using the equation (2.2).

Where:

X is the input vector,

W_j is the weight vector of output neuron *j,* and

M is the total number of output neuron.

3. Let N_m (t) denote a set of indices corresponding to a neighborhood size of the current winner neuron *m.* The neighborhood size needs to be slowly decreased during the training session. The weights of the weight vector associated with the winner neuron *m* and its neighborhood neurons are updated by the equation (2.3)

$$\Delta W_j(t) = \alpha(t)[X(t) - W_j(t)] \quad \text{for } j \text{ E } N_m(t) \tag{2.3}$$

Where α is a positive-valued learning factor, α E [0, 1]. It needs to be slowly decreased with each training iteration. Thus, the new weight vector is given by the equation (4)

$$W_j(t+1) = W_j(t) + \alpha(t)[X(t) - W_j(t)] \quad \text{or } j \text{ E } N_m(t) \tag{2.4}$$

Steps 2 and 3 are repeated for every exemplar in the training set for a user-defined number of iterations.

c) K-means Algorithm

The *k*-means algorithm was introduced by Mac Queen in 1967. The *k*-means algorithm is well known for its efficiency in clustering large data sets. The *k*-

means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters [Han, J. and Kamber, M. 2006] so that the overall sum of square error is minimized.

A major limitation of *k*-mean is to determinate the number of *k* clusters that is to affect the accuracy of the data partition and the efficiency of the clustering processing. As the *k* parameter increase, the data item will be divided into clusters and the processing will need large computational power. On the contrary, the *k* parameter is small, the data item will be divided into less clusters and some significant characteristic of data may be lost. The *k*-means algorithm is described as follows:

1) Determinate the number of *k* clusters.
2) Randomly selects initial represent point a cluster mean.
3) Assign each object to the closest cluster center, based on similarity measure.
4) Computes the new mean for each cluster. This process iterates until object is no change group (cluster).If object change group go to step 3.

2.5. Summary

This chapter has provided background information for education in Thailand. It described the structure and academic environment of education system in Thailand. Moreover, the concepts and related techniques of data mining are presented. In the next chapter, we illustrate the learning process of classification model, the details of data set and software tool techniques applied, and the proposed classification model.

Chapter III Classification Model Construction and Evaluation

This chapter presents the data set which is used; it also illustrates the learning process of classification model. The details of data set and software tool techniques applied for building the classification model are given. Some examples of the classification models built during the learning process are shown. The final version of classification model is also described.

3.1. Introduction

It is significant to build up the classification model of higher-education, particularly undergraduate programs. The datasets which are used in this research project are collected from 52 schools (including public and private schools), and one private university. The datasets were divided into two sub-data, one is model building sub-data, and the other is model testing sub-data. By using of decision tree technique, the classification model had been setup. For the details of data attributes, we have described in Section 3.2. Additionally, we considered the data records of students who have current accumulative grade from 2.5 since it can be implied that the students are qualified to complete their degree and fair enough to continue further education in the same major or related. Moreover, we applied the 10-fold cross validation method for evaluating the classification model.

In addition, in the research, we declare the terminology as follows:

An instance -- a data record

ConfidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).

MinNumObj -- The minimum number of instances per leaf.

3.2. Dataset

This dataset is data about student's education type and student's GPA (Grade Point Average) in his old school. We used this dataset to explore the effect of student's education type and student's GPA in his old school to his studying faculty. The result from this experiment can be used to help students in choosing appropriate faculty for them.

There are 7778 records in dataset, in each record there are 5 attributes. Description of each attribute is shown in Table 3.1. The attribute *faculty* is used to be a class attribute.

Table 3.1. Description of each attribute

Attribute	Description	Possible values of attribute
Gender	Student's gender	F (Female) , M (Male)
OldMajor	The major program that student got.	Art, Thai-art-acting, Business Administrator, Laws, Engineering, Human-Science, Architectural, Statistics, Industrial-Science, Politicial-Science, Householding, Information-Science, Libralian-Science, Education, Supporting-Aggriculture-and-Cooperation, Information-Technology (IT), Economics, Science, Communication-Art, Nursing, Social-Science
OldGPA	GPA that student got from old school.	0.00-1.50, 1.51-2.49, 2.50-3.00, 3.01-3.49, 3.50-4.00
EdType	Student's education type	Diploma , HighSchool
Faculty	Faculty that student is studying, now.	Accounting, Business-Administration, Communication-Art, Business-Informatics, Communication-Art, Laws, Economics, Fine-Arts, Fine-Arts-and-Science, IT, Marketing-Communication, Education, Engineering, Political-and-Administrative-Science, Science

Table 3.1: properties of each attribute

3.3. Software Tool

We used data mining tool called Weka (http://www.cs.waikato.ac.nz/ml/weka/) to build decision tree.

The algorithm that we used is C4.5, which builds a decision tree by recursively selecting attributes on which to split. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and

terminal nodes reflect decision outcomes. The criterion used for selecting an attribute is an information gain. J.R. Quinlan has popularized the decision tree approach with his research spanning more than 15 years. The latest public domain implementation of Quinlan's model is C4.5. The Weka classifier package has its own version of C4.5 known as J48.

3.4. Learning Process of the Classification Model

We proposed to build up a classification model to give a guideline to a student who is applying for higher education, specifically undergraduate programs in private universities. The technique of decision tree was applied. Particularly, the decision tree was developed to determine which major is suitable for a student. It also represents the possibility of further higher education of the student in the levels of Master and Doctoral degrees. The data as described in section 3.2 was prepared through process i.e. extraction, cleaning, transformation, loading and refreshing. During the learning process to build up a decision tree, many trees were created and evaluated by the 10-folder cross validation method. The evaluation results of those decision trees were not convincing. Eventually, the final version of the decision tree returns the results with high accuracy. In the following, we show only three examples (versions 1, 2, and 3) of decision trees built-up during the process. For the final version of the decision tree as the proposed classification model, it is presented in next section.

3.4.1. Version 1

The details of version 1 are following:

ConfidenceFactor = 0.25

MinNumObj = 2

Evaluating method: 10-fold cross validation

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Student

Instances: 7778

Attributes: 5

Gender

OldMajor

OldGPA

EdType

Faculty

Classifier model (full training set)

J48 pruned tree

```
------------------
OldMajor = Medium-Diploma: Accouting (2.0/1.0)
OldMajor = B.Art: Communication-Art (1.0)
OldMajor = Thai-art-acting: Master-of-Education (1.0)
OldMajor = Fundamental-Diploma
|   Gender = F
|   |   OldGPA = 0.00-1.50: Business-Administration (86.0/34.0)
|   |   OldGPA = 2.50-3.00: Accouting (129.0/98.0)
|   |   OldGPA = 3.50-4.00: Accouting (16.0/10.0)
|   |   OldGPA = 3.01-3.49: Accouting (71.0/46.0)
|   |   OldGPA = 1.51-2.49: Business-Administration (338.0/227.0)
|   Gender = M
|   |   OldGPA = 0.00-1.50: Business-Administration (81.0/56.0)
|   |   OldGPA = 2.50-3.00: IT (68.0/51.0)
|   |   OldGPA = 3.50-4.00: Accouting (10.0/8.0)
|   |   OldGPA = 3.01-3.49: Fine-Arts (29.0/22.0)
|   |   OldGPA = 1.51-2.49: Communication-Art (214.0/164.0)
OldMajor = B.BA.(Accounting): Master-of-Business-Administration (1.0)
OldMajor = B.-Laws: Master-of-Laws (248.0/21.0)
OldMajor = B.EG.: Master-of-Science (5.0/1.0)
OldMajor = B.Indurtrial: Master-of-Science (5.0)
OldMajor = High-School(Sc.): Business-Administration (677.0/495.0)
OldMajor = Master-Degree: Doctor-of-BA (40.0/25.0)
OldMajor = Human-Science: Doctor-of-Business-Info. (3.0/2.0)
OldMajor = Bachelor-of-Architectural: Master-of-Science (1.0)
OldMajor = Bachelor-of-Statistics: Master-of-Science (1.0)
OldMajor = Bachelor-of-Industrial-Science: Master-of-Business-Administration
(2.0/1.0)
OldMajor = Bachelor-of-Art(Thai-poem): Doctor-of-Communication-Art (1.0)
OldMajor = B.Politicial-Science: Master-of-Science (1.0)
OldMajor = Politician-Science: Master-of-Science (8.0/5.0)
OldMajor = High-School(Art.): Business-Administration (1121.0/799.0)
```

```
OldMajor = unknown
|   OldGPA = 0.00-1.50: Business-Administration (23.0/14.0)
|   OldGPA = 2.50-3.00: Communication-Art (25.0/19.0)
|   OldGPA = 3.50-4.00
|   |   Gender = F: Doctor-of-Business-Info. (8.0/6.0)
|   |   Gender = M: Business-Administration (3.0/2.0)
|   OldGPA = 3.01-3.49
|   |   Gender = F: Fine-Arts-and-Science (7.0/4.0)
|   |   Gender = M: Master-of-Science (18.0/10.0)
|   OldGPA = 1.51-2.49
|   |   Gender = F: Business-Administration (40.0/28.0)
|   |   Gender = M: IT (19.0/14.0)
OldMajor = Bachelor-of-Householding: Master-of-Business-Administration
(2.0/1.0)
OldMajor = Bachelor-of-Information-Science: Master-of-Business-Administration
(2.0/1.0)
OldMajor = Libralian-Science: Master-of-Science (1.0)
OldMajor = Bachelor-of-Art: Master-of-Business-Administration (43.0/24.0)
OldMajor = B.ED
|   Gender = F: Master-of-Business-Administration (2.0/1.0)
|   Gender = M: Master-of-Science (2.0)
OldMajor = Bachelor-of-Supporting-Aggriculture-and-Cooperation: Fine-Arts-
and-Science (2.0/1.0)
OldMajor = ICT: Master-of-Science (14.0/4.0)
OldMajor = High-Diploma
|   Gender = F: Business-Administration (405.0/193.0)
|   Gender = M
|   |   OldGPA = 0.00-1.50: Business-Administration (84.0/42.0)
|   |   OldGPA = 2.50-3.00: Master-of-Engineering (15.0/7.0)
|   |   OldGPA = 3.50-4.00: Master-of-Engineering (2.0)
|   |   OldGPA = 3.01-3.49: Master-of-Engineering (7.0/3.0)
|   |   OldGPA = 1.51-2.49: Master-of-Engineering (172.0/103.0)
OldMajor = Bachelor-of-Economics: Master-of-Economics (16.0/7.0)
OldMajor = Bachelor-of-Art(Secretorial): Master-of-Business-Administration (2.0)
OldMajor = B.SC: Master-of-Science (48.0/16.0)
OldMajor = Mattayom-5: Master-of-Science (1.0)
OldMajor = Bachelor-of-Communication-Art: Master-of-Communication-Art
(47.0/19.0)
OldMajor = B.EG: Master-of-Science (44.0/5.0)
OldMajor = B.BA.
|   OldGPA = 0.00-1.50: Master-of-Science (7.0/4.0)
|   OldGPA = 2.50-3.00: Master-of-Business-Administration (24.0/4.0)
|   OldGPA = 3.50-4.00: Master-of-Science (50.0/21.0)
|   OldGPA = 3.01-3.49: Master-of-Business-Administration (93.0/40.0)
|   OldGPA = 1.51-2.49: Master-of-Business-Administration (1.0)
OldMajor = Ed.-Special-school: Communication-Art (154.0/90.0)
OldMajor = B.Nursing: Master-of-Science (2.0)
```

