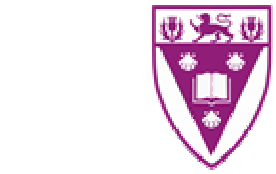
Data Mining with Oracle 10g using Clustering and Classification Algorithms

Nhamo Mdzingwa
Computer Science Department
Supervisor: John Ebden
7th November 2005



RHODES UNIVERSITY
Where leaders learn

Submitted in partial fulfilment of the requirements of the Bachelor of Science (Honours) degree of Rhodes University

Abstract

Deciding on which algorithm to use, in terms of which is the most effective and accurate algorithm in data mining, has always been a challenge for most data miners. This has been made even more complex by the increasing number of data mining tools that are available.

The objective of this research is fundamentally focused at investigating the effectiveness of two Clustering algorithms available in Oracle10g for data mining. These are the K-Means and the O-Cluster algorithms. I intend to provide data miners with adequate information regarding to Oracle data mining algorithm accuracy and effectiveness in building and applying models. Although, the evaluation of unsupervised data mining algorithms is a generally difficult task as the goals of an unsupervised data mining session are frequently not as clear as the goals of supervised learning, I have adopted evaluation techniques proposed by [Roiger et al, 2003]. The second objective is to gather information from the dataset used in the evaluation. Information gathering involves finding predictors of HIV AIDS prevention behaviour attributes. The data was obtained from the Centre for AIDS Development, Research and Evaluation Institute for Social and Economic Research, Rhodes University and is HIV AIDS related

The results obtained are as follows; the first set is concerned with the evaluation of the K-Means and O-Cluster algorithms. Here it was observed that the O-Cluster algorithm builds more accurate models than the K-Means algorithm and also that the models by the O-Cluster algorithm find more accurate clusters when applied to new data. From these results it is evident that the choice of modelling algorithm has significant difference in its efficiency and accuracy. The second set of results involves information gathering from the dataset. Here the attributes *HIV Test* and *Know AIDS* were identified as predictors of prevention behaviour of condom use and abstinence. These were found by distinguishing the clusters found in the dataset.

Acknowledgements

The success of this project can be attributed to the help of a number of people. Firstly, I would like to thank my project supervisor, John Ebden for guiding me through the development process, for his support, technical advice during the course of the project and offering much constructive criticism along the way.

I would also like to thank Kevin Kelly, from the Centre for AIDS Development, Research and Evaluation Institute for Social and Economic Research, Rhodes University, for providing me with the dataset that made this project possible as well as more interesting.

Many thanks go to the Andrew Mellon Scholarship Foundation for providing me with the funds to complete my Honours degree. I also acknowledge the financial and technical support of this project by Telkom SA, Business Connexion, Comverse SA, and Verso Technologies through the Telkom Centre of Excellence at Rhodes University.

Warm thanks also go to my family and friends who have provided support during the challenges, and understanding of my many hours spent at University.

Lastly, recognition should go to my fellow project students. The people I have spent 90% of my life with this year, your help and support has far outweighed the distractions.
Thanks guys!

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

Problem Statement:

The main objective of this project is to evaluate two algorithms available in Oracle 10g data mining with respect to algorithm accuracy and effectiveness. This research project pays attention to the evaluation of clustering algorithms, with the adoption and implementation of techniques proposed by [Roiger et al, 2003]. However, due to the nature of the field of data mining my second objective involves gathering information from the dataset used for the evaluation of the algorithms. This involves finding predictors of HIV AIDS prevention behaviour attributes. Describing and distinguishing the clusters found was necessary in order to do achieve this. Further more, other types of mining algorithms such as association rules (the Apriori) and classification algorithm (Adaptive Bayes Networks) were employed to facilitate further analysis on the data.

1 Background

1.1 Introduction to Data Mining

Data mining is a powerful and new technology that has been steered by the revolutionary progress in digital data acquisition and storage which has resulted in the creation of huge databases. In fact, the production and accumulation of these digital databases is occurring at a faster rate than our ability to comprehend and use them.

The idea behind Data mining is to find patterns of information in these databases such that an organisation such as a business can make use of. This new technology has been defined in almost as many ways as there are authors involved in the field. It has been defined by [Han et al, 2001] and [Mannila et al, 2001] as a process of extracting or mining knowledge from large amounts of data, or simply knowledge discovery in databases. However, because data mining sits as an intermediate between statistics,

computer science, artificial intelligence, machine learning, database management and data visualization, only to mention a few of the fields, the definition rapidly changes depending on the user's perspective.

The main parts of data mining are concerned with the analysis of data and the use of software techniques for finding patterns and regularities in datasets. Fundamentally it is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data. However, the data mining software used by the computer makes use of algorithms that facilitate sifting through the dataset in turn building models. The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of the task to be accomplished and the nature of the available data.

As a summary, data mining is a technology that is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection stored in databases
- Powerful multiprocessor computers
- Data mining algorithms

Techniques in data mining 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 [Mannila et al, 2001]. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions [Han et al, 2001].

1.2 Supervised and Unsupervised Learning

There is a wide range of sources available on data mining and most of these have various ways of implementing data mining processes. Most authors have found it more convenient to categorise data mining corresponding to different objectives to make it easier for the person analysing the data. Due to this the learning from data has been split into mainly two flavours: supervised learning and unsupervised learning.

According to [Roiger et al, 2003], data mining is classified into supervised and unsupervised concept learning methods. Supervised learning builds classification models by forming concept definitions from sets of data containing predefined classes while unsupervised clustering builds models from data without the aid of predefined classes were data instances are grouped together based on a similarity scheme defined by the clustering system.

The supervised learning builds models by using input attributes to predict output attribute values while in unsupervised learning no target attributes are produced but rather gives a descriptive relationship by using an objective function to extract clusters in the input data or particular features which are useful for describing the data. The number of variables in unsupervised learning is usually much higher than in supervised learning. The properties of interest are usually more involved than simple locations (e.g. means, medians) and dispersions (e.g. standard deviations and median absolute deviations) [Roiger et al, 2003]. Typical examples of unsupervised learning tasks are association rules, cluster analysis and principal component analysis, independent component analysis and multidimensional scaling.

[Berry et al, 2000] also categorises data mining into directed data mining and undirected data mining as the two main styles of data mining. According to the authors, the goal in directed data mining is to use the available data set to build a model that describes one

particular variable of interest in terms of the rest of the available data. The authors also point out that directed data mining often takes the form of predictive modelling, where one knows what he wants to predict. Classification, prediction and estimation are the techniques used in directed data mining. In undirected data mining, no variable is singled out as the target. The goal is to establish some relationship among all the variables.

From the above, it is clear that the classified categories described by these authors all involve similar techniques. We can therefore say that directed data mining, supervised learning and predictive modelling of data mining describe similar techniques that can be collectively referred to as supervised learning. Unsupervised learning, undirected data mining and descriptive modelling are techniques in the same category and can therefore be referred to as unsupervised learning.

However, the problem in data mining at the moment is that, although data mining is categorised, there are now too many data mining algorithms available. This makes it difficult for miners to know which of the categorised algorithms builds the most accurate model and whether the model built finds accurate patterns when applied to new data.

1.3 Chapter Summary

This chapter provides a brief introduction to the field of data mining. The field is steered by the ever increasing size of digital data and storage in databases. Data mining is classified into mainly 2 categories, supervised and unsupervised learning. However little knowledge on algorithm accuracy has lead to miner having difficulties when selecting algorithms to mine for data.

The structure of the rest of the paper is as follows: Chapter 2 highlights on research direction indicators elaborating on various work related to the evaluation of data mining algorithms relating to algorithm accuracy and effectiveness. Chapter 3 details the

methods adopted for the investigation detailing on the mining tool used, the algorithms that will be investigated, the dataset and techniques adapted. Chapter 4 is a brief breakdown of the implementation process of the methods suggested while chapter 5 gives an interpretation of results achieved. Chapter 6 details the gathering of information from the dataset by finding predictors of HIV AIDS prevention behaviour attributes. The document concludes with chapter 6 which gives a summary of all findings.

CHAPTER 2

2 Research Direction Indicators

This chapter provides a brief overview of research direction indicators. Basically, it explains why Oracle Data Miner (ODM) was chosen for use from a variety of data mining tools available in the market in order to perform the evaluation of the algorithms. It also explains why clustering algorithms were chosen for evaluation rather than evaluating other algorithms from other categories such as classification or even regression.

2.1 Choice of Mining Tool

There is currently a variety of data mining tools available in the market. Some of these tools are from big software house companies such as IBM, Microsoft, and Oracle. The various tools employ various mining techniques as well as algorithms depending on the functionalities available in these tools. Currently the most popular tools available in the market include the following (only to mention a few):

- *Oracle Data Miner (ODM)* from Oracle
- *Gnome Data Mine Tools* are provided as free open source software
- *Insightful Miner*
- *Clementine*
- *Data mind*
- *Enterprise miner*
- *Intelligent miner for data*

For this research project, Oracle Data Miner (ODM) version 10.1 from Oracle was selected for use to provide a means for the evaluation of clustering algorithms and also in gathering information from the datasets used by describing the clusters found.

The use of ODM mining tool was mainly motivated by research findings from the META Group, a leading provider of Information Technology research, advisory services and strategic consulting, which published its METAspectrum report for Data Mining. The report ranked Oracle Data Mining a leader amongst other data mining tools [Berger, 2004]. The data mining tool rankings are clearly displayed in Figure 1.

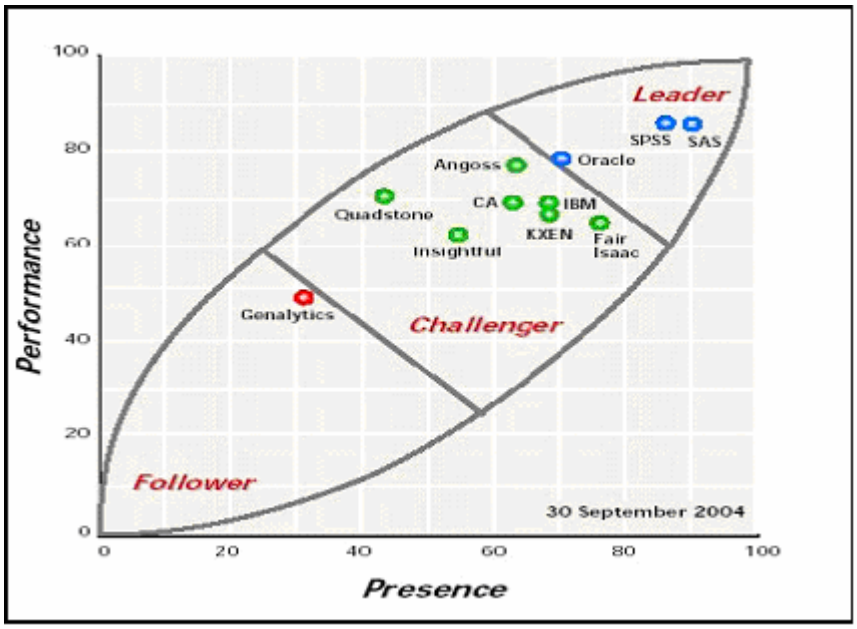


Figure 1: Evaluation of data mining tools from META Group

From this figure we can clearly see that there is tough competition amongst the mining tools in the market. This is mainly influenced by the functionalities that the mining tools possess. These may include the ease of using the tool (whether it provides wizards to help user follow the mining process) and how the tool accesses data from the database.

2.1.1 Oracle Data Miner

Oracle10g Data Miner (ODM) is the mining tool that I selected for use in the evaluation of the algorithms. ODM version 10.1 is a data mining software embedded in the Oracle 10g Database Enterprise Edition (EE) that enables one to discover new insights hidden in data [Berger, 2004]. The Oracle Data Mining suite is made up of two components, the data mining Java API and the Data Mining Server (DMS). The DMS is a server-side, in-database component that performs data mining that is easily available and scalable. The DMS also provides a repository of metadata of the input and result objects of data mining. The Oracle Data Miner is a component that resides within the DMS and is responsible for the actual data mining.

All data mining components have been added to the Oracle Data Miner, this also includes the ODM Browser interface which provides a full set of mining wizards. A data analyst can perform data mining tasks without the need to generate code or to use Oracle Java Developer as it was the case with older versions of Oracle data mining [Oracle 2005].

[Berger, 2004] states that Oracle Data Miner supports supervised learning techniques (classification, regression, and prediction problems), unsupervised learning techniques (clustering, associations, and feature selection problems), attribute importance techniques (for finding key variables), text mining, and has a special algorithm for life sciences sequence searching and alignment problems. The availability of these algorithms provides all the necessary tools required in gathering information from the dataset.

The main advantage about Oracle Data Miner is that all data mining processing occurs within the Oracle database. I took this into consideration when selecting the mining tool to use. Other mining tools force you to extract the data out of the database before the actual mining process which may result in a number of flaws in these mining tools. The aspect of ODM were everything occurs in the database results in a more secure and stable data management which enhances productivity as the data does not have to be extracted from the database before mining it.

Another fact I took into account about this mining tool is that since the ODM model building and model scoring functions are accessible through the Java API and the Oracle Data Miner's graphical user interface (GUI), the combination of the two enables Oracle to provide an infrastructure for one to integrate data mining seamlessly with database applications.

All the functionalities and advantages mentioned here have resulted in Oracle Data Miner being the most preferred mining tool to use when performing the algorithm evaluation and gathering of information from the dataset.

2.2 Choice of Algorithms

The two algorithms I selected for the evaluation are the K-Means and O-Cluster. Both algorithms are clustering algorithms available in ODM. The decision to evaluate these algorithms is influenced by a number of factors which I will discuss in detail.

Firstly, Oracle data mining supports classification, clustering, association and regression algorithms and from these only the classification algorithms available in Oracle Data Miner. These include the Adaptive Bayes Network and the Naïve Bayes which have been evaluated by [Davis, 2004]. This leaves other algorithm classes pending investigation. I therefore found it valuable and thus necessary to investigate the clustering algorithms in ODM.

[Davis, 2004] describes an investigation of the commercial data mining suite, which is available with Oracle9i database software which primarily involves determining classification algorithm effectiveness and efficiency. As a brief conclusion drawn by the author regarding the investigation, Oracle Data Mining provides all the functionality necessary to easily build an effective data mining model. Furthermore she concludes that the Adaptive Bayes Network algorithm produced the most effective data mining model as well as producing the most accurate results when the model is applied to new data.

The work by [Davis, 2004] plays a significant role in the choice of algorithms to investigate. Here I am left with the choice of investigating the Clustering algorithms. However this will provide Oracle data miners with a clear idea concerning the effectiveness of algorithms available in Oracle Data Miner.

It is also worth mentioning that the results by this author are valuable as they give me an indication of which algorithm to use when employing a supervised learning algorithm (classification) when determining cluster quality. This will be discussed in the following chapters.

2.3 Chapter Summary

This chapter has given a brief highlight on some research projects that are closely related to the evaluation of data mining algorithms and how these evaluations have benefited and influenced my research project. The reasons for selecting Oracle Data Mining for this research have been discussed as well as why clustering algorithms will be investigated.

The next chapter will describe the methods adopted for this investigation, as well as how the techniques are to be implemented.

CHAPTER 3

3 Research Approach

This chapter provides a detailed description of the approach that I took in order to perform an effective evaluation. It explains the methods adapted for the evaluation of both the algorithms and in determining the quality of clusters found.

3.1 Mining Process

In the study of data mining literature, many data mining authors felt it is of high importance that data mining should be done in a procedural manner. Many of these authors presented their own mining processes which they felt will develop non-experts to conduct data mining. With the knowledge attained from the various authors, I adapted their mining processes to a process that I then adopted. The resultant process is depicted in Figure.3 in a nutshell.

