DATA MINING FOR ROBUST TESTS OF SPREAD

by

TEH SIN YIN

Thesis submitted in fulfillment of the
requirements for the degree
of Master Science

November 2008

In loving memory of my mother

Law Siew Gaik

ACKNOWLEDGEMENTS

In making this thesis a success, numerous people have helped me considerably by contributing their ideas, advice, support and assistance. Their commitment and persistence have given me added strength to complete this daunting task. I would like to express my utmost appreciation to the many sources of support behind this effort.

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, Associate Professor Dr. Abdul Rahman bin Othman for his invaluable guidance, relentless contribution of ideas, suggestions and information, constant advice and persistent encouragements. His professional expertise and experiences have helped me tremendously by opening the doors for me to go through the many statistical theories thoroughly. He is also my source of inspiration and determination that surge me forward in my task.

My gratitude also goes to the Institute of Graduate Studies (IPS) and School of Distance Education (PPPJJ) for offering me a fellowship. The stipend that I received has eased my financial constraints.

I would like to register my sincere appreciation to Professor H. J. Keselman for making available his data sets. Special thank and appreciation goes to Mr. Shahrier Pawanchik and Mr. Foo Kok Keong, for their kind assistance in my writing, despite the valuable amount of time this work has taken from them. My gratitude also goes to Professor M. Ataharul Islam and Associate Professor Dr. Michael Khoo Boon Chong for allowing me to attend their statistical classes. Their professionalism and knowledge have provided strong foundation on statistical theories which are useful for my research. I would also like to extend my thanks to Associate Professor Dr. Low Heng Chin, my ex-supervisor, for giving me unwavering encouragement and moral support.

Lastly, I am deeply grateful to my loving family members and dearest relatives for their patience, enthusiasm, effort, understanding, sacrifice, constant prayers and endless encouragement at all times during my study. I would like to convey my heartfelt gratitude to my father – Teh Kuang Peng, brother – Teh Cheah Foo, and two sisters – Teh Ming Ching and Teh Sin Ching, for being the pillars of my strength and inspiration. Their continued moral support are a boost to my morale and continuously drive and motivate me to do my very best.

Foremost, I wish to express deepest appreciation to SAS Institute Sdn Bhd, Kuala Lumpur for giving me an excellent training on SAS software. I am indebted to Miss Cheong Cheng Mui and the staffs of SAS for the kindness and hospitality.

In closing, my grateful recognition are due to all friends who lend me their friendship, lecturers who giving me unwavering encouragement and tremendous support, and course mates who endeavor together with me. Once again, my heartfelt appreciation goes to all who directly or indirectly involved and contributed in bringing this study into completion. Words are inadequate to express my gratitude and I thank you all with all my heart.

And may the peace and God's blessings be upon all of you

Teh Sin Yin

TABLE OF CONTENTS

CHAPTER THREE : METHODOLOGY

CHAPTER FOUR : RESULTS AND DISSCUSSIONS

CHAPTER FIVE : CONCLUSION

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AID	Automatic Interaction Detector
ANN	Artificial Neural Networks
ANOVA	Analysis of Variances
CART	Classification And Regression Tree
CHAID	Chi-Square Automatic Interaction Detector
CCR	Correct Classification Rate
DDM	Distributed Data Mining
DISTR	Type of distribution
EDA	Exploratory Data Analysis
GIGO	Garbage In, Garbage Out
GSCOND	Unbalanced group size increments
GSIZE	Total group size
GTE	General Telecommunications and Equipment
L-O-O	Leave-One-Out
MLR	Multiple Linear Regressions
MOMs	Modified One-step M-estimators
NN	Neural Network
OLS	Ordinary Least Squares
SAS	Statistical Analysis Software
SHAPE	Skewness of distribution
TAIL	Kurtosis of distribution
TEST	Test statistics
PROCED	Test of spread procedure
PVAL05	Resultant 0.05 *p*-values

PVAL10	Resultant 0.10 *p*-values
XTRANSF	Transformation on X_{ij}
XTRHINGE	Hinge estimator used on X_{ij}
XTRPARM	Transformation parameters used on X_{ij}
XTRTOTTR	Total trimming proportion on X_{ij}
ZRHINGE	Hinge estimator used on Z_{ij}/R_{ij}
ZRTOTTR	Total trimming proportion on Z_{ij}/R_{ij}
ZRTRIM	Type of trimming on Z_{ij}/R_{ij}

LIST OF PUBLICATIONS & SEMINARS

1 **Teh, S. Y.**, & Othman, A. R. (in press). *Validation of training model for robust tests of spread*. The proceeding of the 3rd International Conference on Mathematics and Statistics, ICoMS, Institut Pertanian Bogor, Indonesia.

2 **Teh, S. Y.**, & Othman, A. R. (2008). *Ensemble method hit ratio for robust tests of spread*. The proceeding of the 22nd Annual SAS Malaysia Forum, KL Convention Center, Kuala Lumpur, 33-37.

3 **Teh, S. Y.**, & Othman, A. R. (in press). *Using the DATA STEP and PROC SORT to split data into training and validation data sets.* The proceedings in the 3rd international conference on Mathematical Sciences – ICM. Al-Ain, United Arab Emirates: University of United Arab Emirates, *Vol. 1*, 133-144.

4 **Teh, S. Y.**, & Othman, A. R. (in press). Using the DATA STEP and PROC SORT to split data into training and validation data sets. *Journal of the Faculty of Science*.

5 **Teh, S. Y.**, Othman, A. R., Keselman, H. J., & Wilcox, R. R. (2007). The biweight midvariance test of spread, In A. I. Md. Ismail, Y. Abu Hassan, A. Mustafa, Z. Zainuddin, H. Kamarul Haili, N. Awang, & M. T. Ismail (Eds.). *Proceedings of the 3rd IMT-GT Regional Conference on Mathematics, Statistics and Applications* (pp. 987-993). Penang, Malaysia: Universiti Sains Malaysia.

6 Othman, A. R., Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., & **Teh, S. Y.** (2007). *Robust Levene test for variance equality*. Bulletin of the International Statistical Institute 56th Session: Proceedings [CD-ROM]. Lisboa, Portugal: Tziranda.

7 **Teh, S. Y.**, & Othman, A. R. (2008, August). *Validation of training model for robust tests of spread*. Poster presented in the 3rd International Conference on Mathematics and Statistics, ICoMS, Institut Pertanian Bogor, Indonesia.