OldMajor = Under-Bachelor: Business-Administration (2.0/1.0)
OldMajor = Bachelor-degree: Master-of-Science (53.0/34.0)
OldMajor = Bachelor-of-Educational
| OldGPA = 0.00-1.50: Master-of-Education (2.0/1.0)
| OldGPA = 2.50-3.00: Master-of-Communication-Art (4.0/2.0)
| OldGPA = 3.50-4.00: Master-of-Education (7.0/3.0)
| OldGPA = 3.01-3.49: Master-of-Science (8.0/2.0)
| OldGPA = 1.51-2.49: Master-of-Education (1.0)
OldMajor = B.PH: Fine-Arts (1.0)
OldMajor = B.Accounting
| OldGPA = 0.00-1.50: Master-of-Business-Administration (1.0)
| OldGPA = 2.50-3.00: Master-of-Business-Administration (5.0/1.0)
| OldGPA = 3.50-4.00: Master-of-Science (7.0/1.0)
| OldGPA = 3.01-3.49: Master-of-Business-Administration (22.0/7.0)
| OldGPA = 1.51-2.49: Master-of-Business-Administration (0.0)
OldMajor = High-School
| Gender = F
| | OldGPA = 0.00-1.50: Business-Administration (210.0/126.0)
| | OldGPA = 2.50-3.00: Fine-Arts-and-Science (405.0/300.0)
| | OldGPA = 3.50-4.00: Fine-Arts-and-Science (163.0/122.0)
| | OldGPA = 3.01-3.49: Business-Administration (266.0/198.0)
| | OldGPA = 1.51-2.49: Fine-Arts-and-Science (805.0/591.0)
| Gender = M: Communication-Art (1255.0/973.0)
OldMajor = Social-Science: Master-of-Communication-Art (1.0)
OldMajor = B.ED.
| OldGPA = 0.00-1.50: Master-of-Education (1.0)
| OldGPA = 2.50-3.00: Master-of-Business-Administration (1.0)
| OldGPA = 3.50-4.00: Master-of-Education (10.0/4.0)
| OldGPA = 3.01-3.49: Master-of-Communication-Art (2.0/1.0)
| OldGPA = 1.51-2.49: Laws (1.0)

Number of Leaves : 87

Size of the tree : 104

Summary

Correctly Classified Instances	5220	67.1124 %
Incorrectly Classified Instances	2558	32.8876 %
Total Number of Instances	7778	

Explanation

Let us take a second look at the first seven lines of the output to see how the tree structure is represented:

```
OldMajor = Medium-Diploma: Accouting (2.0/1.0)
OldMajor = B.Art: Communication-Art (1.0)
OldMajor = Thai-art-acting: Master-of-Education (1.0)
OldMajor = Fundamental-Diploma
|   Gender = F
|   |   OldGPA = 0.00-1.50: Business-Administration (86.0/34.0)
|   |   OldGPA = 2.50-3.00: Accouting (129.0/98.0)
```

Each line represents a node in the tree. The fifth line that starts with a '|', is a child node of the fourth line. The last two lines, those that start with '| | ', are child nodes of the fifth line. In the general case, a node with one or more '|' characters before the rule is a child node of the node that the right-most line of '|' characters terminates at, if you follow it up the page. The next part of the line declares the rule. If the expression is true for a given instance, you either classify it if the rule is followed by a semicolon and a class designation--that designation becomes the classification of the rule--or, if it isn't followed by a semicolon, you continue to the next node in the tree (i.e. the first child node of the node you just evaluated the instance on). If the expression is instead false, you continue to the ``sister'' node of the node you just evaluated; that is, the node that has the same number of '|' characters before it and the same parent node.

Notes that generate a classification, such as

OldMajor = B.Art: Communication-Art (1.0)

are followed by a number (sometimes two) in parentheses. The first number tells how many instances in the training set are correctly classified by this node, in this case 1 is the second number, if it exists (if not, it is taken to be 0.0), represents the number of instances incorrectly classified by the node.

Extracted Rules from decision tree in version 1:

1. If OldMajor = Medium-Diploma Then Faculty = Accounting
2. If OldMajor = B.Art Then Faculty = Communication-Art
3. If OldMajor = Thai-art-acting Then Faculty = Master-of –Education
4. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
5. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Accounting
6. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Accounting
7. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Accounting
8. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
9. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
10. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 2.50-3.00 Then Faculty =ICT
11. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Accounting
12. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Fine-Art
13. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Communication-Art
14. If OldMajor = B.BA(Accounting) Then Faculty = Master-of-Business-Administration
15. If OldMajor = B.-Laws Then Faculty = Master-of-Laws
16. If OldMajor = B.EG. Then Faculty = Master-of-Science
17. If OldMajor = B.Indurtrial Then Faculty= Master-of-Science
18. If OldMajor = High-School(Sc.) Then Faculty = Business-Administration
19. If OldMajor = Master-Degree Then Faculty = Doctor-of-BA

20. OldMajor = Human-Science: Doctor-of-Business-Info.
21. OldMajor = Bachelor-of-Architectural: Master-of-Science
22. OldMajor = Bachelor-of-Statistics: Master-of-Science
23. OldMajor = Bachelor-of-Industrial-Science: Master-of-Business-Administration
24. If OldMajor = Bachelor-of-Art(Thai-poem) Then Faculty = Doctor-of-Communication-Art
25. If OldMajor = B.Politicial-Science Then Faculty = Master-of-Science
26. If OldMajor = Politician-Science Then Faculty = Master-of-Science
27. If OldMajor = High-School(Art.) Then Faculty = Business-Administration
28. If OldMajor = unknown And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
29. If OldMajor = unknown And OldGPA = 2.50-3.00 Then Faculty = Communication-Art
30. If OldMajor = unknown And OldGPA = 3.50-4.00 And Gender = F Then Faculty = Doctor-of-Business-Info.
31. If OldMajor = unknown And OldGPA = 3.50-4.00 And Gender = M Then Faculty = Business Administration
32. If OldMajor = unknown And OldGPA = 3.01-3.49 And Gender = F Then Faculty = Fine-Arts-and-Science
33. If OldMajor = unknown And OldGPA = 3.01-3.49 And Gender = M Then Faculty = Master-of-Science
34. If OldMajor = unknown And OldGPA = 1.51-2.49 And Gender = F Then Faculty = Business-Administration
35. If OldMajor = unknown And OldGPA = 1.51-2.49 And Gender = M Then Faculty = ICT
36. If OldMajor = Bachelor-of-Householding Then Faculty = Master-of-Business-Administration
37. OldMajor = Bachelor-of-Information-Science Then Faculty = Master-of-Business-Administration
38. If OldMajor = Libralian-Science Then Faculty = Master-of-Science
39. If OldMajor = Bachelor-of-Art Then Faculty = Master-of-Business-Administration