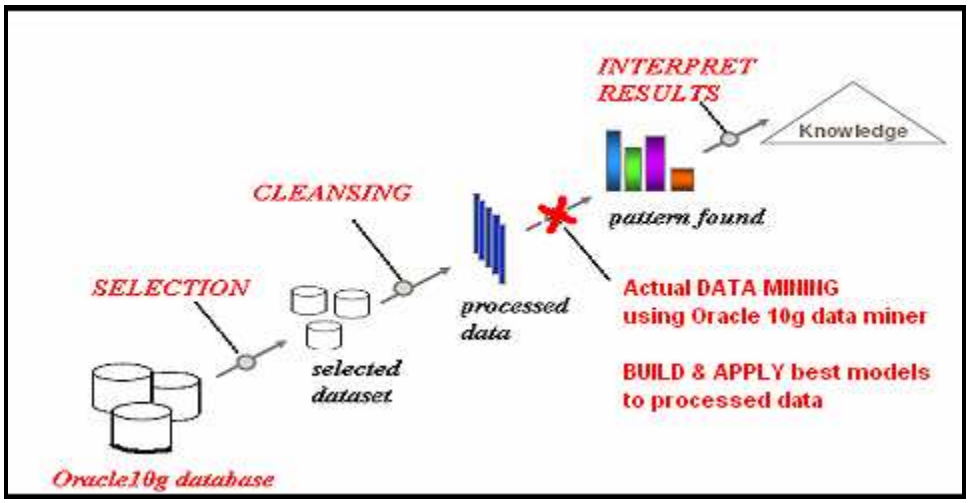


Figure 3: Mining process adopted

Although a more detailed discussion on the mining process is going to be given later in the chapter, Figure 3 simply shows that by using the Oracle 10g database, I select the dataset that I wish to mine. However at this stage proper identification of goals to be

accomplished by the data mining project will help with domain understanding and determining what is to be accomplished. At this point it is necessary to mention that my research project goals involve an effective evaluation of clustering algorithms, for this it was important to find and use understandable data which in turn will be relatively easier to interpret after the mining event.

The next step involves data preprocessing or cleansing of the dataset which primarily deals with noisy data. This stage involves locating duplicate records in the data set, locating incorrect attributes, smoothing the data and dealing with outliers in the data set. It includes data transformation which involves the addition or removal of attributes and instances, normalizing of data and type conversions.

The actual data mining stage then follows; here the models are built by the selected algorithm that sifts through the data set. The resulting model is then applied to new data which is interpreted to determine if the results it presents are useful or interesting. The acquired knowledge is then applied to the problem.

Although not depicted in Figure 3, some algorithm evaluation techniques were adopted from [Roiger et al, 2003] in building models and determining cluster quality, and these will be explained later in the chapter.

3.2 Mining Tool Used

For the purpose of the algorithm evaluation Oracle data miner version 10.1 will be used mainly due to the reasons discussed in chapter 2. Oracle10g database version 10.1.0.2.0 was installed and configured. The data mining tools and software (Oracle data miner 10g version 10.1.0.2.0) was also installed and configured for use with the database.

The Oracle Data Miner 10g is a user interface to Oracle Data Mining (ODM 10.1.) which replaces the Data mining for Java (DM4J) an ODM Component and ODM Browser user interface that was initially used in the older versions of ODM, for instance, Oracle9i data

mining. In addition to Oracle Data Miner, there is the ODM Java Code Generator, which is an Oracle JDeveloper extension [Oracle, 2005].

Although the Oracle data mining tool was installed successfully, there were a number of installation and configuration flaws encountered. These are discussed in Appendix A in detail resulting in a working configuration manual (Appendix B). I also provided an introductory manual before the user can engage to the data mining tutorial for this particular mining tool (Oracle data miner) which is Appendix C. Appendix C is very useful because if the user is new to the tool, he normally finds himself stuck not knowing what to do next.

3.3 Clustering Algorithms

Clustering is used to identify distinct segments of a population and to explain the common characteristics of members of a cluster, and to also determine what distinguishes members of one cluster from members of another cluster. Clustering algorithms make use of the Euclidean distance formula to determine the location of data instances and their position in clusters and so requires numerical values that have been properly scaled [Han et al, 2001].

Oracle Data Miner supports two clustering algorithms and these I selected for investigation in this research project. The reasons why I selected these algorithms for investigation have been discussed in the preceding chapter. The algorithms are, namely:

- *The Enhanced version of **K-Means** and,*
- *Proprietary Orthogonal Partitioning Clustering (**O-Cluster**) algorithm*

Both the Enhanced k-Means (EKM) and Proprietary O-Cluster algorithm support identifying naturally occurring groupings within the data population [Berger, 2004].

The Enhanced version of K-Means algorithm supports hierarchical clusters, handles numeric attributes and will cut the population into the user specified number of clusters. The proprietary O-cluster algorithm handles both numeric and categorical attributes and will automatically select the best cluster definitions [Berger, 2004]. Both algorithms divide the data set into k number of clusters according to the location of all members of a particular cluster in the data. When choosing the number of clusters to create, it is possible to choose a number that doesn't match the natural structure of the data which leads to poor results. For this reason [Berry et al, 2000] suggests that it is often necessary to experiment with the number of clusters to be used.

3.4 Algorithm Settings in Model Building

Each of the two Clustering algorithms in ODM has a setting that is used to tune the algorithm when building models. Both algorithms also have a parameter; the maximum number of clusters (k), this is available so that the user can pre-define the number of clusters that he wishes to find from the dataset.

3.4.1 Building Models by using the K-Means Algorithm

There are two settings for this algorithm when building models, Minimum Error Tolerance and Maximum Iterations. These determine how the parent-child hierarchy of clusters is formed and can be modified experimentally to observe the changes in cluster definitions. Increasing the tolerance or lowering the iteration maximum will cause the model to be built faster, but possibly with more poorly-defined clusters [ODM tutorial, 2004].

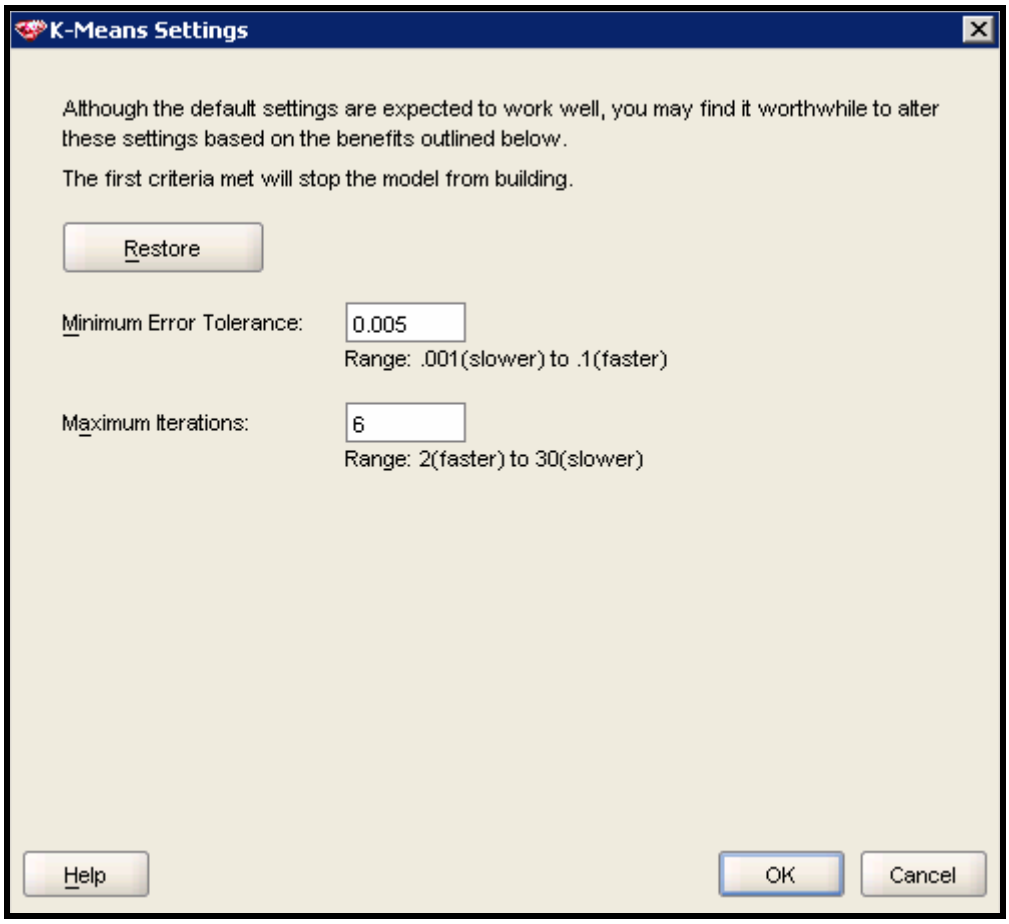


Figure 4: K-Means algorithm settings

3.4.2 Building Models by using the O-Cluster Algorithm

O-Cluster finds natural clusters up to the maximum number entered as a parameter. That is, the algorithm is not forced into defining a user-specified number of clusters, so the cluster membership is more clearly defined. O-cluster has only one setting the Sensitivity; it determines how sensitive the algorithm is to differences in the characteristics of the population. Thus, a higher sensitivity value usually leads to a higher number of clusters [ODM tutorial, 2004].

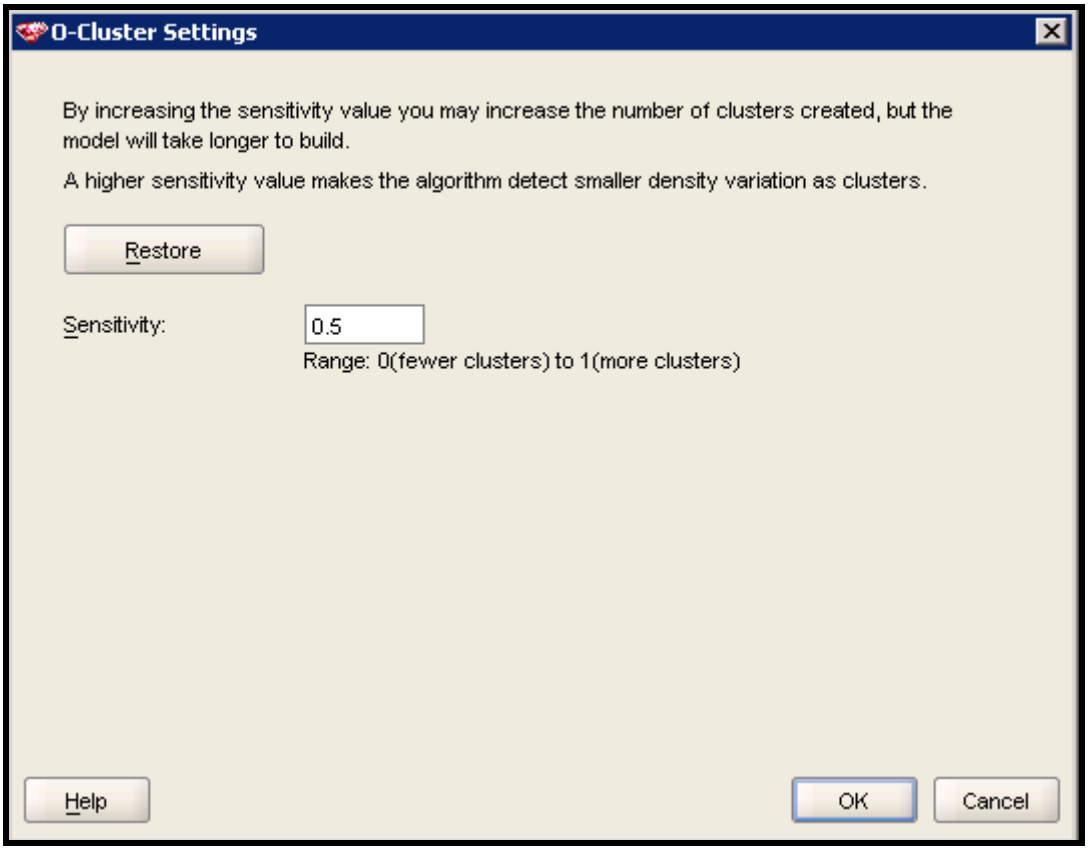


Figure 5: O-Cluster algorithm settings

It must be noted that the main purpose of these algorithm settings is to fine tune the algorithm so as to achieve finer and more accurate clusters. These settings are also useful in this evaluation as they will be used to distinguish the models built as well as to see if the accuracy of an algorithm is affected by the change in these settings. The algorithm settings will be based on trial and error followed by a critical analysis.

3.5 Dataset Used

The dataset used in the evaluation for this research was obtained from the Centre for AIDS Development, Research and Evaluation Institute for Social and Economic Research, Rhodes University. The data is a result of a questionnaire survey relating to HIV AIDS as well as a South African television drama, Tsha Tsha, which is a HIV AIDS

awareness program. A copy of the complete questionnaire can be found in the **Appendix D** on the CD-ROM that accompanies this project.

The survey was conducted by the above research institute around three provinces in South Africa namely, Gautang, Kwazulu Natal and the Eastern Cape. The research institutes' goal is to gather information from the three provinces as a sample of the South African population by using a questionnaire survey, then use statistical tools to determine predictors of HIV AIDS prevention behaviour.

[Berger, 2004] states that in the data mining process the *problem definition* is the most important step. Basically, this is where the domain expert decides the specifics of translating an abstract business objective. In my case, the problem definition is to use the Oracle data mining tool to determine the predictors of HIV AIDS prevention behaviour instead of using statistical tools.

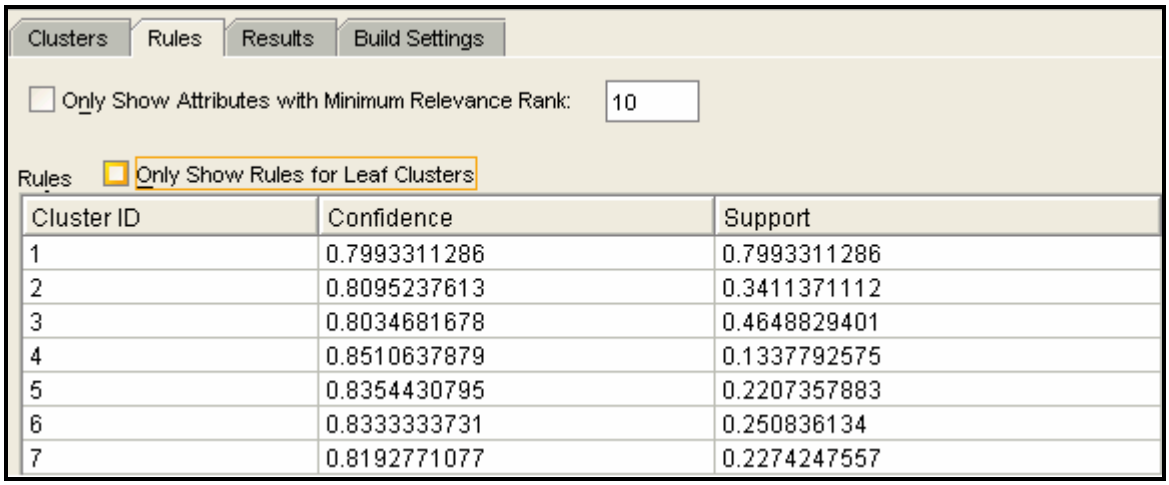
According to the researchers at the Centre for AIDS Development, Research and Evaluation Institute for Social and Economic Research, three rounds of the survey were conducted at different times with the same questionnaire resulting in three sets of data. The final datasets were then compiled by Kevin Kelly, from the above research institution, and loaded into 3 separate excel files namely, Tsha_tsha_round1, Tsha_tsha_round2 and Tsha_tsha_round3. Each of the three datasets consists of 131 attributes and more that 800 rows (cases); this is large enough for the evaluation. These dataset files are also available on the Compact Disc that accompanies this project.