8 **Teh, S. Y.**, & Othman, A. R. (2008, July). *Ensemble method hit ratio for robust tests of spread*. Paper presented in the 22nd Annual SAS Malaysia Forum, KL Convention Center, Kuala Lumpur, Malaysia.

9 **Teh, S. Y.**, & Othman, A. R. (2008, March). *Using the DATA STEP and PROC SORT to split data into training and validation data sets*. Paper presented in the 3rd international conference on Mathematical Sciences – ICM, Rotana Hotel, Al-Ain, United Arab Emirates.

10 **Teh, S. Y.**, Othman, A. R., Keselman, H. J., & Wilcox, R. R. (2007, December). *The biweight midvariance test of spread*. Paper presented in the 3rd IMT_GT regional conference on mathematics, statistics and applications, Gurney Hotel, Penang, Malaysia.

11 Othman, A. R., Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., & **Teh, S. Y.** (2007, July). *Robust Levene test for variance equality*. Paper presented in the International Statistical Institute 56th Session, Lisboa, Portugal.

PERLOMBONGAN DATA DARIPADA PENGKALAN DATA UJIAN TEGUH KEHOMOGENAN VARIANS

ABSTRAK

Data pelbagai dimensi (data simulasi) dalam kuantiti yang besar daripada halaman output SAS bagi enam ratus tiga puluh empat ujian teguh kehomogenan varians tersedia dihasilkan oleh Keselman, Wilcox, Algina, Othman, dan Fradette (dalam pencetakan). Prosedur teguh yang dipergunakan oleh Keselman, et al. (dalam cetakan) bergantung kepada penentuan strategi pemangkasan (trimming) simetrik atau tidak simetrik secara awal atau empirikal. Skor transformasi jenis Levene dan jenis O'Brien digunakan bersama dengan ujian ANOVA *F*, ujian teguh Lee dan Fung (1985) atau ujian Welch. Nilai-*p* untuk ujian-ujian ini dikumpulkan. Suatu ujian dikatakan teguh jika ianya tidak dipengaruhi secara serius oleh kegagalan andaian-andaian. Ujian berkenaan dengan statistik teguh juga adalah ujian yang dapat bertahan di dalam semua keadaan taburan data. Sesuatu ujian ditakrifkan teguh jika nilai-*p*nya bersamaan atau berdekatan dengan 0.05. Dengan adanya data terkumpul tentang nilai-*p* yang dihasilkan di dalam pelbagai situasi yang berbeza, kita berpeluang untuk melombong ciri-ciri prosedur mahupun keadaan simulasi yang menjanjikan nilai-*p* yang berpatutan. Tiga jenis kaedah yang digunakan iaitu regresi logistik, analisis pembezalayan dan satu kaedah penggabungan yang mencantumkan kedua-dua kaedah tersebut. Kita mengelaskan analisis kepada dua kategori iaitu 'keadaan simulasi' dan 'ciri-ciri prosedur'. Keadaan simulasi dikaji dengan tujuh taburan yang berlainan bersama enam reka bentuk. Ciri-ciri prosedur mengandungi maklumat tentang lokasi pusat, jenis pemangkasan, tranformasi, dan ujian statistik. Untuk setiap analisis, data telah dibahagi kepada dua bahagian, iaitu 95% untuk data latihan dan 5% untuk data kesahan. Dalam analisis yang pertama, kita memperolehi kesimpulan yang mengikuti norma statistik, yakni ralat Jenis I teguh diperolehi daripada prosedur yang dijalankan ke atas taburan normal dan jumlah saiz sampel yang besar. Walau bagaimanapun, keputusan dalam regresi logistik menunjukkan bahawa prosedur dengan taburan simetrik platikurtosis atau simetrik leptokurtosis, dan sampel saiz yang berbeza secara sederhana, masih boleh menghasilkan ralat Jenis I yang teguh. Dalam analisis yang kedua, kedua-

dua regresi logistik dan analisis pembezalayan menunjukkan bahawa pemangkasan tidak perlu dilaksanakan terhadap Z_{ij} /R_{ij} untuk memperoleh ralat Jenis I yang teguh. Dalam regresi logistik, ralat Jenis I yang merangkumi [0.045, 0.050] telah diperhati untuk prosedur yang menggunakan min kumpulan dalam transformasi ke atas X_{ij} ; tiada pemangkasan diperlukan ke atas X_{ij} dan tidak memerlukan penganggar 'hinge' untuk X_{ij}. Jika pemangkasan tidak simetrik dilaksanakan maka keputusan yang paling bagus diperolehi apabila 10% pemangkasan diperuntuk ke atas Z_{ij} /R_{ij} dengan H1 sebagai penganggar 'hinge' dan ujian *F* digunakan sebagai ujian statistik. Dalam analisis pembezalayan, kawalan ralat Jenis I yang baik dikesan oleh prosedur yang menggunakan transformasi O'Brien ke atas X_{ij} dan pemangkasan tak simetrik ke atas X_{ij}. Berkenaan dengan peratusan pengkelasan yang betul, walaupun kadar ketepatan model regresi logistik (lebih sedikit daripada 50%) adalah lebih besar daripada daripada analisis pembezalayan dan kaedah campuran, prestasi ramalannya masih lagi rendah.

Kata Kunci: Perlombongan data, ujian kehomogenan varians, ujian statistik teguh, regresi logistik, analisis pembezalayan, penganggar 'hinge'.

DATA MINING FOR ROBUST TEST OF SPREAD

ABSTRACT

Large quantity of multidimensional data (simulation data sets) from SAS output listings of six hundred and thirty four robust tests of spread procedures conducted by Keselman, Wilcox, Algina, Othman, and Fradette (in press) was available. The robust procedures that Keselman, et al. (in press) utilized were either based on prior or empirically determined symmetric or asymmetric trimming strategies. The Levene-type and O'Brien-type transformed scores were used with either the ANOVA *F*-test, a robust test due to Lee and Fung (1985), or the Welch (1951) test. *P*-values from these tests were than collected. A test is robust if it is not seriously disturbed by the violation of underlying assumptions. Robust statistical tests are tests that operate well across a wide variety of distributions. A test can be also considered robust if it provides *p*-values 'close' to the target (usually 0.05) in the presence of (slight) departures from its assumptions. In order to make sense of importance of the features of the procedures on the *p*-values generated, we collated these quantities of data from the output listings into a large SAS data set of 26,628 records and conduct data mining on it. Data mining of simulation conditions and the characteristics of spread procedures that correspond to the target *p*-value of 0.05 or a set of value that is 'close' to 0.05 were then carried out. Three data mining methods were used. They are logistic regression, discriminant analysis, and a composite method combining the two methods. We did separate analyses for the 'simulation conditions' and 'characteristics of the procedures'. The simulation conditions evaluated were seven different distributions by six designs. The characteristics of the procedures contained information of central locations, type of trimming, transformation, and test statistics. For each analysis, data was partitioned using 95% for training and 5% for validation. In the first analysis, our findings agreed with the norm in statistics that robust Type I error rates were obtained from procedures that were run on the standard normal distribution, and with large total sample size. However, findings in association with logistic regression indicated that procedure with symmetric platykurtic distributions or symmetric leptokurtic distributions, and moderately