40. If OldMajor = B.ED And Gender = F Then Faculty = Master-of-Business-Administration
41. If OldMajor = B.ED And Gender = M Then Faculty = Master-of-Science
42. If OldMajor = Bachelor-of-Supporting-Aggriculture-and-Cooperation Then Faculty = Fine-Arts-and-Science
43. If OldMajor = IT Then Faculty = Master-of-Science
44. If OldMajor = High-Diploma And Gender = F Then Faculty = Business-Administration
45. If OldMajor = High-Diploma And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
46. If OldMajor = High-Diploma And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Master-of-Engineering
47. If OldMajor = High-Diploma And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Master-of-Engineering
48. If OldMajor = High-Diploma And Gender = M And OldGPA = 3.01.3.49 Then Faculty = Master-of-Engineering
49. If OldMajor = High-Diploma And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Master-of-Engineering
50. If OldMajor = Bachelor-of-Economics Then Faculty = Master-of-Economics
51. If OldMajor = Bachelor-of-Art(Secretorial) Then Faculty =Master-of-Business-Administration
52. If OldMajor = B.SC Then Faculty = Master-of-Science
53. If OldMajor = Mattayom-5 Then Faculty = Master-of-Science
54. If OldMajor = Bachelor-of-Communication-Art Then Faculty = Master-of-Communication-Art
55. If OldMajor = B.EG Then Faculty= Master-of-Science
56. If OldMajor = B.BA And OldGPA = 0.00-1.50 Then Faculty = Master-of-Science
57. If OldMajor = B.BA And OldGPA = 2.50-3.00 Then Faculty = Master-of-Business-Administration
58. If OldMajor = B.BA And OldGPA = 3.50-4.00 Then Faculty = Master-of-Science

59. If OldMajor = B.BA And OldGPA = 3.01-3.49 Then Faculty = Master-of-Business-Administration
60. If OldMajor = B.BA And OldGPA = 1.51-2.49 Then Faculty = Master-of-Business-Administration
61. If OldMajor = Ed.-Special-school Then Faculty = Communication-Art
62. If OldMajor = B.Nursing Then Faculty = Master-of-Science
63. If OldMajor = Under-Bachelor Then Faculty = Business-Administration
64. If OldMajor = Bachelor-degree Then Faculty = Master-of-Science
65. If OldMajor = Bachelor-of-Educational And OldGPA = 0.00-1.50 Then Faculty = Master-of-Education
66. If OldMajor = Bachelor-of-Educational And OldGPA = 2.50-3.00 Then Faculty = Master-of-Communication-Art
67. If OldMajor = Bachelor-of-Educational And OldGPA = 3.50-4.00 Then Faculty = Master-of-Education
68. If OldMajor = Bachelor-of-Educational And OldGPA = 3.01-3.49 Then Faculty = Master-of-Science
69. If OldMajor = Bachelor-of-Educational And OldGPA = 1.51-2.49 Then Faculty = Master-of-Education
70. If OldMajor = B.PH Then Faculty = Fine-Arts
71. If OldMajor = B.Accounting And OldGPA = 0.00-1.50 Then Faculty = Master-of-Business-Administration
72. If OldMajor = B.Accounting And OldGPA = 2.50-3.00 Then Faculty = Master-of-Business-Administration
73. If OldMajor = B.Accounting And OldGPA = 3.50-4.00 Then Faculty = Master-of-Science
74. If OldMajor = B.Accounting And OldGPA = 3.01-3.49 Then Faculty = Master-of-Business-Administration
75. If OldMajor = B.Accounting And OldGPA = 1.51-2.49 Then Faculty = Master-of-Business-Administration
76. If OldMajor = High-School And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration

77. If OldMajor = High-School And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Fine-Arts-and-Science
78. If OldMajor = High-School And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Fine-Arts-and-Science
79. If OldMajor = High-School And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
80. If OldMajor = High-School And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Fine-Arts-and-Science
81. If OldMajor = High-School And Gender = M Then Faculty = Communication-Art
82. If OldMajor = Social-Science Then Master-of-Communication-Art
83. If OldMajor = B.ED And OldGPA = 0.00-1.50 Then Faculty = Master-of-Education
84. If OldMajor = B.ED And OldGPA = 2.50-3.00 Then Faculty = Master-of-Business-Administration
85. If OldMajor = B.ED And OldGPA = 3.50-4.00 Then Faculty = Master-of-Education
86. If OldMajor = B.ED And OldGPA = 3.01-3.49 Then Faculty = Master-of-Communication-Art
87. If OldMajor = B.ED And OldGPA = 1.51-2.49 Then Faculty= Laws

There are 87 rules in decision tree in test version 1.

Remark:

The number of rules equals to the number of leaves in decision tree.

3.4.2. Version 2

The details of version 2 are following:

ConfidenceFactor = 0.5

MinnumObj = 10

Evaluating method: 10-fold cross validation

Scheme: weka.classifiers.trees.J48 -C 0.5 -M 10

Relation: Student

Instances: 7778

Attributes: 5

Gender

OldMajor

OldGPA

EdType

Faculty

Classifier model (full training set)

J48 pruned tree

OldMajor = Medium-Diploma: Accouting (2.0/1.0)

OldMajor = B.Art: Communication-Art (1.0)

OldMajor = Thai-art-acting: Master-of-Education (1.0)

OldMajor = Fundamental-Diploma

| Gender = F

| | OldGPA = 0.00-1.50: Business-Administration (86.0/34.0)

| | OldGPA = 2.50-3.00: Accouting (129.0/98.0)

| | OldGPA = 3.50-4.00: Accouting (16.0/10.0)

| | OldGPA = 3.01-3.49: Accouting (71.0/46.0)

| | OldGPA = 1.51-2.49: Business-Administration (338.0/227.0)

| Gender = M

| | OldGPA = 0.00-1.50: Business-Administration (81.0/56.0)

| | OldGPA = 2.50-3.00: IT (68.0/51.0)

| | OldGPA = 3.50-4.00: Accouting (10.0/8.0)

| | OldGPA = 3.01-3.49: Fine-Arts (29.0/22.0)

| | OldGPA = 1.51-2.49: Communication-Art (214.0/164.0)

OldMajor = B.BA.(Accounting): Master-of-Business-Administration (1.0)

OldMajor = B.-Laws: Master-of-Laws (248.0/21.0)

OldMajor = B.EG.: Master-of-Science (5.0/1.0)
OldMajor = B.Indurtrial: Master-of-Science (5.0)
OldMajor = High-School(Sc.)
| Gender = F: Business-Administration (420.0/297.0)
| Gender = M
| | OldGPA = 0.00-1.50: Laws (20.0/15.0)
| | OldGPA = 2.50-3.00: Business-Administration (64.0/49.0)
| | OldGPA = 3.50-4.00: Business-Administration (6.0/4.0)
| | OldGPA = 3.01-3.49: Business-Administration (44.0/28.0)
| | OldGPA = 1.51-2.49: IT (123.0/97.0)
OldMajor = Master-Degree
| OldGPA = 0.00-1.50: Doctor-of-BA (19.0/11.0)
| OldGPA = 2.50-3.00: Doctor-of-BA (1.0)
| OldGPA = 3.50-4.00: Doctor-of-BA (11.0/6.0)
| OldGPA = 3.01-3.49: Doctor-of-Business-Administration (8.0/5.0)
| OldGPA = 1.51-2.49: Master-of-Laws (1.0)
OldMajor = Human-Science: Doctor-of-Business-Info. (3.0/2.0)
OldMajor = Bachelor-of-Architectural: Master-of-Science (1.0)
OldMajor = Bachelor-of-Statistics: Master-of-Science (1.0)
OldMajor = Bachelor-of-Industrial-Science: Master-of-Business-Administration (2.0/1.0)
OldMajor = Bachelor-of-Art(Thai-poem): Doctor-of-Communication-Art (1.0)
OldMajor = B.Politicial-Science: Master-of-Science (1.0)
OldMajor = Politician-Science: Master-of-Science (8.0/5.0)
OldMajor = High-School(Art.)
| Gender = F
| | OldGPA = 0.00-1.50: Business-Administration (43.0/28.0)
| | OldGPA = 2.50-3.00: Business-Administration (161.0/116.0)
| | OldGPA = 3.50-4.00: Fine-Arts-and-Science (36.0/23.0)
| | OldGPA = 3.01-3.49: Fine-Arts-and-Science (84.0/57.0)
| | OldGPA = 1.51-2.49: Business-Administration (373.0/259.0)
| Gender = M: Business-Administration (424.0/305.0)