3.6 Techniques Adapted

A couple of techniques were adopted so as to carryout the evaluation. These help in determining the accuracy of models built and determining the quality of clusters found after applying the models to new data. The techniques are detailed below;

3.6.1 Determining Model Accuracy

In order to be able to carry out an effective evaluation of the algorithms, it is valuable to use the best models built by either algorithm. Ideally, the best models should possess the highest probability of finding accurate clusters and must have minimum error .To determine the accuracy of each model built I make use of two parameters that the mining tool provides as output after the building of the model. These parameters are the Confidence and Support. A sample of how these parameters are displayed by the mining tool is shown in Figure 6.



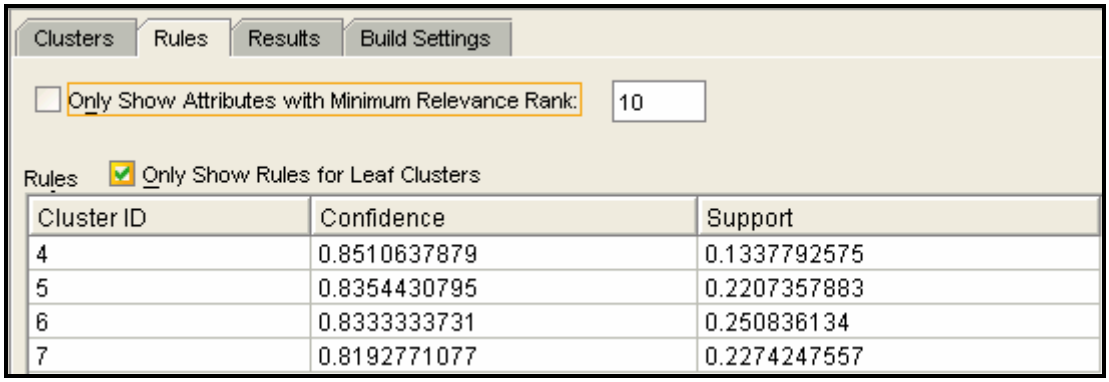
Cluster ID	Confidence	Support
1	0.7993311286	0.7993311286
2	0.8095237613	0.3411371112
3	0.8034681678	0.4648829401
4	0.8510637879	0.1337792575
5	0.8354430795	0.2207357883
6	0.8333333731	0.250836134
7	0.8192771077	0.2274247557

Figure 6: Example of model Confidence and Support values

The Confidence is a measure of the homogeneity of the cluster; that is, how close together are the cluster members [ODM Tutorial, 2004]. I therefore made it a measure of accuracy such that a cluster with the highest confidence value is more accurate and effective than that with a lower value. Thus, computing an average confidence value of all the clusters in a model would determine the accuracy of a model.

The support is a measure of the relative size of a cluster (the total need not be 1.00), such that the higher the value the larger the cluster [ODM Tutorial, 2004]. In this paper it is used as an alterative measure of accuracy to the confidence.

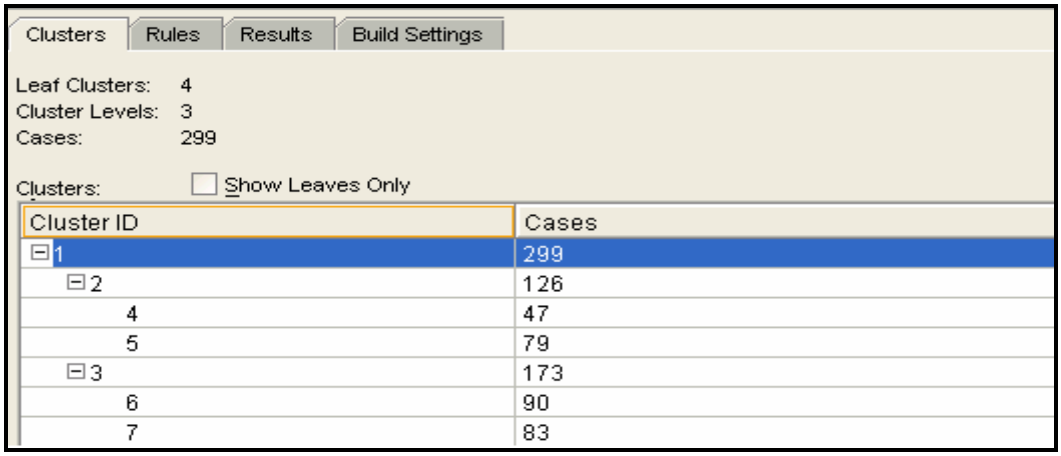
The ClusterID (see Figure 6 and Figure 8) is a value that differentiates the clusters found. The order of numbering used for the ClusterID is as follows; the mining tool generally looks for all clusters in the dataset depending on the algorithm settings. The maximum number of clusters (k) that one sets during model building determines the number of clusters that the mining tool displays as the leaf clusters. When the option *Only show Rules for leaf clusters* is active (see Figure 7), the maximum number of clusters that you selected during model building is displayed with each cluster having a distinct ClusterID. In this case the maximum number of clusters (k) was set to 4.



Cluster ID	Confidence	Support
4	0.8510637879	0.1337792575
5	0.8354430795	0.2207357883
6	0.8333333731	0.250836134
7	0.8192771077	0.2274247557

Figure 7: Example showing option **Only show Rules for leaf clusters** *active*

ClusterID = 1 will always represent the entire dataset, while ClusterID 2 and 3 are two separate clusters in ClusterID 1 (there can be more), this continues until the maximum number of clusters that you will have set is reached (see Figure 8).



Cluster ID	Cases
⊟1	299
⊟2	126
4	47
5	79
⊟3	173
6	90
7	83

Figure 8: A clearer view of the ClusterID structure in ODM

3.6.2 Determining Algorithm Accuracy

The evaluation of unsupervised data mining algorithms is a generally difficult task, since the goals of an unsupervised data mining session are frequently not as clear as the goals of supervised learning. However, [Roiger et al, 2003] suggests and presents a number of techniques for the evaluation of unsupervised algorithms and models. To increase the robustness of this evaluation, I plan to adapt and implement the evaluation techniques by these authors.

The techniques by [Roiger et al, 2003] are explained fully on pages 58 and 232. Here the four main methods that I adopt are as follows:

1) Employing supervised learning to evaluate unsupervised learning.
2) Apply alternative technique's measure of cluster quality.
3) Create own measure of cluster quality (in this case making use of Confidence).
4) Perform a between cluster attribute value comparison.

Although all of the above techniques will be employed at some stage of the evaluation, it is important to mention that the technique of *employing supervised learning to evaluate the unsupervised learning* is the most crucial technique. I will use it to determine cluster quality as I will evaluate cluster accuracy. [Roiger et al, 2003] details the technique as follows:

i. Perform an unsupervised clustering. This step involves the building of models from Cluster algorithms then apply the best model (most accurate) to new dataset. Then designate each cluster found in the dataset as a class and assign each cluster an attribute name. For example, if the clustering technique outputs three clusters, the clusters could be given the class names C1, C2 and C3. The Oracle data mining tool when applying the cluster models to new data includes another attribute the ClusterID, this attribute signifies a cluster digit that an instance belongs to.

ii. Choose a random sample of instances from each of the classes formed as a result of the instance clustering. Each class should be represented in the random sample in the same ratio as it is represented in the entire dataset. The percentage of the total instances to sample can vary, but a good initial choice is two-thirds of all instances.

iii. Build a supervised learner model with class name as the output attribute using the randomly sampled instances as training data. Employ the remaining instances to test the supervised model for classification correctness.

Using this technique I will make use of a classification algorithm (a supervised learning algorithm). I plan to use the Adaptive Bayes Networks (ABN) algorithm for building the classification models. Making use of the ABN is motivated by the results obtained by [Davis, 2004] which concluded that the algorithm is more accurate in predicting attributes for the classification algorithms in Oracle.

As a brief summary of the technique by [Roiger et al, 2003] but applying it to my situation is as follows: I will use Oracle data miner as the mining tool. The evaluation technique will involve taking the resultant table (for instance table APPLY_RESULTS) obtained after applying a cluster model, then pick a random sample of instances (roughly two thirds) from each cluster found from this table and place them in another table that will be used to build a classification model using the Adaptive Bayes Network (ABN) algorithm. The resultant model is then applied to the remaining instances from the table APPLY_RESULTS. The attribute being predicted by the classification model (ABN model) is the ClusterID. The results after applying the ABN model which predicts the ClusterIDs are compared to the ClusterIDs in the table APPLY RESULTS. The ClusterIDs for a particular instance is uniquely identified by the instance's REFNUM (reference number). The ClusterIDs that occur in both tables are counted and a percentage is found. This percentage is used as a measure of cluster accuracy.

3.7 Chapter Summary

This chapter has discussed in detail the approach that will be taken in evaluating the two clustering algorithms available in Oracle data mining 10.1. The next chapter is solely dedicated to the implementation of the methods presented in this chapter.

CHAPTER 4

4 Method Implementations

This chapter provides an insight on how Oracle Data Miner provides data mining functionality as I implement the methods discussed in the preceding chapter. It describes the pre-processing of the data as well as the actual model building. The best models by each algorithm will be identified and applied to new data. The actual evaluation is performed at this stage and the process is explained in this chapter.

4.1 Data Preparation

For this evaluation, the datasets Tsha_tsha_round1 and Tsha_tsha_round2 were used. Tsha_tsha_round1 was used to perform the main evaluation while Tsha_tsha_round2 was used to repeat the whole evaluation process implemented on Tsha_tsha_round1 mainly for verification and confirmation of the results.

Before the datasets were loaded into the database tables initial, pre-processing of the datasets was necessary. This stage involves locating duplicate records in the data set, locating incorrect attributes, smoothing the data, dealing with null values, outliers, inconsistencies and table properties. It also involves the removal of rows with no values at all.

Another good factor about Oracle Data Miner (ODM) is that it also provides an option that helps in the data preparation step. Here on can specify automatic binning for the input data, were the format of the data is automatically converted to one that is understood by the mining tool.

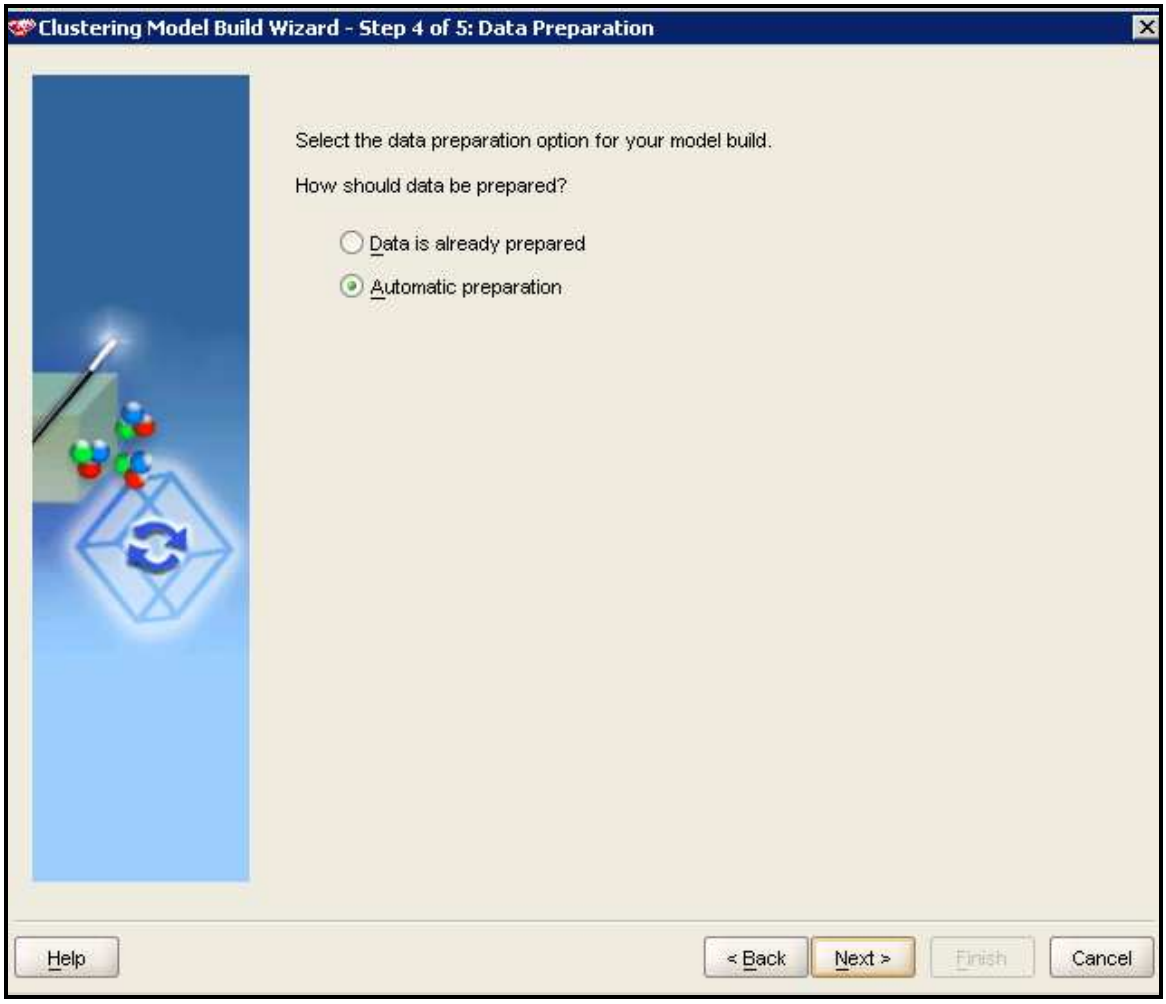
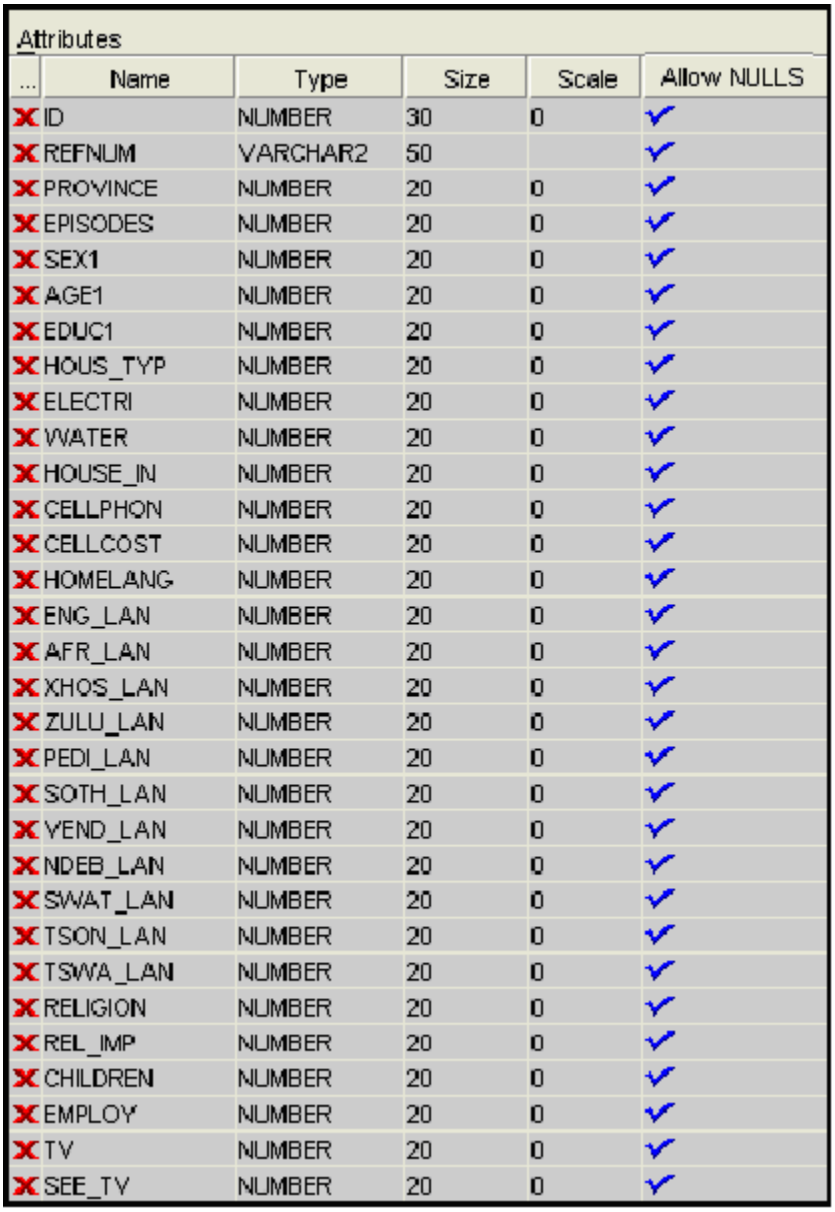


Figure 9: Data preparation

Since the major evaluation process was to be performed on the data Tsha_tsha_round1, the dataset was then partitioned into three datasets (with a random selection of instances). I then created three identical database tables, one that will be used to build models, the other to apply the built models and the third for testing the accuracy of models. The tables are TSHA_TSHA_BUILD1, TSHA_TSHA_APPLY1 and TSHA_TSHA_APPLY2 with each table having 131 attributes. A sample of the resulting structure of the database tables is depicted in Figure 10.