unequal sample size, can still perform well in terms of Type I error rates. In the second analysis, both logistic regression and discriminant analysis revealed that no trimming was needed on Z_{ij} /R_{ij} in order to obtain robust Type I error rates. In logistic regression, Type I error rates falling in [0.045, 0.050] was observed for the procedures that used group means in the transformation of X_{ij} ; no trimming applied on X_{ij} and no hinge estimator used on $X_{ij.}$ If asymmetric trimming was carried out then the best results were observed when 10% trimming when applied on Z_{ij} /R_{ij} with H1 as the hinge estimator and used with usual F test as the test statistic. In discriminant analysis, robust Type I error rates was observed for the procedures that included O'Brien transformation on X_{ij} , and asymmetric trimming applied on X_{ij}. In the overall percentages of correct classification, although the logistic regression model accuracy rate (only slightly more than 50%) was higher than discriminant analysis and the ensemble method, its predictive performance is still very low.

Keywords: Data mining, tests of spread, robust statistical tests, logistic regression, discriminant analysis, hinge estimator.

CHAPTER 1
INTRODUCTION

1.0 Introduction to Data Mining

Data mining is a field that has emerged in response to analysis of large data sets. A wide range of techniques is available to analyst depending on the objective of the analysis at hand. Data mining involves various statistical methods for searching relationships among variables. The methods typically involve analyzing very large (massive) quantities of multi-dimensional data.

Frawley, Piatetsky-Shapiro and Matheus (1991), Fayyad, Gregory, Smyth and Uthurusamy (1996), Cabena, Hadjinian, Stadler, Verhees and Zanasi (1998), Westphal and Blaxton (1998), Thuraisingham (1999), Han and Kamber (2000), Hand, Mannila and Smyth (2001), Kantardzic (2003), and Bozdogan (2004), have defined data mining using slightly different words. Although there is no single definition of data mining that would meet with universal approval, this study summarized all with a sufficiently broad definition of data mining as below:

Data Mining - The process when a large volume of data can be reduced to a much smaller summary form without loss of information.

During the process, various models were constructed, summarized and later derived values from a given collection of data. These can enormously aid the subsequent analysis task. It becomes much easier to do graphical and other checks that give the analyst assurance that predictive models or other analysis outcomes are meaningful and valid. Relevant graphical summaries are perhaps the most important tool at the data analyst's disposal because hidden patterns can be made visible after data have been reduced to manageable size (Bozdogan, 2004).

Statistical data mining is the process of selecting and exploring large amount of complex information using modern statistical techniques and new generation computer algorithms to discover hidden patterns in the data (Bozdogan, 2004). Hence, relative impact and contribution of individual variables on a response need to be accurately measure. This is because the decision-makers will want to understand the basic relationships of the variables as well as prediction output (Christopher, Linda & Linda, 2004). Although there are other data mining techniques that fit these criteria, in this case, logistic regression and discriminant analysis seem to meet the research objectives exceptionally well. However, it is important to understand how regression and discriminant analysis fit in the data mining world.

There is an important philosophical difference between data mining and statistical analysis. For statistical analysis, the ideas (hypotheses) come first, whereas for data mining, the data comes first. Statistical analysis normally comes with a question i.e. “Given this explanation, does the data support it? If so, how confident can I be that it is supported? ”. On the other hand, data mining is a hypothesis-free approach. Statisticians typically have to manually develop the model equations that match the hypotheses. In contrast, the model equations in data mining did not need to be stated in advance.

These can be seen in regression analysis. There were significant differences between regression analysis for statistics and regression analysis for data mining. The statisticians developed regression models and examined the “beta weights” for linear and nonlinear regressions. They also analyzed residuals, constructed confidence intervals, estimated parameter of the models and translated the analysis outcomes back into English to explain what was happening in the data. In contrast, the main search criterion for the data miners was to find the “drivers”, i.e. find the particularly profitable groups and subgroups in the data (inferential modeling). Data miners then had to design models that would predict who are those people that form the profitable groups (predictive modeling).

Furthermore, statistics is also concerned with the process of data collection. However, areas such as design of experiments and design of surveys are not the domain of data mining as data mining is mainly concerned with knowledge discovery in database (Press, 2004). Other than that, the type of data input also helps to explain the difference between data mining and statistics. Statistical techniques usually handle numeric data and one need to make strong assumptions about their distributions; data mining algorithms typically can process a much wider set of data types and one can make fewer assumptions or no assumptions at all about their distributions.

However, researchers believe that there is no single factor which makes data mining an outright winner over the classical statistical approaches. Indeed, statistics play an important role in most data mining environments. The optimal strategy is always to use statistics and data mining as complementary approaches (Cabena, et al., 1998).

In addition, the classical data mining techniques such as Classification And Regression Tree (CART) and Artificial Neural Networks (ANN) are techniques robust to real world data and are also easy to use by many non data miners. By the same token, the technological advances in computer hardware have dramatically raised the ante by several orders, e.g. advances in storage capacity and data processing make some of the most potent data mining techniques practicable today (Berson, Smith, & Thearling, 1999).

The general nature of real world data by itself provides several analytical challenges. Berry and Linoff (2000) provided a good description of the activities involved in data mining. They stated that the major activities include classification, estimation, prediction, affinity grouping or association rules, clustering and visualization. Essentially, these exploratory approaches were intents of finding characteristics of the data sets and the undiscovered patterns in the data sets.

1.1 Advantages of Data Mining

Data mining tools can answer business questions that traditionally are time consuming to resolve. In general, wherever data exist, powerful data mining techniques can help reveal important data patterns that would otherwise remain unnoticed when using simple type of analysis. Data mining techniques scour databases for hidden patterns, finding predictive information that experts may miss. Thus, data mining are applicable in many areas (refer to Figure 1.1) (Thuraisingham, 1999).