OldMajor = unknown
| OldGPA = 0.00-1.50: Business-Administration (23.0/14.0)
| OldGPA = 2.50-3.00: Communication-Art (25.0/19.0)
| OldGPA = 3.50-4.00: Doctor-of-Business-Info. (11.0/9.0)
| OldGPA = 3.01-3.49: Master-of-Science (25.0/16.0)
| OldGPA = 1.51-2.49
| | Gender = F: Business-Administration (40.0/28.0)
| | Gender = M: IT (19.0/14.0)
OldMajor = Bachelor-of-Householding: Master-of-Business-Administration (2.0/1.0)
OldMajor = Bachelor-of-Information-Science: Master-of-Business-Administration (2.0/1.0)
OldMajor = Libralian-Science: Master-of-Science (1.0)
OldMajor = Bachelor-of-Art: Master-of-Business-Administration (43.0/24.0)
OldMajor = B.ED: Master-of-Science (4.0/2.0)
OldMajor = Bachelor-of-Supporting-Aggriculture-and-Cooperation: Fine-Arts-and-Science (2.0/1.0)
OldMajor = ICT: Master-of-Science (14.0/4.0)
OldMajor = High-Diploma
| Gender = F: Business-Administration (405.0/193.0)
| Gender = M
| | OldGPA = 0.00-1.50: Business-Administration (84.0/42.0)
| | OldGPA = 2.50-3.00: Master-of-Engineering (15.0/7.0)
| | OldGPA = 3.50-4.00: Master-of-Engineering (2.0)
| | OldGPA = 3.01-3.49: Master-of-Engineering (7.0/3.0)
| | OldGPA = 1.51-2.49: Master-of-Engineering (172.0/103.0)
OldMajor = Bachelor-of-Economics: Master-of-Economics (16.0/7.0)
OldMajor = Bachelor-of-Art(Secretorial): Master-of-Business-Administration (2.0)
OldMajor = B.SC: Master-of-Science (48.0/16.0)
OldMajor = Mattayom-5: Master-of-Science (1.0)

OldMajor = Bachelor-of-Communication-Art: Master-of-Communication-Art (47.0/19.0)
OldMajor = B.EG: Master-of-Science (44.0/5.0)
OldMajor = B.BA.
| OldGPA = 0.00-1.50: Master-of-Science (7.0/4.0)
| OldGPA = 2.50-3.00: Master-of-Business-Administration (24.0/4.0)
| OldGPA = 3.50-4.00: Master-of-Science (50.0/21.0)
| OldGPA = 3.01-3.49: Master-of-Business-Administration (93.0/40.0)
| OldGPA = 1.51-2.49: Master-of-Business-Administration (1.0)
OldMajor = Ed.-Special-school: Communication-Art (154.0/90.0)
OldMajor = B.Nursing: Master-of-Science (2.0)
OldMajor = Under-Bachelor: Business-Administration (2.0/1.0)
OldMajor = Bachelor-degree
| OldGPA = 0.00-1.50: Master-of-Science (6.0/2.0)
| OldGPA = 2.50-3.00: Laws (10.0/7.0)
| OldGPA = 3.50-4.00: Master-of-Science (9.0/4.0)
| OldGPA = 3.01-3.49: Master-of-Science (20.0/12.0)
| OldGPA = 1.51-2.49: Business-Administration (8.0/5.0)
OldMajor = Bachelor-of-Educational
| Gender = F: Master-of-Education (10.0/6.0)
| Gender = M: Master-of-Science (12.0/6.0)
OldMajor = B.PH: Fine-Arts (1.0)
OldMajor = B.Accounting: Master-of-Business-Administration (35.0/15.0)
OldMajor = High-School
| Gender = F
| | OldGPA = 0.00-1.50: Business-Administration (210.0/126.0)
| | OldGPA = 2.50-3.00: Fine-Arts-and-Science (405.0/300.0)
| | OldGPA = 3.50-4.00: Fine-Arts-and-Science (163.0/122.0)
| | OldGPA = 3.01-3.49: Business-Administration (266.0/198.0)
| | OldGPA = 1.51-2.49: Fine-Arts-and-Science (805.0/591.0)
| Gender = M
| | OldGPA = 0.00-1.50: Communication-Art (236.0/174.0)

| | OldGPA = 2.50-3.00: Communication-Art (241.0/191.0)
| | OldGPA = 3.50-4.00: Business-Administration (59.0/46.0)
| | OldGPA = 3.01-3.49: Business-Administration (133.0/108.0)
| | OldGPA = 1.51-2.49: Communication-Art (586.0/450.0)
OldMajor = Social-Science: Master-of-Communication-Art (1.0)
OldMajor = B.ED.: Master-of-Education (15.0/8.0)

Number of Leaves : 95
Size of the tree : 114

Summary

Correctly Classified Instances	5255	67.5624 %
Incorrectly Classified Instances	2523	32.4376 %
Total Number of Instances	7778	

Time taken to build model: 0.16 seconds

Extracted Rules from Decision Tree in Version 2:

1. If OldMajor = Medium-Diploma Then Faculty = Accouting
2. If OldMajor = B.Art Then Faculty = Communication-Art
3. If OldMajor = Thai-art-acting Then Faculty = Master-of-Education
4. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
5. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
6. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Accounting
7. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Accounting
8. If OldMajor = Fundamental-Diploma And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Accounting
9. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Business-Administration

10. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 2.50-3.00 Then Faculty = ICT
11. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Accounting
12. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Fine-Arts
13. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Communication-Art
14. If OldMajor = B.BA.(Accounting) Then Faculty = Master-of-Business-Administration
15. If OldMajor = B.-Laws Then Faculty = Master-of-Laws
16. If OldMajor = B.EG. Then Faculty= Master-of-Science
17. If OldMajor = B.Indrustrial Then Faculty = Master-of-Science
18. If OldMajor = High-School(Sc.) And Gender = F Then Faculty = Business-Administration
19. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Laws
20. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Business-Administration
21. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Business-Administration
22. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
23. OldMajor = High-School(Sc.) And Gender = M And OldGPA = 1.51-2.49 Then Faculty = ICT
24. If OldMajor = Master-Degree And OldGPA = 0.00-1.50 Then Faculty = Doctor-of-BA
25. If OldMajor = Master-Degree And OldGPA = 2.50-3.00 Then Faculty = Doctor-of-BA
26. If OldMajor = Master-Degree And OldGPA = 3.50-4.00 Then Faculty = Doctor-of-BA

27. If OldMajor = Master-Degree And OldGPA = 3.01-3.49 Then Faculty = Doctor-of-Business-Administration
28. If OldMajor = Master-Degree And OldGPA = 1.51-2.49 Then Faculty = Master-of-Laws
29. If OldMajor = Human-Science Then Faculty = Doctor-of-Business-Info.
30. If OldMajor = Bachelor-of-Architectural Then Faculty = Master-of-Science
31. If OldMajor = Bachelor-of-Statistics Then Faculty = Master-of-Science
32. OldMajor = Bachelor-of-Industrial-Science Then Faculty = Master-of-Business-Administration
33. If OldMajor = Bachelor-of-Art(Thai-poem) Then Faculty = Doctor-of-Communication-Art
34. If OldMajor = B.Politicial-Science Then Faculty = Master-of-Science
35. If OldMajor = Politician-Science Then Master-of-Science
36. If OldMajor = High-School(Art.) And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
37. If OldMajor = High-School(Art.) And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Business-Administration
38. If OldMajor = High-School(Art.) And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Fine-Arts-and-Science
39. If OldMajor = High-School(Art.) And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Fine-Arts-and-Science
40. If OldMajor = High-School(Art.) And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
41. If OldMajor = High-School(Art.) And Gender = M Then Faculty = Business-Administration
42. If OldMajor = unknown And OldGPA = 0.00-1.50 Then Faculty= Business-Administration
43. If OldMajor = unknown And OldGPA = 2.50-3.00 Then Faculty = Communication-Art
44. If OldMajor = unknown And OldGPA = 3.50-4.00 Then Faculty = Doctor-of-Business-Info

45. If OldMajor = unknown And OldGPA = 3.01-3.49 Then Faculty = Master-of-Science
46. If OldMajor = unknown And OldGPA = 1.51-2.49 And Gender = F Then Faculty = Business-Administration
47. If OldMajor = unknown And OldGPA = 1.51-2.49 And Gender = M Then Faculty = ICT
48. If OldMajor = Bachelor-of-Householding Then Faculty = Master-of-Business-Administration
49. If OldMajor = Bachelor-of-Information-Science Then Faculty = Master-of-Business-Administration
50. If OldMajor = Libralian-Science Then Faculty = Master-of-Science
51. If OldMajor = Bachelor-of-Art Then Faculty = Master-of-Business-Administration
52. If OldMajor = B.ED Then Faculty = Master-of-Science
53. If OldMajor = Bachelor-of-Supporting-Aggriculture-and-Cooperation Then Faculty = Fine-Arts-and-Science
54. If OldMajor = IT Then Faculty = Master-of-Science
55. If OldMajor = High-Diploma And Gender = F Then Faculty = Business-Administration
56. If OldMajor = High-Diploma And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
57. If OldMajor = High-Diploma And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Master-of-Engineering
58. If OldMajor = High-Diploma And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Master-of-Engineering
59. If OldMajor = High-Diploma And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Master-of-Engineering
60. If OldMajor = High-Diploma And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Master-of-Engineering
61. If OldMajor = Bachelor-of-Economics Then Faculty = Master-of-Economics