Attributes

...	Name	Type	Size	Scale	Allow NULLS
	ID	NUMBER	30	0	✓
	REFNUM	VARCHAR2	50		✓
	PROVINCE	NUMBER	20	0	✓
	EPISODES	NUMBER	20	0	✓
	SEX1	NUMBER	20	0	✓
	AGE1	NUMBER	20	0	✓
	EDUC1	NUMBER	20	0	✓
	HOUS_TYP	NUMBER	20	0	✓
	ELECTRI	NUMBER	20	0	✓
	WATER	NUMBER	20	0	✓
	HOUSE_IN	NUMBER	20	0	✓
	CELLPHON	NUMBER	20	0	✓
	CELLCOST	NUMBER	20	0	✓
	HOMELANG	NUMBER	20	0	✓
	ENG_LAN	NUMBER	20	0	✓
	AFR_LAN	NUMBER	20	0	✓
	XHOS_LAN	NUMBER	20	0	✓
	ZULU_LAN	NUMBER	20	0	✓
	PEDI_LAN	NUMBER	20	0	✓
	SOTH_LAN	NUMBER	20	0	✓
	VEND_LAN	NUMBER	20	0	✓
	NDEB_LAN	NUMBER	20	0	✓
	SWAT_LAN	NUMBER	20	0	✓
	TSON_LAN	NUMBER	20	0	✓
	TSWA_LAN	NUMBER	20	0	✓
	RELIGION	NUMBER	20	0	✓
	REL_IMP	NUMBER	20	0	✓
	CHILDREN	NUMBER	20	0	✓
	EMPLOY	NUMBER	20	0	✓
	TV	NUMBER	20	0	✓
	SEE_TV	NUMBER	20	0	✓

Figure 10: Sample of Mining Data Table Structure

The technique and tool for loading data into database tables is explained in **Appendix C**. The technique involves the use of an application called sql*loader. Using this application I loaded the three datasets into the tables that I created.

4.2 Cluster Models

The building of cluster models was done in two stages. The first stage involved building models with random algorithm settings as well as distinct cluster size such that each model has distinct parameters. This is to determine if cluster size has any significant effect on algorithm accuracy. The building of models in the second stage was dependent on model accuracy results obtained from the first stage; the second stage of model building involved setting the maximum number of clusters (k) for both algorithms to the same value for all models built. This basically helps in evaluating the resultant model accuracy.

4.2.1 Building of Models

In the first stage ten models were built in total, with five from each algorithm. Since the model settings are based on trial and error, it was necessary to build a large number of models to provide a wide range of models to select the best from. This also helps when analysing model accuracy to determine if the algorithm settings do affect the algorithm's performance.

The ten models all have distinct values of the maximum number of clusters (k). Here, each model built from the K-Means algorithm was named in the form BUILD1_KM_TSHATSHA while models built by the O-cluster algorithm were named in the form BUILD1_OC_TSHATSHA. The digit 1 in these model names denotes the first model, digit 2 second model, and so on. Figure 11 shows the models and their settings in detail,

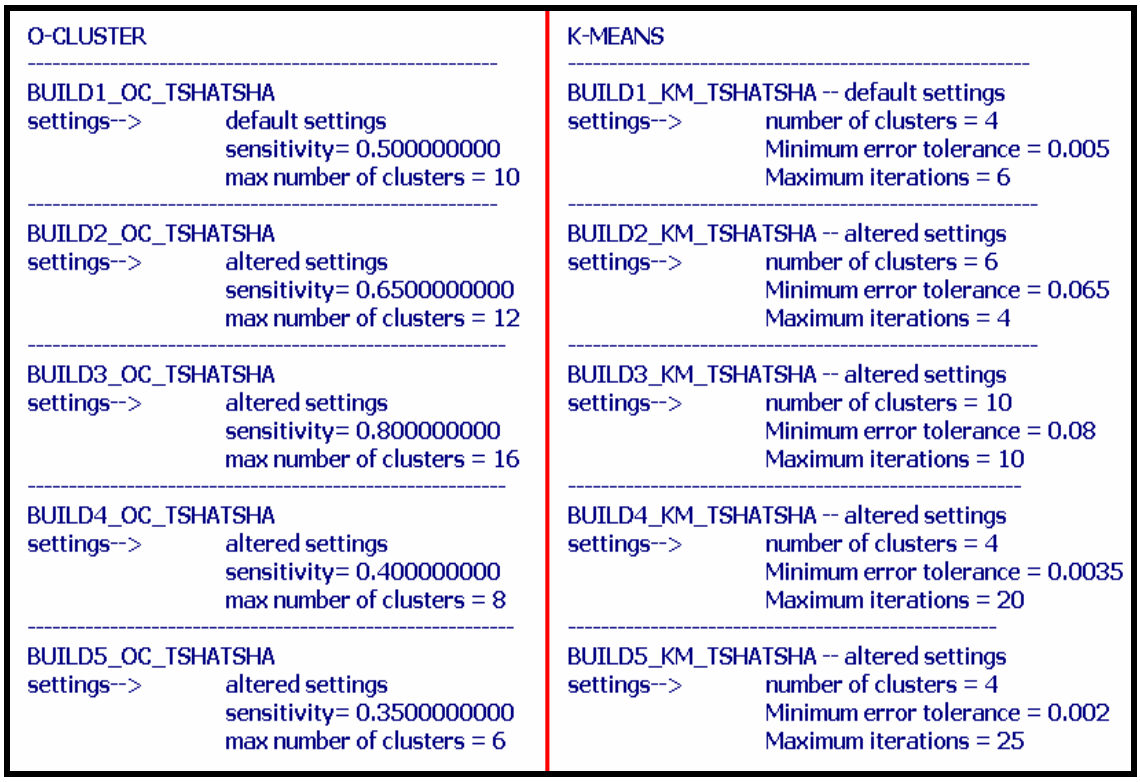


Figure 11: Algorithms, model names and their settings with distinct number of clusters

4.2.2 Interpretation of Initial Model Results

After building the ten models from the algorithm settings in Figure 11, it was observed that each model did actually discover clusters. This is evident from the ClusterIDs, Confidence and Support values that the mining tool depicts on displaying the model results. A sample of the model output was shown in both Figure 6 and 7. For all the models that I built in this stage, I computed an average confidence value and average support which both determine model accuracy; these are depicted in Figure 12.

On analysing these average confidence values computed from the model clusters, I observed a high degree of bias in the results. Here the bias is mainly due to the variation in the value of the maximum number of clusters (k) that was set for each algorithm during model building as shown in Figure 11 (algorithm settings). This makes it difficult to determine the best model built from the two algorithms.

O-Cluster

BUILD1_OC_TSHATSHA -- k=10

id	confidence	support
average	0.891825782	0.0856187292

BUILD2_OC_TSHATSHA -- k=12

id	confidence	support
average	0.9393461558	0.075808249

BUILD3_OC_TSHATSHA -- k=13

id	confidence	support
average	0.984528279	0.0735785951

BUILD4_OC_TSHATSHA -- k=8

id	confidence	support
average	0.84385229	0.09239130

BUILD5_OC_TSHATSHA -- k=6

id	confidence	support
average	0.8028507	0.7792642

K-Means

BUILD1_KM_TSHATSHA -- k=4

id	confidence	support
average	0.83477935	0.208193985

BUILD2_KM_TSHATSHA -- k=6

id	confidence	support
average	0.8638370617	0.14214046667

BUILD3_KM_TSHATSHA -- k=10

id	confidence	support
average	0.889283927	0.0882943151

BUILD4_KM_TSHATSHA -- k=4

id	confidence	support
average	0.83765804	0.20903010

BUILD5_KM_TSHATSHA -- k=4

id	confidence	support
average	0.82141478	0.16555184

Figure12: Average confidence values for the 1st 10 models (biased results)

On analysing Figure 12, I observed a general trend in the average confidence values. The trend shows that increasing the maximum number of clusters for each algorithm results in a higher confidence value. Hence increasing the value of k has a significant effect on the accuracy of the algorithm in model building.

However, observations show that these model results are biased therefore are not suitable for this evaluation. The biasness comes from varying the value of k when building models. When I look at the models built using the O-Cluster algorithm in Figure 12, I observe that increasing k results in more accurate models as the average confidence values increases. This is also the same case with the K-Means algorithm but it is apparent that the O-cluster models are more accurate. Thus it would be false to conclude from these biased results that the O-Cluster algorithm builds better models.

To overcome the problem of bias, I then decided to set the value for the maximum number of clusters (k) to a fixed value. I set the value k to 7 for both algorithms because, the default number of clusters for the *K-Means* algorithm in ODM is 4 and that for *O-Cluster* is 10, therefore setting the value of k for the two algorithms to an average of the two default values was reasonable.

I then re-built 10 more models with the same settings as in Figure 10 but with the maximum number of clusters (k) fixed at 7 for all models. The new model names are in the form BUILD1_OC_TSHATSHA2 for O-cluster and BUILD1_KM_TSHATSHA2 for the K-means, with the digit 2 at the end indicating the second set of models built. These models built at this stage are unbiased. I then made a new computation of average confidence and average support figures as shown in Figure 13.

Figure.13. Computed confidence and support averages for the models built

From Figure 13, it is evident by analysis that the model BUILD3_OC_TSHATSHA2 from the O-Cluster and the model BUILD5_KM_TSHATSHA2 from the K-Means posses the highest values for both the average confidence and support values.

Critical observations of these results shows that changing the settings for the K-Means algorithm has little effect on the clusters found. This results in very small differences in the computed average values for models built by this algorithm. On the other hand, the O-cluster algorithm models were affected greatly by the changes in the settings where the average confidence increases greatly with k.

In conclusion it is evident from Figure 13 that the O-cluster algorithm builds more accurate models than the K-Means algorithm.

4.3 Applying the Best Models

The two models identified in the preceding section as the best and most accurate were BUILD3_OC_TSHATSHA2 from the O-Cluster algorithm and BUILD5_KM_TSHATSHA2 from the K-Means. These two models were applied to the new data TSHA_TSHA_APPLY1.

At this stage the mining tool allows you to select attributes that will be displayed in the output table after applying the models. This facility becomes valuable in this case since the dataset that we are using has a large number of attributes (131). Since the primary goal with this dataset is to determine predictors of HIV AIDS prevention, I then selected as many potential predictors. This filters out those attributes that I felt had little significance to HIV AIDS. Although, this may result in the loss of information it is necessary as it makes the analysis and interpretation of the clusters simpler.

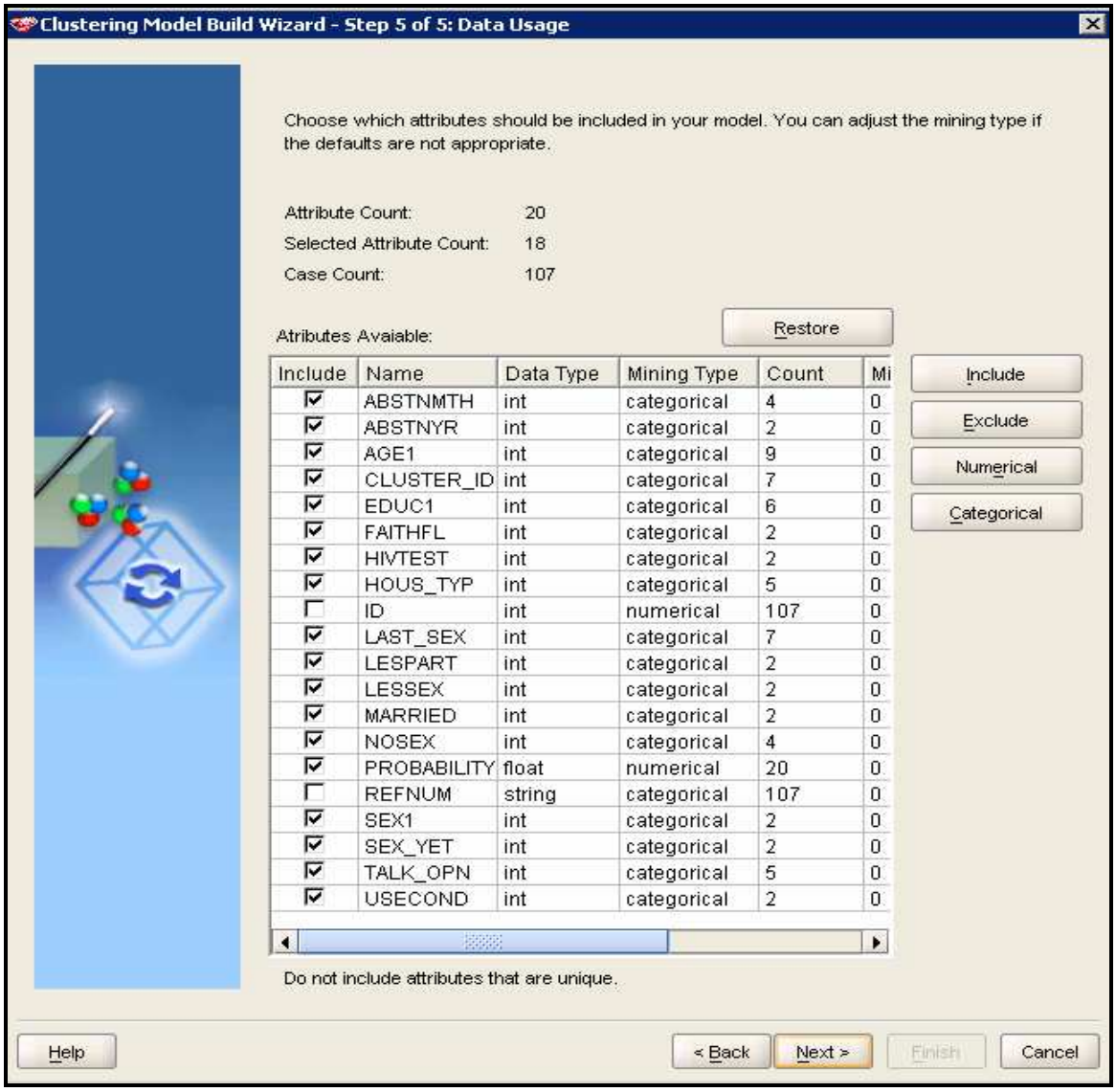


Figure 14: Facility for selecting attributes

The mining tool also provides a wizard which supplies a name for the output table on applying the models; this name may be modified at this stage. On applying the models clusters were discovered in the new data. These cluster results were loaded into the tables APPLY_OC3_TSHATSHA for the O-cluster and APPLY_KM5_TSHATSHA for the K-Means. The results were then exported to spreadsheets to allow further inspection for algorithm evaluation. This will be discussed in the following sections. The interpretation of the clusters will also be discussed in the next chapter after I have given reasons and proved why I selected a particular algorithm as one that provides more accurate results than the other.