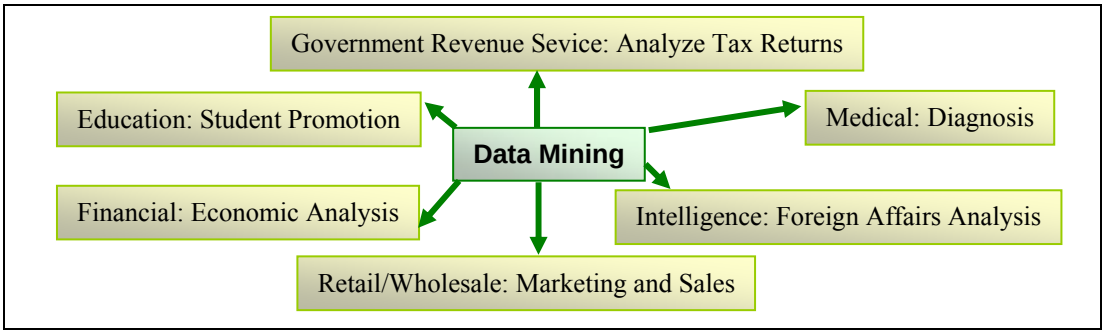

Figure 1.1. Some data mining application areas.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, it can analyze massive databases in minutes. Faster processing means that more models can be automatically experiment to understand complex data. Therefore, high speed makes it practical to analyze huge quantities of data. Larger databases, in turn, improved predictions.

In addition, via data mining, it is possible to discover patterns and build models automatically. The models are both descriptive and prospective. They explain why things happened and predict the future. "What-if" questions can be posted to a data-mining model that cannot be queried directly from the database.

On the other hand, visualization of the data mining output in a meaningful way allows analysts to see the data mining results. Analysts need to view the output of the data mining in a context that they understand. The output can be profiles in the term of graphs or chart. Visualization will enable analysts to see plausible relationship between variables that were tested (Thearling, 1997). By incorporating visualization graphics into the analysis process, analysts will be able to let the results generated by them put into good use. For example, a decision tree will enable analysts to see the interaction between the parent node and child node.

1.2 Data Mining Procedures

The steps in data mining are important keys to a successful data mining of a data set. Data mining projects require substantial initial effort in data preparation. Seventy five percent of total project time goes in data preparation (Michael & Gordon, 2004). In particular, the knowledge discovery process in databases consists of several steps. The overall statistical process, from data sources to model application involved the following data mining process (refer to Figure 1.2) (Han & Kamber, 2000):

- Data Cleaning and Data Quality Assessment
- Data Integration and Consolidation
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

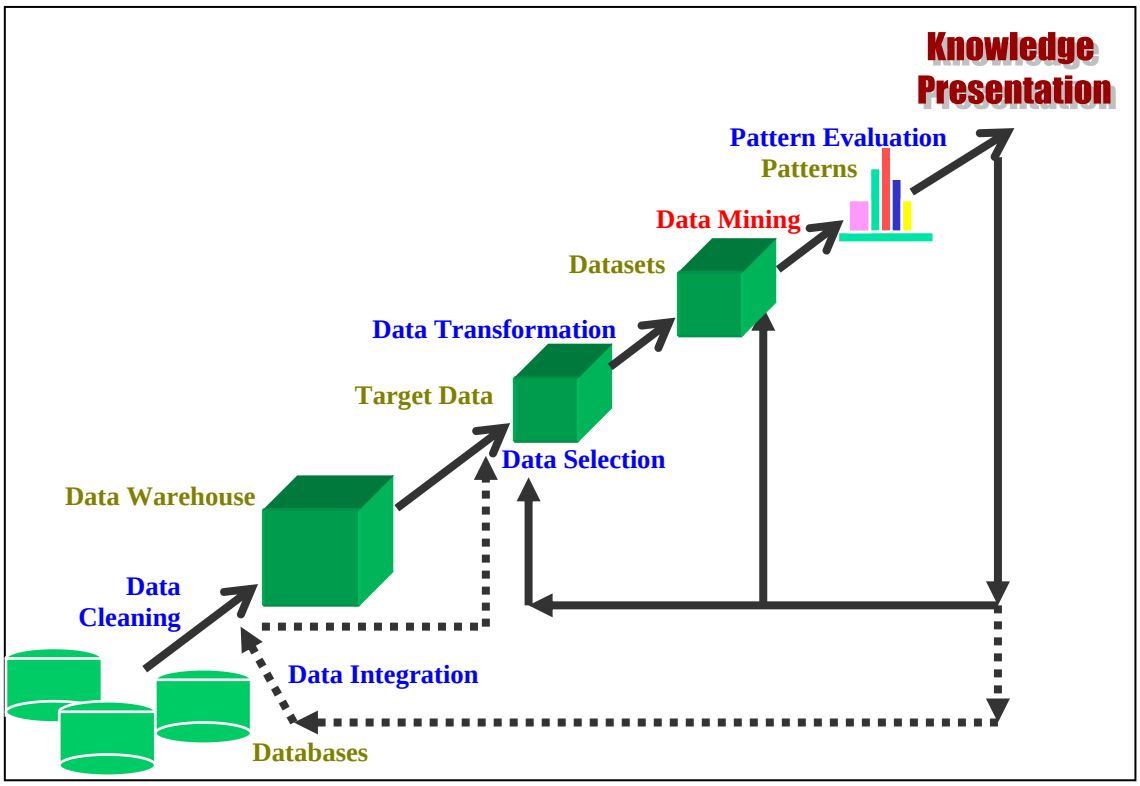

Figure 1.2. An overview of data mining procedures.

1.2.1 Data Cleaning and Data Quality Assessment

Garbage In, Garbage Out (GIGO) is applicable to data mining. In order to get good models, good data was needed. Data cleaning handles noise, errors, missing, and irrelevant data. This preprocessing stage removed not all but certain information that was considered unnecessary and also which are likely to slow down queries. Meanwhile, the data were checked to ensure consistency of formats in the database. However, inconsistent data format is always a possibility because the data were drawn from several different sources.