62. If OldMajor = Bachelor-of-Art(Secretorial) Then Faclulty = Master-of-Business-Administration
63. If OldMajor = B.SC Then Faculty = Master-of-Science
64. If OldMajor = Mattayom-5 Then Faculty = Master-of-Science
65. If OldMajor = Bachelor-of-Communication-Art Then Faculty = Master-of-Communication-Art
66. If OldMajor = B.EG Then Faculty = Master-of-Science
67. If OldMajor = B.BA. And OldGPA = 0.00-1.50 Then Faculty = Master-of-Science
68. If OldMajor = B.BA. And OldGPA = 2.50-3.00 Then Faculty = Master-of-Business-Administration
69. If OldMajor = B.BA. And OldGPA = 3.50-4.00 Then Faculty = Master-of-Science
70. If OldMajor = B.BA. And OldGPA = 3.01-3.49 Then Faculty = Master-of-Business-Administration
71. If OldMajor = B.BA. And OldGPA = 1.51-2.49 Then Faculty = Master-of-Business-Administration
72. If OldMajor = Ed.-Special-school Then Faculty = Communication-Art
73. If OldMajor = B.Nursing Then Faculty = Master-of-Science
74. If OldMajor = Under-Bachelor Then Faculty = Business-Administration
75. If OldMajor = Bachelor-degree And OldGPA = 0.00-1.50 Then Faculty = Master-of-Science
76. If OldMajor = Bachelor-degree And OldGPA = 2.50-3.00 Then Faculty = Laws
77. If OldMajor = Bachelor-degree And OldGPA = 3.50-4.00 Then Faculty = Master-of-Science
78. If OldMajor = Bachelor-degree And OldGPA = 3.01-3.49 Then Faculty = Master-of-Science
79. If OldMajor = Bachelor-degree And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
80. If OldMajor = Bachelor-of-Educational And Gender = F Then Faculty = Master-of-Education

81. If OldMajor = Bachelor-of-Educational And Gender = F Then Faculty = Master-of-Science
82. If OldMajor = B.PH Then Faculty = Fine-Arts
83. If OldMajor = B.Accounting Then Faculty = Master-of-Business-Administration
84. If OldMajor = High-School And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
85. If OldMajor = High-School And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Fine-Arts-and-Science
86. If OldMajor = High-School And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Fine-Arts-and-Science
87. If OldMajor = High-School And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
88. If OldMajor = High-School And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Fine-Arts-and-Science
89. If OldMajor = High-School And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Communication-Art
90. If OldMajor = High-School And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Communication-Art
91. If OldMajor = High-School And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Business-Administration
92. If OldMajor = High-School And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
93. If OldMajor = High-School And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Communication-Art
94. If OldMajor = Social-Science Then Faculty = Master-of-Communication-Art
95. If OldMajor = B.ED. Then Faculty = Master-of-Education

There are 95 rules in decision tree from test version 2.

Remark:

The number of rules equals to the number of leaves in decision tree.

3.4.3. Version 3

The details of version 3 are following:

ConfidenceFactor = 0.5

MinnumObj = 50

Evaluating method: 10-fold cross validation

Scheme: weka.classifiers.trees.J48 -C 0.5 -M 50

Relation: Student

Instances: 7778

Attributes: 5

Gender

OldMajor

OldGPA

EdType

Faculty

Classifier model (full training set)

J48 pruned tree

OldMajor = Medium-Diploma: Accouting (2.0/1.0)

OldMajor = B.Art: Communication-Art (1.0)

OldMajor = Thai-art-acting: Master-of-Education (1.0)

OldMajor = Fundamental-Diploma

| Gender = F

| | OldGPA = 0.00-1.50: Business-Administration (86.0/34.0)

| | OldGPA = 2.50-3.00: Accouting (129.0/98.0)

| | OldGPA = 3.50-4.00: Accouting (16.0/10.0)

| | OldGPA = 3.01-3.49: Accouting (71.0/46.0)

| | OldGPA = 1.51-2.49: Business-Administration (338.0/227.0)

| Gender = M

| | OldGPA = 0.00-1.50: Business-Administration (81.0/56.0)

| | OldGPA = 2.50-3.00: IT (68.0/51.0)
| | OldGPA = 3.50-4.00: Accouting (10.0/8.0)
| | OldGPA = 3.01-3.49: Fine-Arts (29.0/22.0)
| | OldGPA = 1.51-2.49: Communication-Art (214.0/164.0)
OldMajor = B.BA.(Accounting): Master-of-Business-Administration (1.0)
OldMajor = B.-Laws: Master-of-Laws (248.0/21.0)
OldMajor = B.EG.: Master-of-Science (5.0/1.0)
OldMajor = B.Indurtrial: Master-of-Science (5.0)
OldMajor = High-School(Sc.)
| Gender = F: Business-Administration (420.0/297.0)
| Gender = M
| | OldGPA = 0.00-1.50: Laws (20.0/15.0)
| | OldGPA = 2.50-3.00: Business-Administration (64.0/49.0)
| | OldGPA = 3.50-4.00: Business-Administration (6.0/4.0)
| | OldGPA = 3.01-3.49: Business-Administration (44.0/28.0)
| | OldGPA = 1.51-2.49: IT (123.0/97.0)
OldMajor = Master-Degree: Doctor-of-BA (40.0/25.0)
OldMajor = Human-Science: Doctor-of-Business-Info. (3.0/2.0)
OldMajor = Bachelor-of-Architectural: Master-of-Science (1.0)
OldMajor = Bachelor-of-Statistics: Master-of-Science (1.0)
OldMajor = Bachelor-of-Industrial-Science: Master-of-Business-Administration (2.0/1.0)
OldMajor = Bachelor-of-Art(Thai-poem): Doctor-of-Communication-Art (1.0)
OldMajor = B.Politicial-Science: Master-of-Science (1.0)
OldMajor = Politician-Science: Master-of-Science (8.0/5.0)
OldMajor = High-School(Art.)
| Gender = F
| | OldGPA = 0.00-1.50: Business-Administration (43.0/28.0)
| | OldGPA = 2.50-3.00: Business-Administration (161.0/116.0)
| | OldGPA = 3.50-4.00: Fine-Arts-and-Science (36.0/23.0)
| | OldGPA = 3.01-3.49: Fine-Arts-and-Science (84.0/57.0)
| | OldGPA = 1.51-2.49: Business-Administration (373.0/259.0)

| Gender = M: Business-Administration (424.0/305.0)
OldMajor = unknown: Business-Administration (143.0/115.0)
OldMajor = Bachelor-of-Householding: Master-of-Business-Administration (2.0/1.0)
OldMajor = Bachelor-of-Information-Science: Master-of-Business-Administration (2.0/1.0)
OldMajor = Libralian-Science: Master-of-Science (1.0)
OldMajor = Bachelor-of-Art: Master-of-Business-Administration (43.0/24.0)
OldMajor = B.ED: Master-of-Science (4.0/2.0)
OldMajor = Bachelor-of-Supporting-Aggriculture-and-Cooperation: Fine-Arts-and-Science (2.0/1.0)
OldMajor = ICT: Master-of-Science (14.0/4.0)
OldMajor = High-Diploma
| Gender = F: Business-Administration (405.0/193.0)
| Gender = M
| | OldGPA = 0.00-1.50: Business-Administration (84.0/42.0)
| | OldGPA = 2.50-3.00: Master-of-Engineering (15.0/7.0)
| | OldGPA = 3.50-4.00: Master-of-Engineering (2.0)
| | OldGPA = 3.01-3.49: Master-of-Engineering (7.0/3.0)
| | OldGPA = 1.51-2.49: Master-of-Engineering (172.0/103.0)
OldMajor = Bachelor-of-Economics: Master-of-Economics (16.0/7.0)
OldMajor = Bachelor-of-Art(Secretorial): Master-of-Business-Administration (2.0)
OldMajor = B.SC: Master-of-Science (48.0/16.0)
OldMajor = Mattayom-5: Master-of-Science (1.0)
OldMajor = Bachelor-of-Communication-Art: Master-of-Communication-Art (47.0/19.0)
OldMajor = B.EG: Master-of-Science (44.0/5.0)
OldMajor = B.BA.
| OldGPA = 0.00-1.50: Master-of-Science (7.0/4.0)
| OldGPA = 2.50-3.00: Master-of-Business-Administration (24.0/4.0)
| OldGPA = 3.50-4.00: Master-of-Science (50.0/21.0)

| OldGPA = 3.01-3.49: Master-of-Business-Administration (93.0/40.0)
| OldGPA = 1.51-2.49: Master-of-Business-Administration (1.0)
OldMajor = Ed.-Special-school: Communication-Art (154.0/90.0)
OldMajor = B.Nursing: Master-of-Science (2.0)
OldMajor = Under-Bachelor: Business-Administration (2.0/1.0)
OldMajor = Bachelor-degree: Master-of-Science (53.0/34.0)
OldMajor = Bachelor-of-Educational: Master-of-Science (22.0/13.0)
OldMajor = B.PH: Fine-Arts (1.0)
OldMajor = B.Accounting: Master-of-Business-Administration (35.0/15.0)
OldMajor = High-School
| Gender = F
| | OldGPA = 0.00-1.50: Business-Administration (210.0/126.0)
| | OldGPA = 2.50-3.00: Fine-Arts-and-Science (405.0/300.0)
| | OldGPA = 3.50-4.00: Fine-Arts-and-Science (163.0/122.0)
| | OldGPA = 3.01-3.49: Business-Administration (266.0/198.0)
| | OldGPA = 1.51-2.49: Fine-Arts-and-Science (805.0/591.0)
| Gender = M
| | OldGPA = 0.00-1.50: Communication-Art (236.0/174.0)
| | OldGPA = 2.50-3.00: Communication-Art (241.0/191.0)
| | OldGPA = 3.50-4.00: Business-Administration (59.0/46.0)
| | OldGPA = 3.01-3.49: Business-Administration (133.0/108.0)
| | OldGPA = 1.51-2.49: Communication-Art (586.0/450.0)
OldMajor = Social-Science: Master-of-Communication-Art (1.0)
OldMajor = B.ED.: Master-of-Education (15.0/8.0)