4.4 Testing of Model Results

This section primarily deals with determining cluster quality from the cluster results obtained after applying the cluster algorithm models. Ideally, this involves finding out which algorithm model finds more accurate clusters. Although, [Roiger et al, 2003] indicates that the evaluation of clustering algorithms is difficult, I intend to use the technique that these authors proposed as discussed in Chapter 3, section 3.5.2

4.4.1 A Brief Recap of Technique to Determine Cluster Quality

The technique is by [Roiger et al, 2003] and it uses supervised learning evaluation to evaluate unsupervised clustering. Here I make use of a classification algorithm (supervised learning), the Adaptive Bayes Networks (ABN) algorithm with the technique. Making use of the ABN was motivated by the results obtained by [Davis, 2004] which concluded that the algorithm is more accurate in predicting attributes for the classification algorithms in Oracle Data Miner.

Basically, the evaluation technique involves taking the resultant table obtained after applying a cluster model (APPLY RESULTS), pick a random sample of instances (roughly two thirds) from each cluster found, place them in a new database table that will be used to build a classification model, in this case using ABN. The attribute being predicted is identified; in this case it will be the ClusterID. The resultant model is then applied to the remaining instances from the APPLY RESULTS (the one third of instances) with the ClusterIDs removed (stored in excel for comparison later). The results after applying the ABN model which predict the ClusterIDs are then compared to the initial APPLY RESULTS ClusterIDs (the cluster ids in the excel file).

4.4.2 Building Classification Models

Two database tables, *build1_abn_FROM_OC* and *build1_abn_FROM_KM* were created and these are used for building the Classification models with the ABN algorithm. The table *build1_abn_FROM_OC* was loaded with two thirds of instances from the O-cluster model results table (APPLY_OC3_TSHATSHA) created as was explained in section 4.3. Two thirds of instances from the table APPLY_KM5_TSHATSHA (section 4.3) were also loaded into the table *build1_abn_FROM_KM* to cater for the K-Means model results evaluation.

The steps taken in building ABN models are clearly explained in the research by [Davis, 2004] and I did is simply adopt these steps. Since [Davis, 2004] concluded that the ABN algorithm provides more accurate results than the Naïve Bayes algorithm in Oracle for classification algorithms, I then decided to use the default ABN algorithm settings in building these models. These settings included a SingleFeatureBuild model type, a maximum number of predictors of 25, a maximum network feature depth of 10 and no time limit for the running of the algorithm.

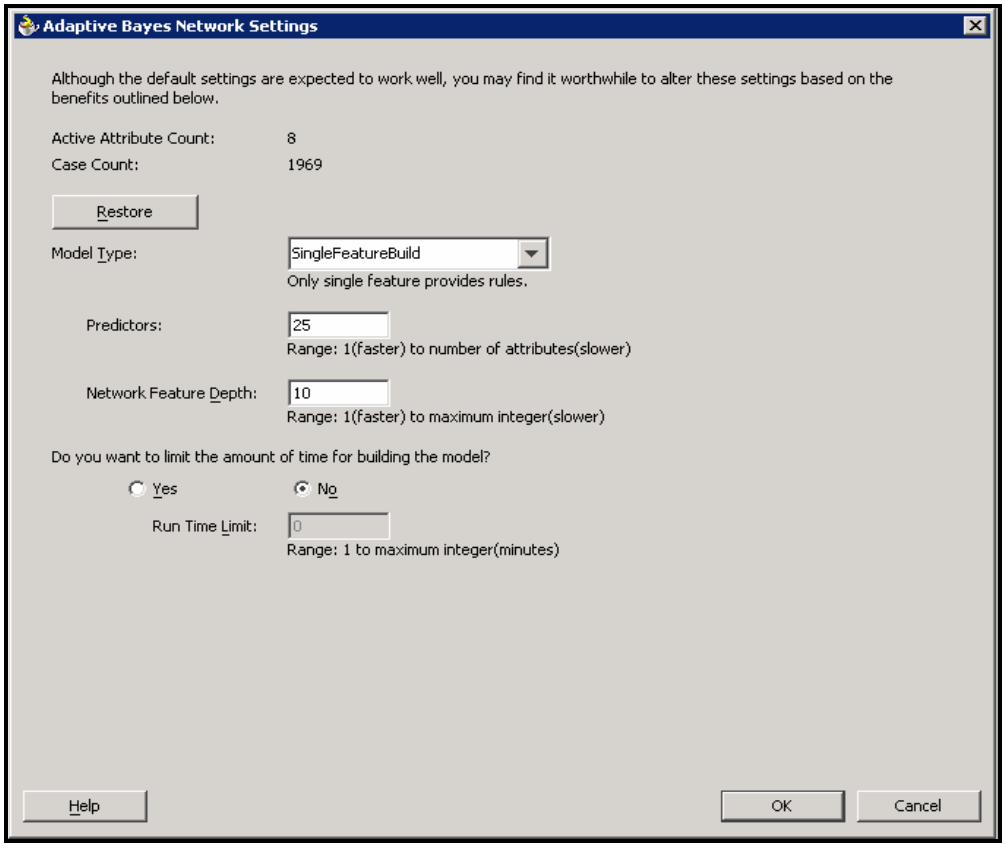


Figure 15: Adaptive Bayes Network default algorithm settings

The two models created were named *OC_abn_Build* from the dataset *build1_abn_FROM_OC* and *KM_ abn_Build* from the dataset *build1_abn_FROM_KM*. Investigating the accuracy of the models built here is unnecessary for this evaluation. This is because any errors or abnormalities found in the algorithm would exert the same effect on both models built since one algorithm with the same conditions (i.e. algorithm settings) is used.

4.4.3 Applying the Adaptive Bayes Network (ABN) Models

The resultant ABN models were applied to the remaining one third of instances from the respective cluster models. Table.1. shows a summary of the datasets used in applying a particular model during the evaluation of the clusters with the ABN algorithm.

DATASET USED TO BUILD MODEL	RESULTANT MODEL NAME	DATASET USED TO APPLY MODEL
build1_abn_FROM_OC	*OC_ abn_Build*	*apply1_abn_FROM_OC*
build1_abn_FROM_KM	*KM_ abn_Build*	*apply1_abn_FROM_KM*

Table 1: Summary of dataset and model naming

The target attribute in both instances when applying the ABN models is the ClusterID. The results were exported to spreadsheets to allow for inspection and comparisons. All the datasets used are provided in the CD ROM that accompanies this research paper.

4.4.4 Comparison of ClusterIDs

The comparison of ClusterIDs is between the classification model results and the cluster model results. Here I wish to find ClusterIDs that appear in both the two distinct model results. In this case I made a comparison of the K-Means model results with the ABN model results and a comparison of the O-Cluster model results with the ABN model results. The comparison and counting was done by importing the resultant tables into a Microsoft Access database followed by the construction of SQL queries to perform the actual comparison and counting. Table.2. gives a summary of the tables that were used in the comparison (the tables contain the ClusterIDs) while Table.3 depicts the outcome of the comparison.

CLASSIFICATION TABLE		CLUSTER TABLE
OC_APPLY_ABN	Vs	APPLY_OC3_TSHATSHA
KM_APPLY_ABN	Vs	APPLY_KM5_TSHATSHA

Table 2: Database tables compared

Table.2. shows that the classification tables from the results of the ABN model are compared with the clustering table results obtained from the cluster algorithm models.

The clustering algorithms basically find clusters in the data; ODM then assigns each data instance to a particular cluster by assigning it a ClusterID. These ClusterIDs are removed from the clustering results and are predicted by the classification algorithm, after which a comparison is made.

DATA SOURCE	CLUSTERID**S IN BOTH TABLES**	**PERCENTAGE OF CLUSTERIDS IN BOTH MODELS**
From O-Cluster results	42 out of 107	39%
From K-Means results	18 out of 107	17%

Table 3: Results from the comparison of cluster and classification **ClusterID**s

Table.3 shows the percentage of the ClusterIDs that appear in both model results. 39% of ClusterIDs appeared in both the cluster model results and classification model results for the O-Cluster algorithm while only 17% for the K-means algorithm. According to this evaluation technique by [Roiger et al, 2003], the percentage outcome is treated as a measure of accuracy for the algorithms, were the greater percentage indicates that that algorithm has more accuracy in finding clusters of high quality. According to Table 3 the O-cluster algorithm has a larger percentage hence is more accurate than the K-Means algorithm.

4.5 Chapter Summary

Initially, ten cluster models were built. However, I then discovered that these initial algorithm settings were biased in order to perform the evaluation. This bias made it difficult to point out the best two models. I then decided to build a further ten models with a slight change in algorithm settings. For the new ten models, I set the number of maximum clusters (k) to seven for all models. This removed the bias encountered in the first ten models making it easier to select the most accurate models. Here the O-cluster algorithm built the most accurate model.

It is also apparent from results that the most accurate models selected had found cluster patterns in the data on applying them to new data. The algorithm evaluation technique proposed by [Roiger et al, 2003] was then implemented. Here it was determined that the model built by the O-Cluster algorithm found more accurate clusters when applied to new data as compare to the K-Means. A further evaluation follows in the next chapter with a conclusive interpretation of the all the results.

CHAPTER 5

5 Interpreting Evaluation Results

This chapter provides an interpretation of the results obtained so far during the evaluation process. The interpretation is from all the data mining models built using similar techniques but different algorithm settings. In this chapter I extract important facts that will help and support in describing the conclusions reached regarding the algorithm that has been determined as the most accurate algorithm. Each model comparison is made then interpreted detailing the facts for the results obtained.

5.1 Comparing the 1st Ten Cluster Models

It has already been explained in preceding chapters that the results in the 1st ten models are biased therefore they provide little information for the evaluation of the two algorithms. In general the average confidence values displayed in Figure 12 (chapter 4) show that increasing k the maximum number of clusters results in the average confidence increasing for an algorithm model. The bias is that the O-Cluster seems to be producing more accurate that the K-Means when the value of k is increased or rather varied.

Although the confidence for each cluster increases with k, from a logical point of view, the clusters found tend to become less complex resulting in less interesting clusters. In simple terms the clusters loss their value. As the number of clusters increase, the data is divided into many groupings and this result in loss of attribute relationships. Therefore this should be kept in mind when increasing the value of k. However it is also difficult to indicate the right k value to accommodate both accuracy and cluster meaning.

The default maximum number of clusters for the O-Cluster algorithm is 10 while that for the K-means is 4. In Figure 12 the models BUILD1_OC_TSHATSHA from the O-Cluster and BUILD1_KM_TSHATSHA from the K-Means algorithm were built from the

algorithm default settings. On analysing their average confidence values one can see that the O-cluster algorithm produced a more accurate cluster model. Further more; it is also clear that the O-cluster algorithm seems to be producing models with a much higher average confidence value as compared to the K-Means models with increased k.

5.2 Comparing the 2nd Ten Cluster Models

Figure 13 in chapter 4 shows the average confidence and the support values computed for each model. The major difference in the 2nd model building phase in the algorithm settings for these ten models as compared to the settings for the 1st ten models is in the maximum number of clusters that I set to seven for all the models. Making this parameter, the maximum number of clusters, to be the same for all the models made the evaluation process unbiased.

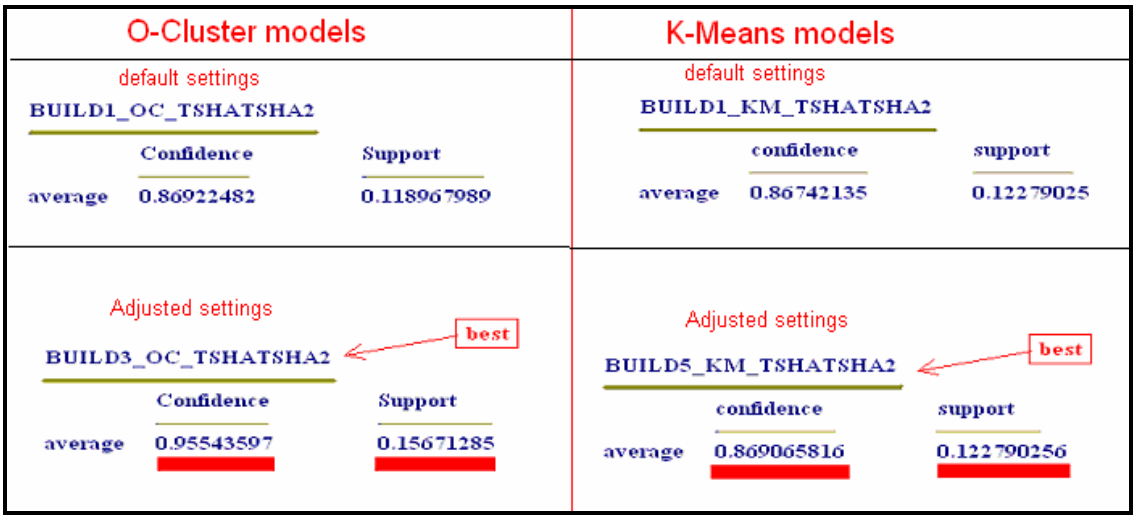


Figure 16: Comparison of models with default and adjusted settings

When the comparison is inspected from Figure 16 (shortened version of Figure 13), it is clear that the average confidence for the O-Cluster models is greater than that for the K-Means in both cases. Since the models were all built with the maximum number of clusters set to seven it is evident that the O-Cluster algorithm builds more accurate models as compared to the K-Means algorithm. This conclusion is primarily based on the

results obtained after the building of models with a variation in algorithm settings for both algorithms.

5.3 Accuracy of Algorithms

In order to determine algorithm accuracy a technique proposed by [Roiger et al, 2003] was adapted. The technique basically employs supervised learning algorithms to evaluate unsupervised algorithms. Although the technique determines the final outcome regarding algorithm accuracy by indicating which algorithm produces more accurate cluster results, another factor was considered. This deals with how accurate an algorithm is in building models. This makes it possible to determine whether model accuracy gives a good indication of model performance when applied to new data. It is evident as discussed in sections 5.1 and 5.2, that from the models built and the average confidence figures computed, the O-Cluster algorithm resulted in the building of more accurate models.

Also to be emphasised is the effect that the maximum number of clusters (k) has on the performance of the algorithm during model building. It is clear that for the K-Means algorithm, the variation of k, the maximum number of clusters, has very little effect on the accuracy of the resultant model while the opposite is true for the models built by the O-cluster.

Once the models had been applied to new data, the results of this step had been used in the building of Adaptive Bayes Network (ABN) algorithm models. This was followed by applying the models to the remaining one third of instances (basic implementation of unsupervised evaluation using technique proposed by [Roiger et al, 2003]). From these results it is significant that the best model built by the O-Cluster algorithm produces more accurate cluster results when applied to new data. The implementation step is discussed in section 4.3.4.

The conclusions drawn here are that the O-Cluster algorithm produces both more accurate models and more accurate clusters when applied to new data as compared to the K-Means algorithm in Oracle data mining.

5.4 Chapter Summary

It is apparent in this chapter that from the results obtained from the different models that the most effective model was built using the O-cluster algorithm. The results also show that the model built using the O-cluster finds more accurate cluster when applied to new data. This was determined by employing supervised learning evaluation for unsupervised learning, a technique by (Roiger et al, 2003).

The next chapter involves gathering information from the dataset. This will need distinguishing the clusters found by the O-Cluster model which provides more accurate cluster results.