The aim of data cleaning is to find out what forms of data summary are likely to be useful, while losing minimum information from the data. It may often be reasonable to base analysis on one or more random samples of the data. Where data cleaning is a large chore, there may be a trade-off between time spent on data cleaning and time spent on analysis. It may then make sense to limit cleaning to a random sample of the data, allowing more time for analysis.

1.2.2 Data Integration and Consolidation

After removing noise from the observation in data cleaning and data quality assessment stage, data integration and consolidation were required. Collecting the target variables and attendant information from many different data sets is an essential step because all the data sets were stored in independent files. The uniformity in variable coding and the scale of measurements should be verified before combining the different variables and observations from different data sets. Then, all the data sets were merged and validated to ensure everything is in order.

1.2.3 Data Selection

In the data selection step, this research aims to extract information needed by the data mining step from the database. Hence, data which are relevant to the analysis task are retrieved based on the objectives of this study. Base on one's data selection criteria, one may wish to include or exclude other sets of data. Data selection is important because the existence of a lot of data for the data mining process. If inappropriate data were retrieved, the process could be time consuming and more money will be spend and also increase risk. Data selection is different from sampling the database and choosing predictor variables. It is an elimination of irrelevant or unneeded data during data analysis to achieve the research objectives.

1.2.4 Data Transformation

Data transformation is the application of a mathematical modification to the values of a variable. There are three purposes for data transformation most commonly discussed in data mining and statistics:

1) mathematical functions,
2) statistical ranking, and
3) missing value imputation.

Data transformation for functional purpose is to better satisfy the fundamental assumptions of statistical analysis underlying the multivariate techniques:

a) to stabilize the variance of the dependent variable, if the homoscedasticity assumption is violated;

b) to normalize the dependent variable, if the normality assumption is noticeably violated; and

c) to linearize the regression model, if the original data suggest a model that is nonlinear in either the regression coefficients or the original variables (dependent or independent).

Fortunately, if the original data does not satisfy (a)-(c), one can find the same transformation that helps to accomplish the first two and sometimes even the third assumptions. Patterns of the variables will suggest specific transformations (Hair, Anderson, Tatham & Black, 1998; Kleinbaum, Kupper, Muller & Nizam, 1998).

The second purpose of data transformation is statistical ranking. Data transformation for statistical ranking purpose allows the researchers to convert the continuous or interval variable into rank or ordinal variable that are better match for the model. This can improve the relationship (correlation) between variables. Thirdly, data transformation was done when there were missing values present in large proportions of some important variables in the data sets. In this case, data transformation is actually imputation of the missing values (Hair, et al., 1998). Imputation meant that missing values were estimated by existing values.

1.2.5 Data Mining

Having very large databases is becoming standard practice. Hence, it is impossible for researchers to mine the data manually to search for interesting patterns. Data mining is an intelligent method applied to extract data patterns, i.e. applying a concrete algorithm to find useful and novel patterns in the data. Generally, data mining is the stage of finding universal patterns or principles that summarize and explain a set of observations.

1.2.6 Pattern Evaluation

The splitting of the large preprocessed data set into training, validation and test subsets allows potential models to be assessed. Training data set was used to build models that were assessed on the validation (holdout) data set. If a model does not perform satisfactorily in the validation phase, it is restrained. One can partition the data by applying continuous iteration between training and validation data sets until the performance with validation data achieved satisfactory. This naïve strategy is based on the assumption that the training set and the validation set are chosen as representatives of the same, unknown distribution of data. Hence, validation determines whether the model is "built right" (Johnson & Wichern, 2002).

However, in common practices, researchers specified the percentage of the data allocated to training and validation without iteration (Efron & Gong, 1983; Picard & Berk, 1990; SAS Institute Inc., 2004). Specifically, when they used programming software to partition data, default percentage will be used to generate the partition (i.e. for SAS EMiner default split percentage is 40% training, 30% validation, and 30% test data sets). Moreover, different methods for splitting or sampling the data would have different percentage of split. For example, Efron and Gong (1983) suggested a split between 25%-50% of the data as an optimal partition for validation when cross-validation method was used; Azuaje

(2003) suggested 95%-5% partition for training-validation when leave-one-out (L-O-O) method was used.

Next, a trained and validated model was accessed using the test data. To ensure an unbiased assessment of model performance, the test data set is ordinarily used only once at the end of the modeling process. In practice, the test data step was omitted and instead the validation sample or cross-validation sample was accepted as the final assessment (Hand, et al., 2001; Larose, 2006). This research compared outcomes of the model to suit this study and if the model meets the objectives then the model will be used. Once established, its performance was monitored.

1.2.7 Knowledge Presentation

Data mining is the process that transforms immense data into certain knowledge presentation, in which researchers are able to find meanings. One of the important roles of data mining is that knowledge presentation has to be developed according to the research objectives. The knowledge can be used directly, incorporated into another system for further action, or simply documented and provided the report to interested parties.

The knowledge that is discussed earlier can be a rule in decision tree, regression tree, and profiles in the form of graphs or charts (Elder & Abbott, 1998). For example, in decision trees, knowledge presentation is in the term of a rule that classifies a large amount of data. The results of judgment regarding conditions are displayed as branches and the whole data is rendered as a tree. Another example is given by Fukuda, Morimoto, Morishita and Tokuyama (1996). They expressed the distribution of data by colors. Abundant data was successfully visualized. Such knowledge presentations enable analysts to describe knowledge freely (Anzai, 1989), and also serve as significant role in the discovery of knowledge.

1.3 Introduction to Robust Tests of Spread

Large quantities of multidimensional data (simulation data sets) from Statistical Analysis Software (SAS) output listings of 634 robust tests of spread procedures conducted by Keselman, Wilcox, Algina, Othman, and Fradette (in press) are available. The robust procedures that Keselman, et al. (in press) utilized were either based on prior or empirically determined symmetric or asymmetric trimming strategies. The Levene-type and O'Brien-type transformed scores were used with either the Analysis of Variance (ANOVA) *F*-test, robust test carried out by Lee and Fung (1985), or the Welch test (Welch, 1951). The procedures created will be presented in Chapter 3.

The two main purposes of this study were to create a database of all *p*-values for tests of spread procedures from Keselman, et al. (in press) and to conduct data mining of simulation conditions and characteristics of the spread procedures that correspond to the target *p*-value of 0.05 or a set of value that is 'close' to 0.05. A test is robust if it is not seriously disturbed by the violation of underlying assumptions. It provides *p*-values 'close' to the true ones in the presence of (slight) departures from its assumptions. Robust statistical tests are tests that operate well across a wide variety of distributions.