Number of Leaves : 81
Size of the tree : 95

Summary

Correctly Classified Instances	5293	68.0509 %
Incorrectly Classified Instances	2485	31.9491 %

Total Number of Instances 7778

Time taken to build model: 0.03 seconds

Extracted Rules from Decision Tree in Version 3

1. If OldMajor = Medium-Diploma Then Faculty = Accouting
2. If OldMajor = B.Art Then Faculty = Communication-Art
3. If OldMajor = Thai-art-acting Then Faculty = Master-of-Education
4. If OldMajor = Fundamental-diploma And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
5. If OldMajor = Fundamental-diploma And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Accounting
6. If OldMajor = Fundamental-diploma And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Accounting
7. If OldMajor = Fundamental-diploma And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Accounting
8. If OldMajor = Fundamental-diploma And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
9. If OldMajor = Fundamental-diploma And Gender = M And OldGPA = 0.00-1.50 Then Business-Administration
10. If OldMajor = Fundamental-diploma And Gender = M And OldGPA = 2.50-3.00 Then ICT
11. If OldMajor = Fundamental-diploma And Gender = M And OldGPA = 3.50-4.00 Then Accounting
12. If OldMajor = Fundamental-diploma And Gender = M And OldGPA = 3.01-3.49 Then Fine-Arts
13. If OldMajor = Fundamental-diploma And Gender = M And OldGPA = 1.51-2.49 Then Communication-Art
14. If OldMajor = B.BA.(Accounting) Then Faculty = Master-of-Business-Administration
15. If OldMajor = B.-Laws Then Faculty = Master-of-Laws
16. If OldMajor = B.EG. Then Faculty = Master-of-Science
17. If OldMajor = B.Indrustrial Then Faculty = Master-of-Science

18. If OldMajor = High-School(Sc.) And Gender = F Then Faculty = Business-Administration
19. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Laws
20. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Business-Administration
21. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Business-Administration
22. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
23. If OldMajor = High-School(Sc.) And Gender = M And OldGPA = 1.51-2.49 Then Faculty = ICT
24. If OldMajor = Master-Degree Then Faculty = Doctor-of-BA
25. If OldMajor = Human-Science Then Faculty = Doctor-of-Business-Info.
26. If OldMajor = Bachelor-of-Architectural Then Faculty = Master-of-Science
27. If OldMajor = Bachelor-of-Statistics Then Faculty = Master-of-Science
28. If OldMajor = Bachelor-of-Industrial-Science Then Faculty = Master-of-Business-Administration
29. If OldMajor = Bachelor-of-Art(Thai-poem) Then Faculty = Doctor-of-Communication-Art
30. If OldMajor = B.Politicial-Science Then Faculty = Master-of-Science
31. If OldMajor = Politician-Science Then Faculty = Master-of-Science
32. If OldMajor =High-School(Art.) And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
33. If OldMajor =High-School(Art.) And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Business-Administration
34. If OldMajor =High-School(Art.) And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Fine-Arts-and-Science
35. If OldMajor =High-School(Art.) And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Fine-Arts-and-Science

36. If OldMajor =High-School(Art.) And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
37. If OldMajor =High-School(Art.) And Gender = M Then Faculty = Business-Administration
38. If OldMajor = unknown Then Faculty = Business-Administration
39. If OldMajor = Bachelor-of-Householding Then Faculty = Master-of-Business-Administration
40. If OldMajor = Bachelor-of-Information-Science Then Faculty = Master-of-Business-Administration
41. If OldMajor = Libralian-Science Then Faculty = Master-of-Science
42. If OldMajor = Bachelor-of-Art Then Faculty = Master-of-Business-Administration
43. If OldMajor = B.ED Then Faculty = Master-of-Science
44. If OldMajor = Bachelor-of-Supporting-Aggriculture-and-Cooperation Then Faculty = Fine-Arts-and-Science
45. If OldMajor = IT Then Faculty = Master-of-Science
46. If OldMajor = High-School And Gender = F Then Faculty = Business-Administration
47. If OldMajor = High-School And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
48. If OldMajor = High-School And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Master-of-Engineering
49. If OldMajor = High-School And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Master-of-Engineering
50. If OldMajor = High-School And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Master-of-Engineering
51. If OldMajor = High-School And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Master-of-Engineering
52. If OldMajor = Bachelor-of-Economics Then Faculty = Master-of-Economics
53. If OldMajor = Bachelor-of-Art(Secretorial) Then Faculty = Master-of-Business-Administration

54. If OldMajor = B.SC Then Faculty = Master-of-Science
55. If OldMajor = Mattayom-5 Then Faculty = Master-of-Science
56. If OldMajor = Bachelor-of-Communication-Art Then Faculty = Master-of-Communication-Art
57. If OldMajor = B.EG Then Faculty = Master-of-Science
58. If OldMajor = B.BA. And OldGPA = 0.00-1.50 Then Faculty = Master-of-Science
59. If OldMajor = B.BA. And OldGPA = 2.50-3.00 Then Faculty = Master-of-Business-Administration
60. If OldMajor = B.BA. And OldGPA = 3.50-4.00 Then Faculty = Master-of-Science
61. If OldMajor = B.BA. And OldGPA = 3.01-3.49 Then Faculty = Master-of-Business-Administration
62. If OldMajor = B.BA. And OldGPA = 1.51-2.49 Then Faculty = Master-of-Business-Administration
63. If OldMajor = Ed.-Special-school Then Faculty = Communication-Art
64. If OldMajor = B.Nursing Then Faculty = Master-of-Science
65. If OldMajor = Under-Bachelor Then Faculty = Business-Administration
66. If OldMajor = Bachelor-degree Then Faculty = Master-of-Science
67. If OldMajor = Bachelor-of-Educational Then Faculty = Master-of-Science
68. If OldMajor = B.PH Then Faculty = Fine-Arts
69. If OldMajor = B.Accounting Then Faculty = Master-of-Business-Administration
70. If OldMajor = High-School And Gender = F And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
71. If OldMajor = High-School And Gender = F And OldGPA = 2.50-3.00 Then Faculty = Fine-Arts-and-Science
72. If OldMajor = High-School And Gender = F And OldGPA = 3.50-4.00 Then Faculty = Fine-Arts-and-Science
73. If OldMajor = High-School And Gender = F And OldGPA = 3.01-3.49 Then Faculty = Business-Administration

74. If OldMajor = High-School And Gender = F And OldGPA = 1.51-2.49 Then Faculty = Fine-Arts-and-Science
75. If OldMajor = High-School And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Communication-Art
76. If OldMajor = High-School And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Communication-Art
77. If OldMajor = High-School And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Business-Administration
78. If OldMajor = High-School And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
79. If OldMajor = High-School And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Communication-Art
80. If OldMajor = Social-Science Then Faculty = Master-of-Communication-Art
81. If OldMajor = B.ED. Then Faculty = Master-of-Education

There are 81 rules in decision tree version 3.

Remark:

We extract rules from decision tree by traversing through the tree. The number of rules equals to the number of leaves in decision tree.

3.5. The Classification Model

We have improved the decision tree model. The decision tree model is built up with significant result of evaluation.

3.5.1. Dataset description

This dataset is data about student's education type and student's GPA (Grade Point Average) in his old school. We used this dataset to explore the effect of student's education type and student's GPA in his old school to his studying

faculty. The result from this experiment can be used to help students in choosing appropriate faculty for them.

In order to efficiently measure the accuracy, we have reduced the size of dataset. In particular, there are 3751 records in dataset, in each record there are 5 attributes. The "faculty" attribute is used to be a class attribute. Moreover, the faculties, which focusing students are now majoring, are Accounting, Business-Administration, Communication-Art, Economics, Fine-Arts, Fine-Arts-and-Science, Information Technology, and Laws. Description of each attribute is showed in table 3.2.