CHAPTER 6

6 The Gathering of Information

This chapter is mainly concerned with gathering information from the dataset. Here I need to find predictors of HIV AIDS prevention behaviour. In order to this it is necessary to define what we mean by predictors of prevention behaviour. This makes it easier and feasible to know what we are actually looking for exactly. Therefore my definition is as follows:

HIV AIDS predictors of prevention behaviour are attributes within our dataset that influence an individual to:

a) *use a condom when he/she decides to be sexually active,*
b) *lead to abstain from having sexual intercourse for at least a year or more and*
c) *attributes that lead one to having fewer sexual partners. These are the attributes that I want to find from the dataset.*

Now that I have a definition of what the predictors of prevention are, how do I intend to find these attributes? I firstly intend to use a trial and error mechanism that involves making use of the O-Cluster algorithm to find clusters that contain either one of the attributes **use_ condom** (which means that the person either makes use of a condom or not during sexual intercourse), **abstnyr (**meaning person has abstained for a year or more**)** or **lespart (**meaning has decided to have less or fewer partners**).**

The other approach that I used in order to determine the predictors was to employ the association rule algorithm (the Apriori). Association rule mining searches for interesting relationships among items in a given data set.

6.1 Determining Predictors by Distinguishing Clusters

The approach I took was to build a cluster model with the O-cluster algorithm because from my algorithm evaluation, I determined that the O-Cluster algorithm builds more accurate results than the K-Means algorithm and that this model by the O-Cluster algorithm also finds more accurate clusters when applied to new data. Here I simply selected for use the model that I determined as the most accurate model by the O-Cluster algorithm during the evaluation process and this was the model **Build3_OC_tsha_tsha2,** this model was built from the dataset **Tsha_tsha_Build2**. I then applied the model to new data to find the clusters and the new data was in table **Tsha_tsha_Apply1**.

After I applied the model to new data it is apparent that cluster patterns were found and this is shown in the resultant table **APPLY_OC3_TSHATSHA**, this is evident from the allocation of ClusterIDs to the dataset instances. The results are exported to spreadsheets once the cluster model had been applied to new data. The spreadsheet file name is **APPLY_OC3_TSHATSHA**. Table.4 shows the distribution of instances in the clusters found.

CLUSTER_ID	***NUMBER OF INSTANCES***	***PERCENTAGE OF INSTANCES (%)***
4	55	18
6	54	18
8	9	3
10	61	20
11	47	16
12	50	17
13	23	8
Totals	299	100

Table 4: Summary of clusters in APPLY_OC3_TSHATSHA

To meet my goal of finding predictors I first have to distinguish the clusters found. However from this resultant table **APPLY_OC3_TSHATSHA** it proves to be difficult due to the large number of attributes in the table (number of attribute is 21). To overcome this I simply re-apply the same O-Cluster model to the same dataset, but in the output

table I removed any attribute that I felt had little contribution to solving this problem. This step was repeated nine times with each predictor that was mentioned in my definition of prevention predictors being in each table. The outcome is detailed below corresponding to each table result.

1. Table **ANALYSE_CLUSTER_1**

This is the first output table after applying the model to new data and was named table ANALYSE_CLUSTER_1 (simply indicating 1st set of cluster analysis). This table includes the predictor **use_condom** from the definition above. The other attributes that were included here were **EDUC1 (education level), LESPART (less/fewer partners)** and **SEX1 (which is gender).** This was to see if these attributes have any relation or influence to why one would result in using a condom to prevent from being affected by the HIV virus.

CLUSTER_ID	PROBABILITY	EDUC1	SEX1	USECOND
8	1	4	2	0
6	1	4	2	0
4	0.9724	4	2	0
13	1	5	1	0
11	0.9743	3	2	0
13	0.9932	3	2	0
10	1	4	2	0
12	0.7693	4	1	0
12	0.9923	4	2	0
12	0.9316	4	2	0
10	0.9995	4	1	0
10	1	6	2	0
12	1	4	2	0
11	0.9797	5	2	1
10	1	6	2	1
4	0.9993	4	2	1
10	1	2	1	1
11	1	4	1	1
10	1	2	2	1
4	0.9879	6	1	1
11	0.5822	6	1	1
11	1	4	2	1
12	1	4	2	1
4	0.9998	4	2	1
11	0.9901	6	2	1
11	1	6	2	1
4	1	4	1	1

Figure 17: Sample of output table ANALYSE_CLUSTER_1

On analyzing this table after sorting the table according to the attribute **use_condom,** I observe that the people who used condoms (signified by use_comdom=1) are fewer than those who do not. **Therefore it becomes difficult to draw conclusions on what attributes influence condom use. This is mainly due to the small dataset or sample size.**

However from these results it is observed that education level has very little influence to condom use. The distribution shows that even though people have at least been to high school some do to make use of condoms while the others do not. Therefore this table was discarded, had little information thus resulting in the mining of a new table.

2. Table **ANALYSE_CLUSTER_2**

This table also has the attribute **use_condom** but now includes attributes **EDUC1 (education level), EMPLOY** (if employed or not)**, HIV_TST** (if person would have an HIV TEST) and **HIVTEST** (if person has had an HIV TEST). In this table semi-clear relations were established, it was apparent that all those people who had an HIV TEST did make use of a condom. This indicates that if one had an HIV Test (although we don't know the outcome of test results) it is apparent that that person eventually makes use of a condom when having sexual intercourse.

CLUSTER_ID	PROBABILITY	EDUC1	EMPLOY	HIV_TST	HIVTEST	USECOND
4	1	4	4	2	0	0
4	1	4	1	4	1	1
4	1	4	1	3	0	0
4	1	4	2	4	0	0
4	0.9979	4	2	4	0	0
4	1	4	2	1	0	0
4	1	4	1	1	1	0
4	0.999	4	1	5	1	1
4	0.9953	4	2	2	1	1
4	0.9998	4	4	2	1	1
4	1	4	2	3	0	0
4	1	4	1	4	0	1
4	0.9417	4	1	2	0	0
4	0.9993	4	2	2	0	1
4	1	5	1	2	1	0
4	1	6	3	2	1	0
4	0.996	4	2	2	1	1
4	0.9984	4	1	1	1	1
4	1	6	1	3	1	0
4	0.9785	4	1	2	0	0
4	1	4	2	2	0	0
4	0.9999	4	1	2	0	0
4	0.9928	4	1	2	0	0
4	0.9724	4	2	5	1	1
4	1	4	1	4	0	0
4	1	6	1	1	1	1
4	1	4	2	1	0	0
4	0.9983	4	2	2	0	0
4	1	4	1	2	1	1
4	0.7599	4	2	3	1	0
4	1	4	2	5	0	1
4	0.7501	2	1	2	1	0
4	1	4	2	1	0	0
4	0.9537	4	1	2	1	1
4	1	5	4	2	0	0
4	0.9975	4	6	2	0	0
4	1	4	6	4	1	1
4	0.9924	5	1	3	1	0

Figure 18: Sample of output table ANALYSE_CLUSTER_2

Therefore, the attribute HIVTEST in this case influences condom use, hence is a predictor of prevention behaviour.

3. Table **ANALYSE_CLUSTER_3**

In this table I am now looking for attributes that influence one to abstain from having sexual intercourse for a year or more. I therefore included the following attributes in the resultant table: **ABSTNMTH** (abstain for a month), **ABSTNYR** (abstain for a year or more), **FAITHFL** (faithful to partner), and **HIVTEST** (have an HIV test).

CLUSTER_ID	PROBABILITY	ABSTNMTH	ABSTNYR	FAITHFL	HIVTEST
8	1	1	0	0	1
8	1	2	0	0	1
6	1	2	1	0	0
6	1	1	1	0	0
6	1	2	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	2	0	0	1
6	0.9086	2	0	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	3	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	1	1	0	0
6	1	2	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	2	1	0	0
6	1	4	1	0	0
6	1	2	1	0	0
6	1	1	1	0	0
6	1	1	0	1	0
6	1	1	1	0	0
6	1	2	1	0	0

Figure 19: Sample of output table ANALYSE_CLUSTER_3

On analysis it is observed that ClusterID = 6 is the one cluster that contains the majority of people who managed to abstain for a year or more. However, it was also clear that of those who managed to abstain none of them could be faithful and none had been tested for HIV. This clearly shows that testing for HIV has no influence in resulting one to abstaining (although it has an influence on condom use as explained in **ANALYSE_CLUSTER_2)**.

4. Table **ANALYSE_CLUSTER_4**

EMPLOY (employment status), **HIV_TST** (if person would go for HIV test), KNOWAIDS (if person knows AIDS), **KNOWHIV** (if person knows HIV), **LESPART** (if person has decided to have fewer partners) and **TALK_OPN** (if person talks openly about the virus) were attributes included in this table.

CLUSTER_ID	PROBABILITY	EMPLOY	HIV_TST	KNOWAIDS	KNOWHIV	LESPART	TALK_OPN
12	0.9707	2	4	0	0	0	5
10	1	1	1	1	0	0	5
11	0.9986	2	3	0	0	0	5
4	1	1	4	0	0	0	5
10	0.9988	2	1	1	0	0	5
11	0.6431	1	2	0	0	0	1
11	1	1	1	0	0	0	2
6	1	5	1	1	0	0	5
6	1	1	4	0	0	0	3
6	1	1	4	0	0	0	5
4	0.9417	1	2	1	0	0	1
11	1	3	2	1	0	0	5
12	1	0	3	0	0	0	5
10	1	1	3	1	0	0	5
11	1	5	2	1	0	0	2
8	0.9402	1	5	1	0	0	2
10	1	1	1	0	0	1	1
12	0.9999	2	2	1	0	1	1
12	0.9394	1	1	0	0	1	1
11	0.5822	2	2	0	1	1	1
11	1	1	3	0	0	1	2
6	1	4	1	0	0	1	3
13	1	4	4	0	0	1	1
11	1	1	1	1	0	1	1
8	1	1	2	1	0	1	2
10	1	1	1	1	0	1	3
4	0.999	1	5	1	0	1	5
6	1	2	5	0	0	1	1
11	0.9901	1	1	0	0	1	1
13	0.904	4	1	1	0	1	5
11	1	4	1	1	0	1	1
10	1	1	2	1	0	1	5
6	1	2	1	0	0	1	1
6	0.9086	1	2	1	0	1	2
11	1	4	2	1	0	1	5
12	0.9582	2	2	1	0	1	2
11	0.9998	1	1	0	0	1	5
11	1	3	1	1	1	1	1

Figure 20: Sample of output table ANALYSE_CLUSTER_4

Here focus is on the attribute **LESPART** from my definition of prevention predictors: Observations in this table show that these results are not really useful. This is because 277 out of 299 people in this table did not decide in having fewer partners thus leaving only 20 people out of 299 who decided to have fewer partners. However from the 20 only 2 seem to know about AIDS. Thus no conclusions can be draw due to small sample size.

5. Table **ANALYSE_CLUSTER_5**

Attributes included in this table are **ABSTNYR** (abstain for a year or more), KNOWAIDS (person knows AIDS), **KNOWHIV** (person knows HIV) and **MARRIED** (marital status).

CLUSTER_ID	PROBABILITY	ABSTNYR	KNOWAIDS	KNOWHIV	MARRIED
4	1	0	0	0	1
4	1	0	0	0	1
4	0.8358	0	1	0	1
4	0.972	0	1	0	1
4	0.9934	0	1	0	0
6	1	0	1	0	1
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	1	0	0	1
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	0.9999	1	0	0	1
6	1	1	0	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	1	0	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	1	0	0	0
6	1	1	1	0	0
6	1	1	1	0	0
6	1	1	0	0	0
6	1	1	0	0	0
6	1	1	1	0	0
6	1	1	0	1	0
6	1	1	1	1	0
6	1	0	0	0	1
6	1	0	0	0	0
6	1	1	1	0	0

Figure 21: Sample of output table ANALYSE_CLUSTER_5

Attribute in to focus on is **ABSTNYR**: Observations here draw an interesting pattern from the results. It is evident that ClusterID = 6 is the only cluster that contained people who have managed to abstain for a year or more. And from this same cluster it is evident that the majority of the people know about AIDS. This alone brings me to the conclusion that one abstains because he knows AIDS (the dangers and effects of it). ***Therefore I conclude that Knowing about AIDS somehow leads to abstinence.***

6. Table **ANALYSE_CLUSTER_6**

From the results obtained in previous table I then decided to continue investigating the attribute ABSTNYR (abstain for a year or more) hence included different attributes to see what outcome I get if I keep in mind that ClusterID = 6 contains only people who managed to abstain for a year or more. These other attributes are **AGE1** (persons age), **HIVTEST** (if been tested for HIV), **MARRIED** (marital status), **SEX1** (gender) and **YTHHELP** (if person has volunteered to help care for people who have been infected with the virus).

CLUSTER_ID	PROBABILI...	ABSTNYR	AGE1	HIVTEST	MARRIED	SEX1	YTHHELP
6	1	1	23	0	0	1	3
6	1	1	20	0	0	2	4
6	1	1	18	0	0	1	0
6	1	1	19	0	0	1	2
6	1	1	17	0	0	1	1
6	1	1	24	1	0	2	1
6	1	1	18	0	0	1	2
6	1	1	20	0	0	1	3
6	1	1	20	0	0	2	1
6	1	1	18	0	0	2	0
6	1	1	21	0	1	1	1
6	1	1	17	0	0	1	2
6	0.9086	0	21	0	0	2	3
6	1	1	19	0	0	2	2
6	1	0	19	0	1	1	2
6	1	1	19	0	0	2	2
6	1	1	24	0	0	1	2
6	1	1	18	0	0	1	2
6	1	1	23	0	0	2	1
6	1	1	22	0	0	1	1
6	1	1	17	0	0	2	1
6	1	1	24	0	0	2	2
6	1	1	19	0	0	2	5
6	1	0	20	0	1	1	2
6	1	0	19	1	0	2	4
6	1	1	18	0	0	1	2
6	1	1	19	0	0	1	1
6	1	1	18	0	0	1	1
6	1	1	18	0	0	2	5
6	1	1	24	0	0	1	1
6	1	1	20	0	0	2	1
6	1	1	20	0	0	2	1
6	1	1	23	0	0	1	2
6	1	0	18	0	0	2	1
6	1	1	19	0	0	2	1
6	1	1	19	0	0	2	1
6	1	1	18	0	0	2	1
6	1	1	18	0	0	2	1

Figure 22: Sample of output table ANALYSE_CLUSTER_6

Since the attribute in focus here is **ABSTNYR**: Observations show that the people who abstained (mainly in ClusterID = 6) were not influenced by having an HIV test, by their marital status or gender. This therefore brings a conclusion that other attributes not included in this table may have influenced these people to abstain.

7. Table **ANALYSE_CLUSTER_7**

For this table I decided to focus on the use of condoms again by including the following attributes**: USECOND, EDUC1** and **SEX_YET** (if person has had sex before or not). From this table I was wishing to establish a relationship between condom use and education level, but it has proven impossible to determine whether literacy (education level) plays a role in why one may use a condom. These results are thus not really clear to draw any conclusions.

CLUSTER_ID	PROBABILITY	EDUC1	SEX_YET	USECOND
11	0.4978	5	0	0
11	1	5	0	0
6	1	4	0	0
11	0.9998	4	1	0
12	1	5	0	0
13	0.5823	3	1	0
4	1	4	1	0
4	0.972	3	1	0
8	0.9884	4	1	0
6	1	6	0	0
10	1	2	1	1
11	1	4	1	1
4	0.9879	6	1	1
11	1	4	1	1
11	1	5	0	1
10	0.9997	6	0	1
11	0.9901	6	0	1
11	1	6	0	1
4	0.9953	4	1	1
11	0.9797	5	1	1
11	0.5822	6	1	1
11	1	4	1	1
12	1	4	1	1
4	1	0	1	1
4	1	4	1	1
4	0.9998	4	0	1
4	1	4	1	1
10	0.9988	6	1	1
4	0.9993	4	0	1
10	0.9991	5	1	1
11	1	4	1	1
10	1	6	1	1
11	0.9998	4	0	1
4	1	4	0	1
10	1	4	1	1
10	0.9999	4	1	1
10	1	2	1	1
4	0.9934	5	0	1

Figure 23: Sample of output table ANALYSE_CLUSTER_7

8. Table **ANALYSE_CLUSTER_8**

This table has the following attributes **LESPART** (fewer partners), **KNOWAIDS** and **KNOWHIV**. Here I turn my attention to the attribute fewer partners, where I wish to determine if knowing both HIV and AIDS leads to one resulting in having fewer partners. The problem here though is that the sample size of those people who decided (for whatever reason) to have fewer partners is very small (23 out of 299). Therefore it is difficult to draw conclusions here.