This study evaluated the specific conditions (run over seven different distributions by six one-way independent group designs) under which a test will give satisfactory result. It means that the empirical Type I error rate of these tests should be 'close' to $\alpha = 0.05$. This evaluation was an investigation on the impact of these conditions on tests of homogeneity of variances that resulted in *p*-value close to $\alpha = 0.05$.

The simulation conditions (four variables) that were evaluated in the J=3 study, where J = number of groups, were:

(a) **total sample size** - total sample sizes are manipulated by N=60 and N=120,

(b) **degree of sample size inequality** - there were three conditions of sample size equality or inequality investigated**,**

(c) **shape of the population distribution** - seven distributions were used to compare the procedures, and

(d) **percentages of trimming** – two different sets of four values of total trimming, i.e. symmetric trimming (10%, 20%, 30% and 40%), and asymmetric trimming (10%, 15%, 20% and 25%) were evaluated.

For each combination of these conditions, five thousand data sets were simulated and run on the 634 procedures mentioned earlier. The p-values of these procedures applied on these data sets were collected. Further discussion on how these conditions were chosen is in Chapter 3.

1.4 Problem Statements

Base on Section 1.3, the problem statements for this study are

i) What are the important simulation conditions that will produce p-value close to or equal to 0.05?

ii) What are the important characteristics of the spread procedures?

1.5 Objective of the Thesis

Two primary goals of this research were:

i) Create a database of all *p*-values for tests of spread procedures from Keselman, et al. (in press) and

ii) Determine important simulation conditions and characteristics of the spread procedures that correspond to the target *p*-value of 0.05 or a set of value that is 'close' to 0.05 using the predictive modeling approach.

1.6 Significance or Contribution of Study

This study enables one to examine the simulation conditions and characteristics of procedures that were able to produce robust Type I error rates procedures simultaneously. In addition to this an ensemble method that combined logistic regression with discriminant analysis was developed. There have been some works done on neural network ensembles (Hansen & Salamon, 1990) and ensemble methods in machine learning (Dietterich, 2000), but, there is no work done on combining logistic regression and discriminant analysis together in an ensemble. This proposed technique is useful and applicable for combined model predictions to form a potentially stronger solution in an ensemble system. Hence, this thesis explores the potential of an ensemble method with this new combination.

1.7 Organization of the Thesis

Briefly, data mining is a search for relationship between the variables using various computer-intensive methods. The methods typically involve analyzing massive quantities of multidimensional data. Data mining allows us to identify hidden pattern of relationships in large databases. Therefore, in order to explain the proposed methodology, this thesis is organized as follows:

Basically, Chapter 1 is a general introduction to data mining; advantages of data mining; procedures for data mining; and introduction to tests of spread. The primary objectives of this study are stated too.

In Chapter 2, literature of past and contemporary research on data mining will be reviewed. Important data mining and statistical terminologies will also be reviewed. Next, the supervised learning methods for data mining will be discussed. Then, the data mining techniques which were applied to determine the important simulation conditions and the important characteristics of the procedures in the generation of *p*-values will be illuminated. The techniques were logistic regression, discriminant analysis and a composite method that combined the two methods mentioned earlier. The criterion of robustness, a topic relevant to the tests of spread; the tests of spread themselves: namely, the Levene and O'Brien tests were discussed too. In the last section, detail information of the 633 statistics procedures will be stated.

In Chapter 3, the methodology for this study is presented. First, the conditions and designs on conducting data mining were studied. In this chapter, the transformation of the dependent variable, the information regarding categorical independent variables available for entry, and the application of dummy variables were presented. Next, randomization technique was performed to randomly split observations into training and validation data sets. Then, the aggregation of training and validation data sets by restructuring the data was discussed. The variables were separated according to 'simulation conditions' and 'characteristics of the procedures'. After splitting the observations into training and validation data sets, the performance of the training model was evaluated with the validation data set.

The findings are presented in Chapter 4. Here, the focus was on the analysis and interpretation of outputs. Visualization tools were utilized to help in the research interpretation. This study analyzed the data set by separating the variables into ‘simulation conditions’ and ‘characteristics of the procedures’. The discussion focused on logistic regression, discriminant analysis, and a composite method that combined the two methods mention earlier.

In Chapter 5, the findings were summarized with conclusions for ‘simulation conditions’ and ‘characteristics of the procedures’, respectively. Last but not least, the thesis is finalized with the lessons learnt from this study and recommendations for further research are provided.

CHAPTER 2
LITERATURE REVIEW

2.0 Introduction

This chapter reviewed literature of past and contemporary research on data mining. Important data mining and statistical terminologies will be reviewed. Next, the supervised learning methods for data mining were discussed. Then, the data mining techniques that were applied to determine the important of the simulation conditions and characteristics of the procedures in the generation of *p*-values were illuminated. The techniques were logistic regression, discriminant analysis and a composite method that combined the two methods mentioned earlier. The criterion of robustness; a topic relevant to the tests of spread; the tests of spread themselves: namely, the Levene and O'Brien tests were discussed too. In the last section, detail information of the 633 statistics procedures was given.

2.1 Genesis of Data Mining

The proliferation of information technology over the last decade has caused almost all organizations to automate their business practices. This has provided them with large quantities of data. Though available these were fragmented for analysis. The existence of these massive data is becoming prevalent in many industries and their applications (Christopher, et al., 2004). In response to meet the needs of industries for these large datasets to be analyzed, a field called data mining came into being. Data mining encompasses a wide range of techniques that are available to the analyst and the analysis is based on the objectives.

Naisbitt (1982, p.12) noted that "we are drowning in information but starved for knowledge". Data mining provides an effective means to analyze the uncontrolled and unorganized data and turn them into meaningful knowledge. Traditionally, organizations use data tactically to manage operations. To have a competitive edge, strong organizations need

to use data strategically to expand their businesses, improve profitability, reduce overhead costs, and market their products and services more effectively (Graettinger, 1999). Data mining creates information assets that an organization can leverage to achieve these strategic objectives.

From the philosophy of science perspective, data mining actually follows the traditional scientific method put forward by Galileo (1564-1642). In review, the scientific methodology proposed by Aristotle (b 384BC) and advocated by Bacon (1561-1626) involving collecting large quantities of data; searching for patterns and then making hypotheses about the patterns (refer to Figure 2.1). Galileo advocated continuing in the same style but suggested that scientists should also do experiments to validate the hypotheses. This Galileo scientific method was widely accepted within the scientific community for about three hundred years. It was still used commonly throughout the nineteenth century (Press, 2004).