Table 3.2: Properties of each attribute in dataset

Attribute	Description	Possible values of attribute
Gender	Student's gender	F (Female) , M (Male)
OldMajor	The major program that student got.	Medium-Diploma, Fundamental-Diploma, High-School(Sc.), High-School(Art.), High-School, High-Diploma, Ed.-Special-school, Diploma-of-Business-Admin, Diploma-of-Accounting, Diploma-of-Art, High-Diploma-of-Business-Admin, High-Diploma-of-Accounting
OldGPA	GPA that student got from old school.	0.00-1.50, 1.51-2.49, 2.50-3.00, 3.01-3.49, 3.50-4.00,
EdType	Student's education type	Diploma , HighSchool
Faculty	Faculty that student is studying, now.	Accounting, Business-Administration, Communication-Art, Economics, Fine-Arts, Fine-Arts-and-Science, IT, Laws

3.5.2. Experiment Details

The details are following:

ConfidenceFactor = 0.25

MinnumObj = 2

Evaluating method: 10-fold cross validation

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Student

Instances: 3751

Attributes: 5

Gender

OldMajor

OldGPA

EdType

Faculty

Classifier model (full training set)

J48 pruned tree

OldMajor = Medium-Diploma: Accounting (1.0)

OldMajor = Fundamental-Diploma

| Gender = F: Fine-Arts-and-Science (218.0/130.0)

| Gender = M

| | OldGPA = 0.00-1.50: Business-Administration (47.0/29.0)

| | OldGPA = 1.51-2.49: Communication-Art (133.0/92.0)

| | OldGPA = 2.50-3.00: IT (40.0/28.0)

| | OldGPA = 3.01-3.49: Business-Administration (20.0/15.0)

| | OldGPA = 3.50-4.00: Communication-Art (4.0/2.0)

OldMajor = High-School(Sc.): IT (648.0/277.0)

OldMajor = High-School(Art.)

| Gender = F: Communication-Art (653.0/360.0)

| Gender = M
| | OldGPA = 0.00-1.50: Laws (41.0/11.0)
| | OldGPA = 1.51-2.49: Business-Administration (192.0/138.0)
| | OldGPA = 2.50-3.00: Laws (57.0/33.0)
| | OldGPA = 3.01-3.49: Business-Administration (16.0/7.0)
| | OldGPA = 3.50-4.00: Business-Administration (3.0/1.0)
OldMajor = High-School: Business-Administration (918.0/147.0)
OldMajor = High-Diploma: IT (111.0/23.0)
OldMajor = Ed.-Special-school: Communication-Art (69.0/24.0)
OldMajor = Diploma-of-Business-Admin: Business-Administration (145.0/1.0)
OldMajor = Diploma-of-Accounting: Accounting (87.0)
OldMajor = Diploma-of-Art: Communication-Art (40.0)
OldMajor = High-Diploma-of-Business-Admin: Business-Administration (231.0)
OldMajor = High-Diploma-of-Accounting: Accounting (77.0)

Number of Leaves : 22
Size of the tree : 27
Time taken to build model: 0.06 seconds

Summary

Correctly Classified Instances	3155	**84.1162 %**
Incorrectly Classified Instances	595	15.8838 %

Extracted rules from decision tree

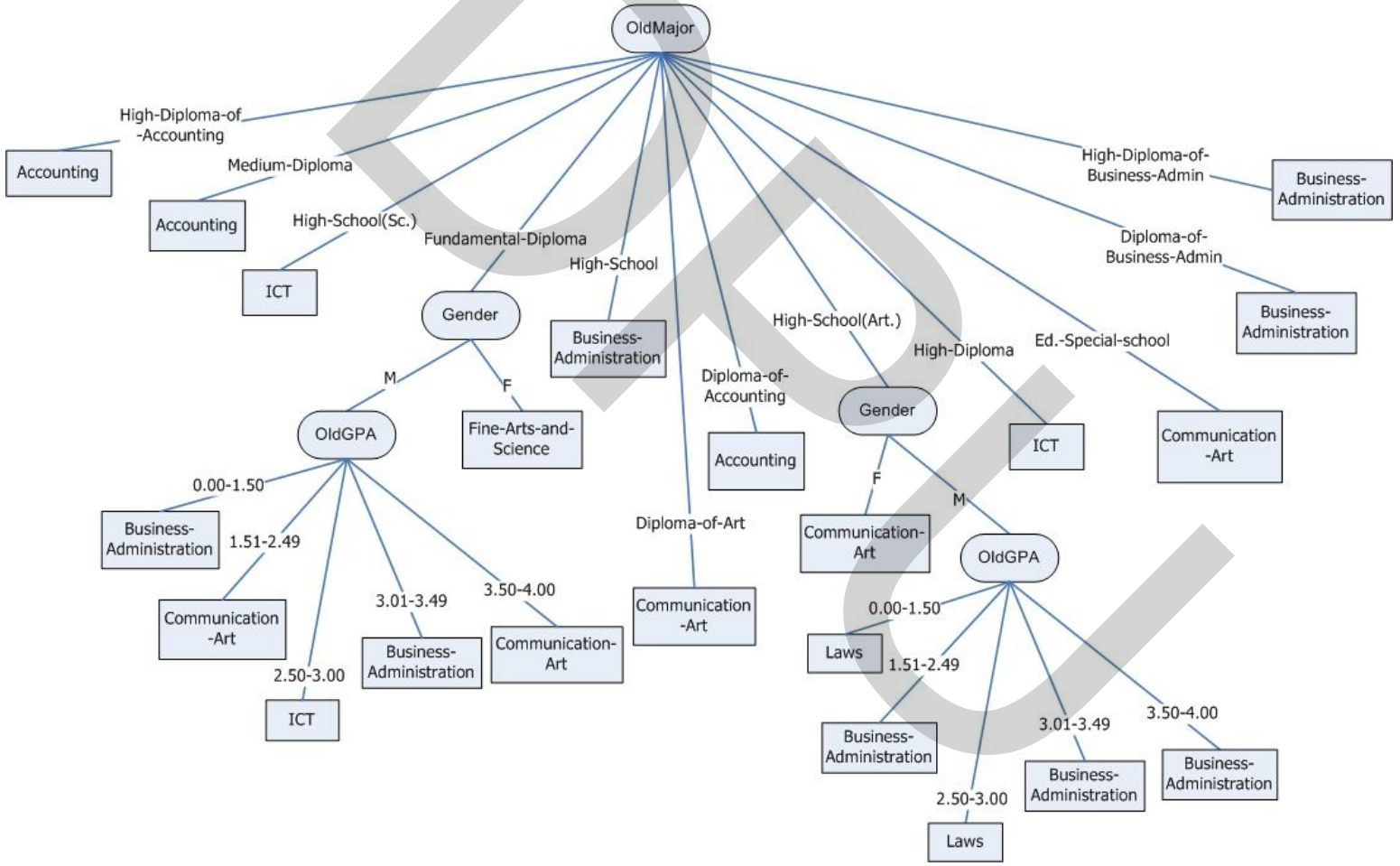
1. If OldMajor = Medium-Diploma Then Faculty = Accounting
2. If OldMajor = Fundamental-Diploma And Gender = F Then Faculty = Fine-Arts-and-Science
3. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Business-Administration
4. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Communication-Art

5. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 2.50-3.00 Then Faculty = ICT
6. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
7. If OldMajor = Fundamental-Diploma And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Communication-Art
8. If OldMajor = High-School(Sc.) Then Faculty = ICT
9. If OldMajor = High-School(Art.) And Gender = F Then Faculty = Communication-Art
10. If OldMajor = High-School(Art.) And Gender = M And OldGPA = 0.00-1.50 Then Faculty = Laws
11. If OldMajor = High-School(Art.) And Gender = M And OldGPA = 1.51-2.49 Then Faculty = Business-Administration
12. If OldMajor = High-School(Art.) And Gender = M And OldGPA = 2.50-3.00 Then Faculty = Laws
13. If OldMajor = High-School(Art.) And Gender = M And OldGPA = 3.01-3.49 Then Faculty = Business-Administration
14. If OldMajor = High-School(Art.) And Gender = M And OldGPA = 3.50-4.00 Then Faculty = Business-Administration
15. If OldMajor = High-School Then Faculty = Business-Administration
16. If OldMajor = High-Diploma Then Faculty = ICT
17. If OldMajor = Ed.-Special-school Then Faculty = Communication-Art
18. If OldMajor = Diploma-of-Business-Admin Then Faculty = Business-Administration
19. If OldMajor = Diploma-of-Accounting Then Faculty = Accounting
20. If OldMajor = Diploma-of-Art Then Faculty = Communication-Art
21. If OldMajor = High-Diploma-of-Business-Admin Then Faculty = Business-Administration
22. If OldMajor = High-Diploma-of-Accounting Then Faculty = Accounting

According to the evaluation result, the decision tree is reliable and returns significant results. We then proposed the decision tree as the classification model

for undergraduate program selection. The decision is shown in Figure 3.1. Each oval is represented for a factor to be concerned in classification. A rectangle is represented for a major. The classification to find out the most-appropriate major for a student is concerned several factors: *gender, old major in previous program, accumulative GPA in high school, type of studying (i.e. diploma, or high school)*, and *faculty to be studying*. The possible values of each factor are described in section 3.5.1.

Figure 3.1. A decision tree model to classifying a major for undergraduate program applicant



3.6. Summary

This chapter has described the learning process and presented the proposed classification model. In next chapter, we present the experiments developed to demonstrate the work and analyses the experimental results of using the model.

Chapter IV Case Study and Usage Analysis

This chapter presents the examples of using the classification model and testing to show that the classification model, developed based on the learning process, can be applied to work and support the education issues.

4.1 Introduction to Case Study

We have created two test cases in order to identifying at risk students. For each of test case we used data set encompassing students' profiles and applied with the model proposed in Chapter 3. Test case 1 is aimed to evaluate how the decision tree model can give the useful information to unsuccessful students. Test case 2 is aimed to analyze the accuracy of model in predicting poor academic performance. Each case is described below:

4.2 Test Case 1

Purpose and Goals

- To analyze the accuracy of model in predicting poor academic performance
- To use the processed information to be a suggestion for students who are applying higher education

Data Used for Model Development

- Juniors and seniors(2005 – 2006)
- 3,751 cases
- Average High School GPA of 2.33
- 56% female and 44% male

Variables Used for Model Development

- Educational status: Third-year and Fourth-year

- High School Performance
  - High school GPA
  - Old major
- Student Demographics
  - Gender
- Type of Feeder High School (Public, Private)

Methods

- Apply the decision tree model developed and described in Chapter 3
- Run nine experiments with different confidence and minimum number of object.