CLUSTER_ID	PROBABILITY	KNOWAIDS	KNOWHIV	LESPART
6	1	0	0	0
4	1	1	0	0
10	0.9998	0	0	0
11	0.9983	0	0	0
4	1	0	0	0
11	1	1	0	0
6	1	0	0	0
12	1	0	0	0
13	0.5823	1	0	0
4	1	1	0	0
4	0.972	1	0	0
10	1	1	0	0
12	0.9707	0	0	0
10	1	0	0	0
4	1	0	0	0
10	1	0	0	0
11	1	1	0	1
12	0.9999	1	0	1
12	0.9394	0	0	1
8	1	1	0	1
13	1	0	0	1
11	0.9901	0	0	1
13	0.904	1	0	1
11	1	1	0	1
11	0.5822	0	1	1
10	1	1	0	1
10	1	1	0	1
6	1	0	0	1
12	0.7693	0	0	1
10	1	0	0	1
11	1	0	0	1
6	1	0	0	1
11	1	1	0	1
4	0.999	1	0	1
6	1	0	0	1
6	0.9086	1	0	1
11	0.9998	0	0	1
11	1	1	1	1

Figure 24: Sample of output table ANALYSE_CLUSTER_8

9. Table **ANALYSE_CLUSTER_9**

From table **ANALYSE_CLUSTER_5**, I observed that the ClusterID = 6 is different from all other clusters in that it only contains people who have managed to abstain. So in this table (**ANALYSE_CLUSTER_9**) I intend to further find other attributes that have closely influenced an individual to be able to abstain for a year or more. To do this I included the following attributes together with **ABSTNYR** (abstain for a year or more), and these are **EDUC1** (education level), **EMPLOY** (employment status), **FAITHFL** (faithful to partner), **HOUS_TYP** (your house or home type), **KNOWAIDS, KNOWHIV**, **MARRIED, SEX1** (gender) and **YTHHELP**.

CLUSTER...	PROBABIL...	ABSTNYR	EDUC1	EMPLOY	FAITHFL	HOUS_TYP	KNOWAIDS	KNOWHIV	MARRIED	SEX1	YTHHELP
6	1	1	4	1	0	3	0	0	0	2	3
6	1	1	4	1	0	3	0	0	0	2	3
6	1	1	4	2	0	3	0	0	0	1	3
6	1	1	4	2	0	1	0	0	0	2	1
6	1	1	4	2	0	1	0	0	0	1	2
6	1	1	6	5	0	1	1	0	0	2	2
6	1	1	5	1	0	2	0	0	0	2	4
6	1	1	6	6	0	1	1	0	0	1	1
6	1	1	4	2	0	3	1	0	0	1	2
6	1	1	4	1	0	2	0	0	0	2	1
6	1	1	4	1	0	3	0	0	0	1	1
6	1	1	3	1	0	5	0	1	0	2	5
6	1	0	5	1	0	1	0	0	1	1	2
6	1	1	4	1	0	4	0	0	0	2	5
6	1	1	6	2	0	1	0	0	0	1	2
6	1	1	4	1	0	1	0	0	0	1	1
6	1	1	4	1	0	1	0	1	0	2	2
6	1	1	4	2	0	1	1	0	0	1	2
6	1	1	6	2	0	1	0	0	0	2	1
6	1	0	4	2	1	1	0	0	0	2	1
6	1	1	4	2	0	4	1	0	0	2	1
6	1	1	4	1	0	3	0	0	0	2	1
6	1	1	4	2	0	1	0	0	0	2	1
6	1	1	4	2	0	3	0	0	0	1	2
6	1	1	5	1	0	1	1	0	0	2	1
6	1	1	5	1	0	1	1	0	0	2	1
6	0.9086	0	4	1	0	1	1	0	0	2	3
6	1	1	5	3	0	3	1	0	0	1	3
6	1	1	6	2	0	1	0	0	0	2	1
6	1	1	5	1	0	1	0	0	0	2	2
6	1	1	3	2	0	2	1	0	0	1	3
6	0.9999	1	4	2	0	3	0	0	1	2	1
6	1	1	1	1	0	2	0	0	0	2	2
6	1	1	4	2	0	1	0	0	0	1	1
6	1	1	4	2	0	1	0	0	0	1	1
6	1	1	6	4	0	1	0	0	0	2	1
6	1	1	4	2	0	3	0	0	0	1	2
6	1	1	4	2	0	1	1	0	0	1	1

Figure 24: Sample of output table ANALYSE_CLUSTER_9

On analyzing this table, I observed that the majority of the people who abstained (see ClusterID = 6) have at least been to high school. However this cannot be conclusive as it is evident that for other clusters (containing people who failed to abstain) had a similar

distribution. It is also clear that gender and the other attributes included here did not make significance to why people abstain.

6.1.1 Conclusions drawn from the Analyse Cluster tables

Observations in table ANALYSE_CLUSTER_2 make it evident that the attribute HIVTEST influences condom use. However, ANALYSE_CLUSTER_5 clearly concludes that knowing about AIDS leads to abstinence. Therefore from these table observations I have concluded that the attributes **HIVTEST** and **KNOWAIDS** have been clearly been identified as predictors of prevention behaviour.

6.2 Determining Predictors by using Association Rules

According to [Al-Attar, 2004], association rules are similar to decision trees and association rule induction is the most established and effective of the current data mining technologies. This technique involves the definition of a business goal and the use of rule induction to generate patterns relating this goal to other data fields. The patterns are generated as trees with splits on data fields. This technique allows the user to add their domain knowledge to the process and decide on attributes for generating splits [Han et al, 2001].

In order to employ the association rules, I simply applied the Apriori algorithm which is the only association rule available in ODM to the datasets TSHA_TSHA_APPLY1 and APPLY_OC3_TSHATSHA. The resulting tables from the algorithm sifting through these datasets were named ASSOCIATION_MODEL for dataset APPLY_OC3_TSHATSHA and ASSOCIATION_MODEL2 for dataset TSHA_TSHA_APPLY1 and these were exported to spreadsheet files for analysis. The spreadsheets were also named as ASSOCIATION_MODEL and ASSOCIATION_MODEL2 and these are available on the CD ROM that accompanies the project.

The dataset APPLY_OC3_TSHATSHA was used to find rules because it the resultant data that contains the ClusterIDs found using the O-Cluster model. Analysing the rules in ASSOCIATION_MODEL I observed the following.

The majority of relationships here lead to people not having less partners, not using condoms, being unfaithful to partners in an event to prevent HIV infection. Most of the relationships are however meaningless.

From the table it is however evident that ASSOCIATION_MODEL is giving relationships of attribute behaviour from a negative perspective (for example, if one is married and talks openly about HIV AIDS then they will not be able to abstain for a year). This is making the identification of predictors difficult. By interpreting these relationships, I observe that even though individuals may talk openly about HIV AIDS and married, this will not result them in having less partners or abstain for a year. Therefore, in this association table it is difficult to identify predictors.

The dataset TSHA_TSHA_APPLY1 is the dataset that was used when applying the cluster models. Analysing the outcome of this dataset (outcome is in ASSOCIATION_MODEL2) I observe the following; an individual who talks openly about HIV AIDS and hasn't had sex before is most likely to use a condom when he/she decides to have sex. Table results also show that if one had an HIV Test and had sex before, he/she is most likely to abstain for a year. The majority of the association rules here show that one will have fewer partners, if one has at least been to high school, has had an HIV Test and talks openly about HIV.

Here ASSOCIATION_MODEL2 is giving relationships from a positive perspective, for example, an individual who talks openly about HIV AIDS and hasn't had sex before is most likely to use a condom when he/she decides to have sex. .Therefore for this table it is evident that the most likely predictors are having an *HIV test* and *talking openly* about HIV AIDS.

It must be mentioned that the identification of HIV AIDS prevention predictors is a very difficult process which needs reasoning and domain knowledge. From the associations the attributes *HIV test* and *talk openly* have been identified as predictors for condom use as prevention. Although, the justification of these predictors is mainly based on the clusters and association rules it must be mentioned that the interpretation of the relationships in the associations is solely dependent on domain knowledge of the field.

6.3 Chapter Summary

Analysis of the Clusters found was made in a trial and error approach. However it was evident that there was a significant distinction between some clusters, namely ClusterID=4 and ClusterID = 6. These two clusters made it possible to identify predictors. In the cluster analysis the attributes HIVTEST and KNOWAIDS have been clearly been identified as predictors of prevention behaviour of condom use and abstinence.

By employing an association rule algorithm, the Apriori, the attributes *HIV test* and *talk openly* were identified as HIV AIDS predictors of prevention of condom use.

The following chapter will summarise the conclusions that have be drawn from all the results obtained. It also gives a conclusion regarding Oracle Data Miner as a mining tool.

CHAPTER 7

7 Summary and Conclusions

In this chapter, the main results of this work are summarised. This project has been a very successful one, with the major project aims and goals met. There were two main objectives in this research. Firstly, it was aimed at analysing and evaluating the effectiveness of two Clustering algorithms, the O-Cluster and K-Means in terms of accuracy in building models and finding clusters when applied to new data. The second objective was to find predictors of HIV AIDS prevention behaviour attributes from the dataset obtained from the Centre for AIDS Development, Research and Evaluation Institute for Social and Economic Research, Rhodes University.

7.1 Key Results of the Work

7.1.1 Conclusions Regarding Algorithm and Model Accuracy

[Roiger et al, 2003] discusses aspects of evaluating the performance of models built during data mining. According to these authors, when evaluating performance it is necessary to consider whether the results of the data mining can be interpreted and whether the results can be used with confidence. In regard to this it was observed that the O-Cluster algorithm builds more accurate models as compared to the K-Means algorithm. The outcome of these accuracies is depicted by average confidence and support values in Figure 12 and Figure 13.

The most effective model built was the model built by using the O-cluster algorithm as it possesses an accuracy figure of 95,5% (Figure 13). On applying this model to new data and using the evaluation technique by [Roiger et al, 2003] to analyse the accuracy the algorithm model has in finding clusters, it was evident that 42 out of 107 (39%) ClusterIDs were correctly predicted in the outcome of the Adaptive Bayes Network

model. This shows that the model built using the O-Cluster algorithm performs better than the model built by the K-Means algorithm which only managed to correctly predict 18 out of 107 (17%) ClusterIDs.

In order for an algorithm to perform well, it was necessary to tune the algorithm by adjusting its settings. According to the model results obtained from the algorithm evaluation phase, it is evident that tuning in some cases results in the algorithm performing well and sometimes performing badly in terms of model accuracy.

Observations indicated that tuning the K-Means algorithm had very little effect on increasing the accuracy of the models built by the algorithm. On the other hand, tuning the O-Cluster algorithm proved to make the algorithm perform better thus building the most accurate model as depicted in Figure 12 and 13. The setting to enable tuning for the O-Cluster is the Sensitivity while those for the K-Means are the Minimum error tolerance and Maximum Iterations.

7.1.2 Conclusions Regarding Information Gathered from Dataset

A few attribute predictors of HIV AIDS prevention were found from the dataset. The process of finding these involved critical analysis of the results as well as good reasoning. These were the attributes *HIV test* and *Know Aids* have been clearly been identified as predictors of prevention behaviour of condom use and abstinence. By employing the Apriori algorithm, the attributes *HIV test* and *talk openly* were identified as HIV AIDS predictors of prevention of condom use.

- *HIV test – if one has had an HIV test*
- *Know Aids – if one knows about AIDS*
- *Talk openly – if one talks openly about HIV AIDS or not*

7.2 Oracle Data Miner and Available Algorithms

The fact that Oracle Data Mining's data mining functionality is embedded inside the Oracle Database makes the data gathering and preparation process simpler. It also makes model building and model applying very simple as this is evident from the step by step wizards provided by the mining tool. ODM provides easy and reliable access to the database and the tables stored in the database. This makes it possible to search for a specific data set during the data mining process. This also allows the user to view summaries of the data including distributions of attributes in the data set, which is of use during the data preparation phase. Any results obtained can also be exported to spreadsheets allowing increased accessibility and ensuring they can be easily worked with.

ODM supports the following algorithms as stated by [Berger, 2004] and the evaluation results of some of these algorithms investigated seems to be interesting.

- Enhanced k-Means Clustering (clustering)
- Orthogonal Partitioning (clustering)
- Adaptive Bayes Network supporting decision trees (classification)
- Naive Bayes (classification)
- Model Seeker (classification)
- Apriori (association rules)

The results obtained in this research when making an evaluation of the two Clustering algorithms available in ODM have shown that the O-Cluster algorithm builds more accurate models as well as in finding more accurate clusters when the model is applied to new data as compared to the K-Means. On the other hand, [Davis, 2004] also proves in her research that for the evaluation of Classification algorithms between the Adaptive Bayes Network and the Naive Bayes, the ABN algorithm gives more accurate results when used to build models as well as in applying the models to new data.

It must be emphasised that the ease and speed of building a model using the wizards allows for a number of models to be built at the same time. This approach is recommended in order to ensure the most effective model possible is produced.

7.3 Conclusion

Following on from the conclusions and recommendations of the theory research, I have managed to conclude that the O-Cluster algorithm has been identified as the most accurate clustering algorithm in Oracle data mining 10g. The results achieved by the evaluation fulfil the aims of this project. However considerable time was spent laying the groundwork for the practical stages were different proposals were considered as to how to implement theory suggestions in a practical way. Eventually, a novel solution was adapted and hence leads to these results.

REFERENCES:

[Berger, 2004] *Oracle Data Mining (Know More, Do More, Spend Less) - An Oracle White Paper, October 2004* by Berger,C.URL:http://www.oracle.com/technology/products/bi/odm/pdf/bwp_db_odm_10gr1_1004.pdf, Accessed: 14 April 2005

[Berry et al, 2000] *Mastering Data Mining: The Art and Science of Customer Relationship Management* by Michael J.A. Berry and Gordon S. Linoff, USA, Wiley Computer Publishing, 2000

[Bradley et al, 1998] *Refining Initial Points for K-Means Clustering* by P. Bradley and U. Fayyad: ICML 1998

[Davis, 2004] *An Evaluation of Commercial Data Mining* by Emily Davis. Oracle Data Mining Honours Research Project 2004. URL: http://www.cs.ru.ac.za/research/previous/g01d1801/ Accessed: 14 September 2005

[Elder et al, 1998] *Fourth International Conference on Knowledge Discovery & Data Mining Research* - Friday, August 28, 1998 by John F. Elder IV & Dean W. Abbott Elder, http://www.datamininglab.com Accessed: 11 September 2005

[Han et al, 2001] *Data mining: concepts and techniques* by Jiawei Han and Micheline Kamber, San Francisco, California, Morgan Kauffmann, 2001.

[Insightful Miner, 2003] *Insightful Corporation Seattle, Washington, Insightful Miner 3.0 User Guide*: http://www.insightful.com/support/iminer30/getstart.pdf, June 2003- Accessed: 11 October 2005

[Mannila et al, 2001] *Principles of data mining*, by David Hand, Heikki Mannila and Padhraic Smyth, , Cambridge Massachusetts, MIT Press, 2001.