Pre-Galileo Scientific Method

Collect data → Searching for patterns → Hypothesizing

Galileo Scientific Method

Collect data → Searching patterns → Hypothesizing → Experiments check hypotheses

Confirmatory Science Method

Postulated theory → Collect experimental data support the theory

Exploratory Science (Data Mining)

Collect data → Find patterns → Hypothesizing → Experiments check hypotheses

Figure 2.1. Data mining from science perspective.

However, in the twentieth century, an important shift in the way the scientific method was practiced took place. The Galileo scientific method was retained, and it was advocated that the theory should be postulated first. The experimental data should be collected so as to support the theory (refer to Figure 2.1). This has been called confirmatory

science. For example, an analyst in a sales department create a hypothesis that there will be sales increase for certain type of products around Father's Day or a price discount, is the primary effect for sales increase. Data mining today is akin to the nineteenth century scientific method, in that the data generate the theory. This data mining approach which primarily looks at some patterns of the data is called exploratory science (Press, 2004).

Early work on data mining were focused on issues such as data collection, followed by access strategies, database systems, data model and then data analysis. Modern data mining have all of the early works with more comprehensive improvements. Developments in data mining were discussed in Grossman (1997) and Thearling (2006). The two researchers agreed in the stages of data mining development but they have discussed them in different ways in terms of the period of development. Grossman (1997) classified data mining into three generations, whereas Thearling (2006) classified them in terms of specific time frame.

According to Grossman (1997), the first generation develops single or a collection of data mining algorithms to mine vector-valued data. The second generation supports mining of larger data sets and data sets in higher dimensions. It also includes developing data mining scheme and data mining languages to integrate mining into database management systems. The third generation provides Distributed Data Mining (DDM) focused on database or data warehouse of information which is located in different places or in different physical locations.

Chronologically, Thearling (2006) explained that data mining started in the 1960s as a data collection, followed by data access in the 1980s. In the 1990s, the data mining evolved into data warehousing and decision support stages.

Essentially, Grossman (1997) and Thearling (2006) statements indirectly agreed with Thuraisingham's (1999) view that data mining emerged as a field in late 1980s and early 1990s but then research in machine learning, statistics and database management has actually gone on for a long time. According to Thuraisingham (1999), data mining as a field is partly due to "knowledge discovery in databases" workshop series that started in the late 1980s and then evolved into "knowledge discovery in databases" conference series.

It is clear that data mining techniques are a result of a long process of research and product development for many decades. Although data mining algorithms embrace techniques that have been in existence for at least 10 years, only recently these techniques are implemented as mature, reliable, understandable tools that consistently do better than the older statistical methods. The older data mining techniques are closest to traditional data analysis methods. In fact, it is fair to say that traditionally statistics have been used for many data mining analysis, such as building predictive models or discovering associations in databases.

At present, most commonly used data mining techniques are ANN and decision trees such as CART. These techniques are used to mine large databases to capture useful patterns and relationships. They are useful for either discovering new information within large databases or building predictive models. They represent the vast majority of the techniques that are often mentioned in the popular press. Though CART is gaining wider acceptance, the older decision tree technique such as **Ch**i-Square **A**utomatic **I**nteraction **D**etector (CHAID) is frequently used too (Sarma, Gupta & Vadhavkar, 2004). Further discussion for ANN, CART and CHAID were in Section 2.3.1, Section 2.3.2, and Section 2.3.3, respectively.

Data mining appears to have had its genesis in business, in particular, the marketing, finance, advertising, and other subfields. The patterns identified by the data mining solutions can be transformed into knowledge, which can then be used to supporting business decision making. For example, **G**eneral **T**elecommunications and **E**quipment (GTE) laboratory has worked on data mining focusing on producing specific reports (Berry & Linoff, 2000). Then the approach has spread widely throughout the social sciences. Lately, data mining is also applicable to medical cost mining (Kantardzic, 2005).

In fact, data mining is a multi-disciplinary subfields at the confluence of statistics, pattern recognition, computer science, machine learning, **A**rtificial **I**ntelligence (AI), and database technology, and perhaps a few other subfields too (refer to Figure 2.2). To date, each subfield has its own terminology and jargon. Example, unlike the statisticians, non-statistical workers refer dependent variables as features and independent variables as attributes. They actually refer to the same things.

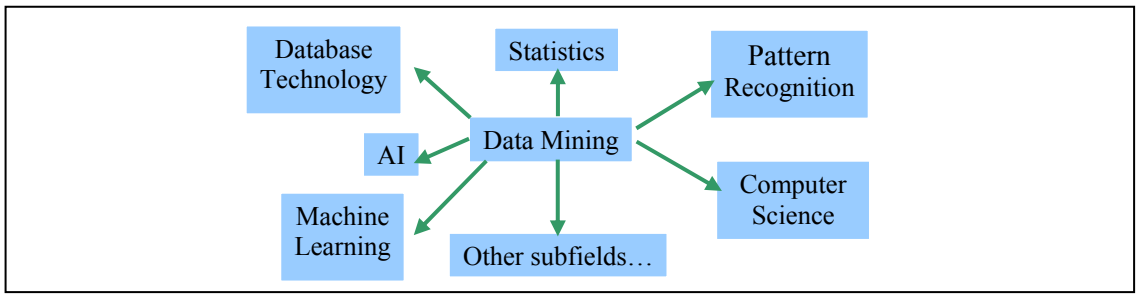

Figure 2.2. Data mining confluence of multi-disciplinary subfields.

Data mining methodology draws upon collection of data mining tools some of which are common in the various subfields. These tools include various types of traditional statistical methods of multivariate analysis. Some of the traditional statistical methods are classification, regression, clustering, contingency table analysis, principal components analysis, correspondence analysis, multi-dimensional scaling, factor analysis, and latent structure analysis. In addition, computer-based methods are AI, and machine learning (Press, 2004). Other then traditional statistical methods and computer-based methods, data mining

not developed from statistics are tree building and tree pruning, support vector machines, link analysis, genetic algorithms, market basket analysis and neural network analysis.