Results

According to the nine experiments, the average results of test case 1 show that correctly identified is 87.937% of at-risk students while misclassifying is 47.6%. The details of each experiment are shown in the following table.

Table 4.1: Percentage of correctly classified, incorrectly classified, and misclassifying

Confidence	Minimum number of instance	Correctly identifying	Incorrectly identifying	Misclassifying
0.25	2	91.45%	08.55%	41.02%
0.25	10	90.42%	09.58%	42.8%
0.25	50	89.63%	10.36%	47.05%
0.50	10	88.15%	11.85%	49.28%
0.50	30	88.02%	11.98%	49.55%
0.50	50	87.95%	12.05%	50.03%
0.75	5	86.45%	13.55%	47.85%
0.75	20	85.42%	14.58%	49.8%
0.75	50	83.95%	16.05%	51.02%

Implications for Deployment

The percentage of correctly identified is high (87.3%). It implies the model performs the classification of program selection for students with the high performance. The percentage of misclassifying is fairly low (47.6%). It implies the model failed to identify the class label, particularly suitable program, for a

student. This is due to incomplete data e.g. missing some attribute values, incorrect data.

4.3 Test Case 2

Purpose and Goals

- To analyze the accuracy of model in predicting poor academic performance
- To use information/data that occur before students enroll
- To lower cost by determining a group of potential students

Data Used for Model Development

- Freshmen (2004 – 2006)
- 4,652 cases
- Average High School GPA of 2.38
- 52% female and 48% male

Variables Used for Model Development

- Educational status: First-year
- High School Performance
  - High school GPA
  - Old major
- Student Demographics
  - Gender
- Type of Feeder High School (Public, Private)

Methods

- Apply the decision tree model developed and described in Chapter 3
- Run nine experiments with different confidence and minimum number of object.

Results

According to the nine experiments, the average results of test case 2 show that correctly identified is 83.2% of at-risk students while misclassifying is 55.2%. The details of each experiment are shown in the following table.

Table 4.2: Percentage of correctly classified, incorrectly classified, and misclassifying

Confidence	Minimum number of instance	Correctly identifying	Incorrectly identifying	Misclassifying
0.25	2	89.3%	10.7%	52.1%
0.25	10	87.25%	12.75	53.07%
0.25	50	85.4%	14.6%	53.18%
0.50	10	84.95%	15.05%	54.28%
0.50	30	83.6%	16.4%	54.55%
0.50	50	82.15%	17.85%	55.03%
0.75	5	79.88%	20.12%	57.35%
0.75	20	78.82%	21.18%	57.9%
0.75	50	77.45%	22.55%	59.34%

Implications for Deployment

The percentage of correctly identified is fairly high (83.2%). It implies the model can give the useful information to students about programs that they should take. The percentage of misclassifying is average (55.2%). It may be implied that at-risk students are identified with a high misclassification cost.

4.4 Summary

This chapter has presented two test cases created in order to present how precise the decision tree model is and give some ideas of use of the decision tree model to earn some benefits. The majority of benefits are fallen into the students.

Chapter V Conclusions and Future Work

This chapter provides the conclusions, some useful suggestions for future study and , future work of this research. Section 5.1 presents the overall conclusions. The future work are described in Section 5.2.

5.1 Conclusions

The issue we address may be one of the most important problems in education management today. Sophisticated system to support higher education, particularly, undergraduate programs has a lot of profit improvement potential, both by increasing opportunities of students and by making possible better academic plans for students.

We applied the technique which is a supervised learning technique that classifies data items composed of several attributes, for example, old major at pre-university, home address, school location, studying faculty, and GPA, into pre-defined class labels. This appropriate technique builds a classifier model that predicts future data trend. The required data is quite simple, and data set is collected from many sources such as university and old schools of students.

According to the research project, it shows that it is a challenge to an academic institute to adopt the techniques of data mining to decrease risk students and improve the quality of education system. A good system including the mining techniques, i.e. classification model, may provide the users with important competitive information. The users here include prospective students, current students, and academic staffs. In addition, the model could be incorporated with available management information system to tackling related issues in education system in Thailand.

5.2 Future Work

The following issues are interesting directions for future work:

1. Automatic process

   At present, the model is used in a semi-automatic way to acquire data from the operational source systems as a data set. The operational database within academic institutes periodically updated in monthly and yearly according to term period. The automatic process is expected to support data extraction, data transformation in data set format, and loading to the data set periodically. So, the researchers who are responding to the system can acquire the updated data for helping to planning and analyzing data which is consequence to rapidly making decision.

2. Association Rule Mining

   Association Rule Mining is another potential and challenge technique to mine data to study relationship of students within the same group of studying major. It is believed that the approach could benefit by taking each data group that obtained from classifying technique to find association with group of systems and taking that association to analyzed trend or to predicted students' behavior go on.

3. Information Visualization

   After a process of data mining, the researchers may be analyzed data to extract new knowledge. Thus, presentation of data is significant to uses. Now, the statistic results are represented the data result after data mining process, but it is not enough to display scatter group of data. Accordingly, in the future we can represent the scatter group of data in scatter plot format to help user forget the overall group of data obviously.

References

Anonymous. 2002. *Summary of Statistical Data of Students 1997-2001*. [Online]. Available : http://www.stat.mua.go.th/ebook/. Retrieved January 25, 2009.

Anonymous. 2004. *Compendium of Programs, Fields of Study, and Degrees.* [Online]. Available : http://www.stat.mua.go.th/ebook/. Retrieved January 25, 2009.

Agrawal, R., Imielinski, T., Swami A. 1993. *Mining Association Rules Between Sets of Items in Large Databases.* SIGMOD Conference, pp. 207-216.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth, Monterey, CA.

Champoaunjam, J. 2005. *Higher Education Data and Information Commission on Higher Education 2005*. [Online]. Available: http://www.stat.mua.go.th/ebook/. Retrieved January 25, 2009.

Chaudhuri, S. and Dayal, U. 1997. *An Overview of Data Warehousing and OLAP Technology*. SIGMOD Record 26(1), pp.65-74.

Deza, E., Deza, M. 2006. *Dictionary of Distances*. Elsevier. ISBN 0444520872 . Also Available at http://en.wikipedia.org/wiki/Distance. Retrieved August 18, 2009.

Giacinto, G., and Roli, F. 1997. *Ensembles of Neural Networks for soft classification of remote sensing images.* Proceeding of the European Symposium in Intelligent Technique, March 20-21, Bari, Italy, pp. 166-170.

Han, J. and Kamber, M. 2006. *Data Mining Concepts and Techniques (2nd ed.).* San Francisco: Morgan Kaufmann publications.

Hand, David J., Heikki Mannila, and Padhraic Smyth. 2001. *Principles of Data Mining (Adaptive Computation and Machine Learning)*, The MIT Press.

Inmon, W.H. 2005. *Building the Data Warehouse (4th ed.)*. Indianapolis: Wiley Publishing Inc.

Kimball, Ralph and Margy Ross. 2002. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling,* 2nd Edition, John Wiley & Sons.

Kohonen, T. 2001. *Self-Organizing Maps*, Berlin-Heidelberg.

Kotsiantis, S.B. 2007. *Supervised Machine Learning: A Review of Classification Techniques*. Informatica 31. pp. 249-268.

Marakas, George M. 2003. *Modern Data Warehousing, Mining, and Visualization: Core Concepts*, Prentice Hall.

Mingers, J. 1989. *An empirical comparison of pruning methods for decision-tree induction.* Machine Learning, 4(2), pp. 227-243.

Office of Higher Education Commission. 2006. *Research Report: Standard of Higher Education.* Parbpim Press. Bangkok. Thailand

Open Intelligence. 2005. *Open Source Web Application for OLAP Reporting*. Retrieved October 7, 2007, From http://openi.sourceforge.net/openi_product.html

Ponniah, P. 2001. *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*, John Wiley & Sons.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, California.

Raub, G.B; and Chen, W.W. 2005. A Cluster Analysis Approach to Describing Tax Data, *In proceedings of American Statistical Association Conference (ASA) 2005*.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*, Addison-Wesley.

Wilkinson, G. G., Fierens, F., and Kanellopoulos, I. 1995. *Integration of neural and statistical approaches in spatial data classification.* Geographical Systems, 2, pp. 1-20.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2nd Edition, Morgan Kaufmann.

Biography

Name	Asst. Prof. Dr. Waraporn Jirapanthong
Education Background	• PhD. in Computer Science, Software Engineering Group, City University, London, UK. • MSc. in Computer Science (Best Science Student with the Highest GPA Award from Professor Taeb Nilanithi Foundation, Thailand, 2001), Faculty of Science, Mahidol University, Thailand. • BSc. in Computer Science (First Class Honours), Faculty of Science, Thammasat University, Thailand.
Employment	Lecturer, Faculty of Information Technology, Dhurakij Pundit University