Through evolution, data mining tools from different subfields are supporting each other in data mining. Typically, clerks spend years recording the data and analysts go through them detecting various patterns. This was the beginning of "**statistics**". However, getting the data organized was a big problem. So, to solve this problem, a new field (**computer science**) was being involved. They started storing the data in computerized files and databases. This was the first big step toward data mining. Subsequently, **AI** was developed with new and improved searching and learning techniques. Then, **database technology** comes to improve the way of storing and retrieving the data.

Following this the evolution of support between statistical methods and computer science were looked into. Statistical methods have resulted in various statistical packages to compute sums, averages and distributions. These packages are now being integrated and stored in databases for mining. While, machine learning in computer science is all about learning rules and patterns from the data. Machine learners need some amount of statistics to carry out machine learning. According to Thuraisingham (1999), statistical methods and machine learning are the key components to data mining. An example of such support can be found in ANN. ANN now adopted a statistical method known as cross tabulation to produce "classification table".

From the literature review, it is clear that a great deal of time and effort has been spent on developing data mining and improving better data mining techniques in multi-disciplinary subfields. However, this hot cutting-edge research area of data mining will need the contributions from all subfields.

2.2 Terminologies

The terms that being used in data mining and statistical analysis in certain concepts, ideas, variables, processes and the rest, may or may not be the same. The discussion that follows will briefly highlight similarity and differences, and determine the terms that will be use through out the writing of thesis.

2.2.1 Classification

Classification has the same meaning in data mining and statistics. Kantardzic (2003) mentioned that classification is the first and most common task in inductive learning. This is a learning function that classifies a data item into one of several predefined classes (Hand, 1981; McLachlan, 1992; Weiss & Kulikowski, 1991). In other words, classification aims to indicate the common characteristics of the group to which each case belongs. This pattern can be used to understand the existing data and predict how new instances will behave.

A training data set can be seen in Figure 2.3. The samples belong to different classes and therefore different graphical symbols were used to visualize each class. And, the final result of classification is the curve (separator) shown in Figure 2.3, which best separates the samples into two different classes. By applying this function, every new sample can be classified into one of the two classes. Similarly, when the problem is specified with more than two classes, it results in more complex functions.

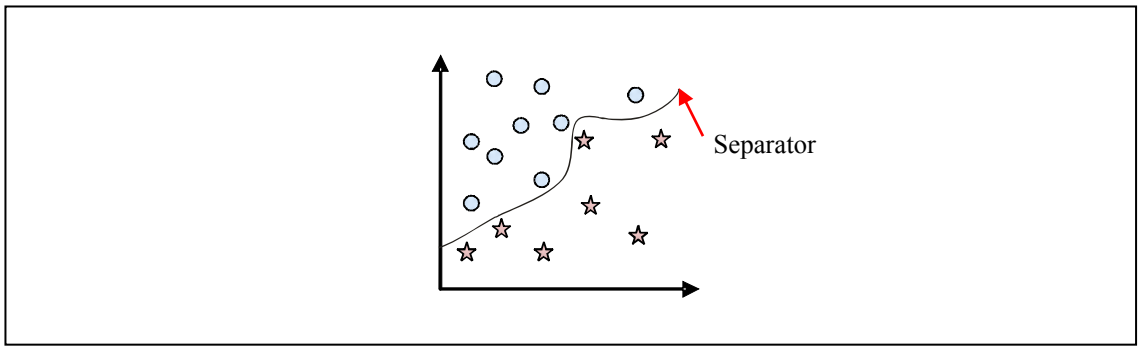

Figure 2.3. Graphical interpretation of classification.

2.2.2 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is a terminology used in data mining and statistics to explain an approach or philosophy for data analysis that employs graphical and descriptive statistical techniques to uncover underlying structure, maximize insight into a data set, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models and determine optimal factor settings. The objectives of EDA are to create hypotheses on the causes of observed phenomena; assess the assumptions on which statistical inference will be based; support the selection of appropriate statistical tools and techniques, and provide a basis for further data collection through surveys or experiments (Hoaglin, 1985; Pyle, 1999).

2.2.3 Predictive Models

The objective of a model which is built from the data mining process is to classify/predict new instances correctly. Hence, predictive models involve models predicting an output variables based on several predictors or independent variables (Cabena, et al., 1998; Hand, et al., 2001; Tan, Steinbach & Kumar, 2006). Common data mining predictive modeling techniques are regression, decision tree and neural network. Predictive model in statistics serve the same purpose too. For example, by collecting customers' demographics information of the past and present, one can predict the financial situation of any customer; whether or not the customer is likely to be a bankrupt in future.

2.2.4 Dependent Variables and Independent Variables

The variables were classified based on whether a variable is intended to describe or be described by other variables. Such a classification depends on the study objectives. If the variable under investigation is to be described in terms of other variables (Y variable), it is called a *dependent variable, a response, an explained variable, a predicted variable, a target variable* or a *regressand*. The variable (X_i) used in conjunction with other variables (X_1, ..., X_{i-1}, X_{i+1}, ..., X_p) to describe a given response variable (Y) are called *independent variables,*

predictors, explanatory variables, control variables or *regressors* (Kleinbaum, et al., 1998). These terms were used interchangeable in this thesis writing.

2.2.5 Regression

In statistics, regression analysis is helpful in ascertaining the probable relationship between variables, and the ultimate objective. It is to predict or estimate the value of one variable corresponding to a given value of another variable. In data mining, Kantardzic (2003) mentioned that regression is a second inductive learning task. This learning function in regression maps a data item into a real value prediction variable.

The ideas of regression were first elucidated by the English scientist Sir Francis Galton (1822-1911) in his research on heredity – first on sweet peas and later on human stature. He described the probability of children being either short or tall depended on the height of their parents. He first used the word 'reversion' to describe this phenomenon and later changed the terminology to 'regression' (Chatterjee, Hadi & Price, 2000).

Hence, regression analysis in simple terms is finding a mathematical function which best describes the relationship between a dependent variable and one or more independent variables. In a simple regression analysis, it is limited to situations in which the dependent variable is a continuous variable. For linear regression the relationship is constrained to a straight line and the least-squares analysis was used to determine the best fit line. However, in many circumstance, the dependent variable is dichotomous. The type of regression analysis that is employed where the dependent variable is dichotomous is logistic regression, and likelihood functions are used to find the best relationship (Draper & Smith, 1981; Menard, 2002). Similarly in data mining, the function was called regression function. Using the regression function, it is possible to estimate the value of a prediction variable for each new sample (Kantardzic, 2003). Look at the example given in Figure 2.4. The figure shows the regression function which was generated based on some predefined criteria.