[Oracle, 2005] *The Oracle Home Page.* Revised February 2005. Oracle 10g Data Mining FAQ http://www.oracle.com/technology/products/bi/odm/odm_10g_faq.html#api Accessed: 17 May 2005

[Oracle Installation, 2004] Oracle Database Installation Guide 10*g* Release 1 (10.1.0.2.0) for 64-bit Windows- http://download-east.oracle.com/docs/html/B13805_02/toc.htm Accessed: 2 July 2005

[ODM Release Notes, 2004] *Oracle Data Miner 10.1Release Notes and Installation Instructions November 2004.* http://www.oracle.com/technology/products/bi/odm/odminer_install.htm Accessed: 2 August 2005

[ODM tutorial, 2004] *Oracle Data Miner Tutorial Part 1 & Part 2, 1 September, 2004.* http://www.oracle.com/technology/products/bi/odm Accessed: 17 August 2005

[Roiger et al, 2003] *Data mining: a tutorial- based primer* by Richard J. Roiger and Michael W. Geatz, Boston, Massachusetts, Addison Wesley, 2003, pg 58 and 232.

Appendix A:

Installation Problems Encountered in ODM

Oracle10g version 10.1.0.2.0 and Oracle Data miner 10g version 10.1.0.2.0 were installed on the server machine – Athena.ict.ru.ac.za using the installation guide from [ODM Release Notes, 2004]

Two errors have been obtained as a result of this installation guide with the one discussed first being the major problem.

Since the Oracle database 10g and ODM are both version 10.1.0.2, this installation guide strongly recommends that you upgrade to the following patch sets:

- Oracle database 10.1.0.3 and
- Oracle Data Mining 10.1.0.3.1

On installing and configuring all the required software as described in this installation guide the following errors appear when logging into the data mining server:

- **odm api is incompatible with odm api for the schema and the required is odm api 10.1.0.2.0**

Figure 25: Error obtained after installations

Discussion of Problem and Solution:

Installing the above patches was hoped to upgrade the Oracle database from version 10.1.0.2.0 to version 10.1.0.3 but that was not the case. The patches would not upgrade the database to the higher version as required while the ODM does upgraded to version 10.1.0.3.1. When upgrading, the `odmapi.jar` used by Oracle Data Miner, must be the same version as that installed in the database. So the patching of ODM 10.1.0.3 results in the loading of a new `odmapi.jar` into the database. The result of the patching makes the database and ODM versions to be incompatible with each other.

The database version can be obtained by logging into the database using the Enterprise manager then run the following sql query from an application iSQL plus provided within the enterprise manager (iSQL plus is similar to SQL plus, difference is that iSQLplus runs on a web browser while SQLplus is an application). Refer to Appendix C for instruction on how to logon to the enterprise manager on the machine Athena.

```
->SQL> select COMP_ID, COMP_NAME, VERSION, STATUS from dba_registry
where COMP_ID = 'ODM'
returns:

COMP_ID
------------------------------
COMP_NAME
----------------------------------------------------------------
VERSION                        STATUS
------------------------------ -----------
ODM
Oracle Data Mining
10.1.0.2.0                     VALID
```

The solution is to de-install the patches and replace the original ODM API in both the ODM and Oracle database. The original **odmapi.jar** for this database is for the version 10.1.0.2.0. This api can be obtained from the Oracle website or by the reinstallation of the Oracle 10g database version 10.1.0.2.0 (and can be found in the database lib directory: C:\oracle\product\10.1.0\em_2\dm\lib in this database I installed on athena). The same

version of the odmapi.jar has to be placed in the Oracle data miner lib directory (C:\odminer\lib).

I therefore recommend that when installing and configuring the database and the data miner not to install or patch with the Oracle database 10.1.0.3 and Oracle Data Mining 10.1.0.3.1 patches because this results in the error discussed above.

Another error that I encounter was the same as that depicted by Figure.15. The error will be obtained if the Oracle database is not properly configured for data mining. Here a data mining user account in the database is configured so that it can be linked to and used by the Oracle data miner. The proper way to configure the database for data mining is discussed in detail in Appendix B, were I provide all the necessary steps to installing and configuring the database and Oracle data miner for data mining.

Appendix B:

Configuring ODM

This configuration is the Oracle 10g database version 10.1.0.2 and Oracle data miner version 10.1.0.2 on a Windows Platform (Microsoft Windows Server 2003, Standard Edition). Basically the Oracle Data Mining server runs against an Oracle 10g Database. Both the database and ODM can be downloaded from the Oracle website.

STEP 1-*Install Database*

For the installation of the Oracle 10g database refer to the Oracle Database 10*g* Release 1 (10.1) Documentation [Oracle Installation, 2004].

It is recommended to install the Enterprise version of the database as it contains all the necessary all the data mining tools needed by the Oracle data miner software.

After a successful installation, all the ODM software is located in the directory:
C:\oracle\product\10.1.0\em_2\dm

During installation of the database the passwords for all default users was set and in this case it was set to **oracle10g** for all users. The username for super user is **sys** and the corresponding password is **oracle10g**. To log into the database you use the Enterprise manager (EM). The EM for oracle10g is accessed through a web browser and the URL for this particular database is URL:http://athena.ict.ru.ac.za:5500/em/

Enter the credentials, username-**sys**, and password-**oracle10g** then Connect as-SYSDBA

When I logged onto the database I then create my own data mining user account:
Username-**odm_use** AND password-**datamine** then Connect As normal user.

STEP 2-*Install Data miner*

After downloading the data miner version 10.1.0.2.0 (odminer.zip file) unzip the entire contents of odminer.zip to the directory C:\odminer.

To run the Data miner, you run (double click) the odminerw.exe executable file located in the directory where Oracle Data Miner was installed. In this case it can be found in the directory; **C:\odminer\bin\odminerw.exe**

STEP 3-*Configure Data miner user*

To configure the database for a data miner user the script **odmuser.sql** is executed in SQL*Plus which is located in located in the directory containing all the Data mining software (**C:\oracle\product\10.1.0\em_2\dm\admin**). Therefore to be able to run the script you need to logon to the database as a super user (in this case as sys).

Start SQL*Plus and login as follows:

>**SQLPLUS sys/oracle10g as sysdba**

Run the script **odmuser.sql** with the following command:

SQL> @ C:\oracle\product\10.1.0\em_2\dm\admin\odmuser.sql

The output is as follows:

```
============================================================
Enter value for 1: odm_use   // data mining user account we are giving mining permissions in the database
Enter value for 2: datamine    //corresponding password
Enter value for 3: odm          // default tablespace for this user
old   1: drop user &USERNAME cascade
new   1: drop user odm_use cascade
drop user odm_use cascade
*
old   1: create user &USERNAME identified by &PASSWORD default tablespace
&TBSNAME temporary tablesp
new   1: create user odm_use identified by datamine default tablespace odm temporary
tablespace temp
create user odm_use identified by datamine default tablespace odm temporary tablespace
temp quota un
 *
old   9: to &USERNAME
new   9: to odm_use
```

Grant succeeded.

old 1: grant execute on ctxsys.ctx_ddl to &USERNAME
new 1: grant execute on ctxsys.ctx_ddl to odm_use

Grant succeeded.

old 1: grant DMUSER_ROLE to &USERNAME
new 1: grant DMUSER_ROLE to odm_use

Grant succeeded.

STEP 4-*load ODM sample datasets*

This step is option if you have your own datasets to perform the data mining, but it becomes necessary if one needs to follows the data mining tutorial which will be discussed in Appendix C. In is also important to note that all data mining is performed by the data mining user granted permission as shown above and all the data that is intended to be used during mining is loaded into that user's table space. The database only grants one data mining user.

To load the sample datasets execute the following scripts:

First run the *odmtbs.sql* script to create a tablespace and specify the location of the tablespace file to be used by the data mining users.

- SQL>C:\oracle\product\10.1.0\em_2\dm\admin\odmtbs.sql TEMP01.DBF C:\oracle\product\10.1.0\oradata\orc2\ TEMP01.DBF

Run the *dmuserld.sql* script to load the sample data mining datasets into the specified Data Mining User schema as follows:

- SQL> C:\oracle\product\10.1.0\em_2\dm\admin\dmuserld.sql odm_use datamine C:\oracle\product\10.1.0\oradata\orc2\ TEMP01.DBF

The ODM installation guide specifies that for Oracle10, ODM Java and PL/SQL sample programs use datasets shipped with the common schema SH. Therefore it is necessary to execute the following script to grant necessary SH access privileges. As mentioned all default database user password was set to oracle10g. To grant these privileges run the script *dmshgrants.sql* as follows:

- SQL> @ C:\oracle\product\10.1.0\em_2\dm\admin\dmshgrants.sql oracle10g odm_user

Then logout of SQL*Plus as superuser

- SQL> exit

Re-Login as the Data mining user (odm_use)

- sqlplus odm_user/datamine

Execute the script *dmsh.sql* to create related views and tables in the SH Schema.

- SQL> C:\oracle\product\10.1.0\em_2\dm\admin\dmsh.sql

STEP 5-*launching the Oracle data miner*

Double click the executable→ ***odminerw.exe***

Since this is the first time you are launching ODM Data Miner (odminerw.exe from the directory we installed the data miner: ***C:\odminer\bin\odminerw.exe***), you will be prompted to create a new database connection. At this point simply enter the information for the user ***odm_use*** as follows:

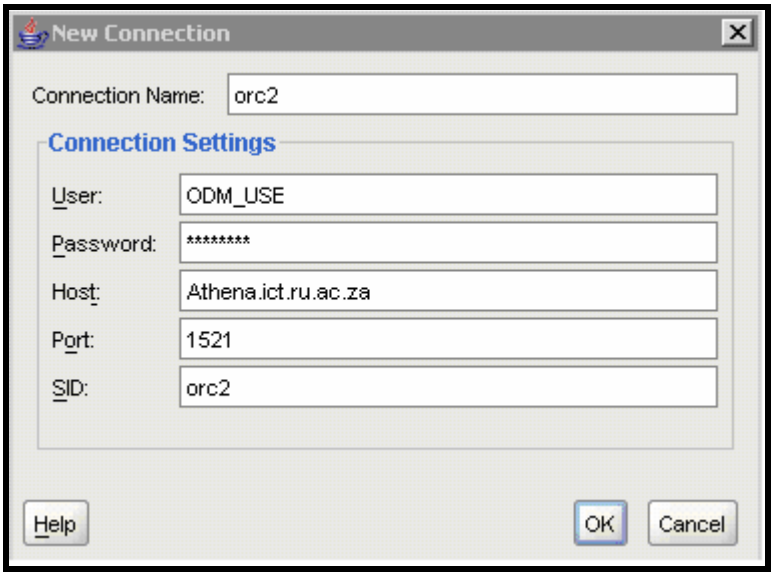


Figure26: Connection settings for odm_use

Note that the *Connection name* can be anything but in this case I set it to orc2 (the database name)

NB: if all the above installation and configuration steps are implemented successfully, the following screens should be seen when logging on to the ODM server.

Figure 27: Select orc2 connection name

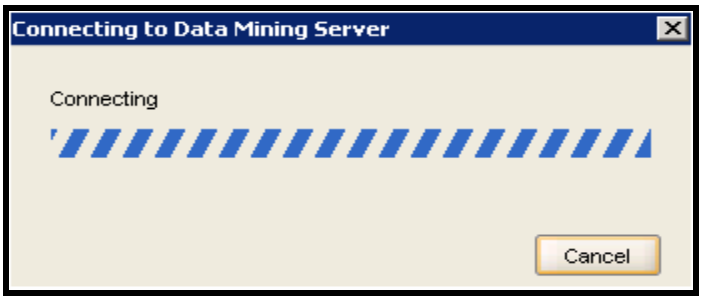


Figure 28: connecting to the ODM server

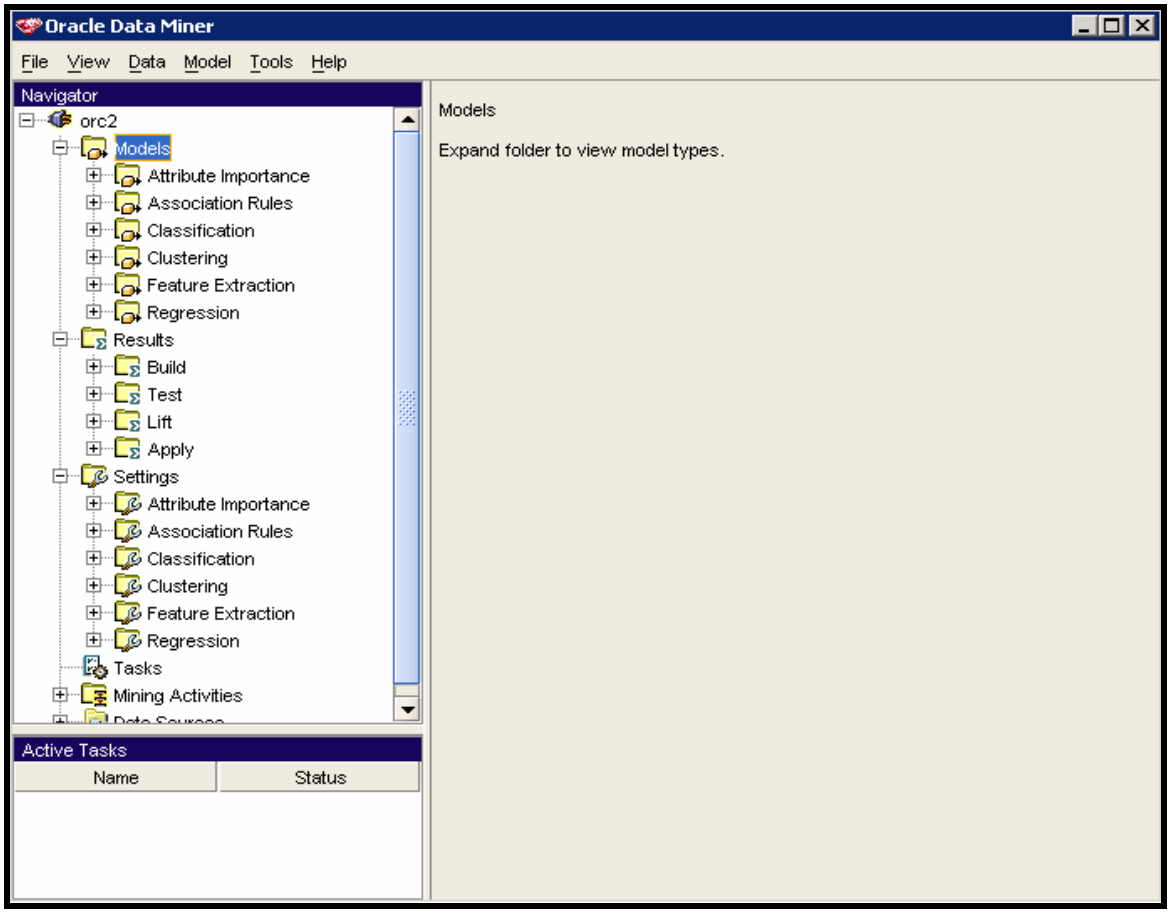


Figure 29: ODM server interface when connected

Appendix C:

ODM Tutorials

This appendix aims to provide the reader with information on how to use the Oracle data miner used in the evaluation of the algorithms in this research project. I assume the reader has access to the *Oracle10g Data Mining Tutorials* which are available on the CD-ROM that accompanies this project. It is also assumed the user has access to the server machine Athena.ict.ru.ac.za which has the Oracle10g Release 1 version 10.1.0.2 and Oracle data miner version 10.1.0.2 installed and configured for use.

C.1 Preparing for Data Mining

The server machine was installed with the mining tools such that any user can make use of the mining tools. However if the reader has no credentials to the machine, it is possible to login to the machine with the username – **g05m5125** and the password – **final22**.

- Launch the Oracle Data mining server, this is located in the following directory:

  C:\odminer\bin\odminerw.exe

  Double click the executable ***odminerw.exe*** *(look at Appendix C step 5)*

- If it is required to load any datasets into the database it will be necessary to create database tables first. This is achievable using the Enterprise manager (EM). The EM for oracle10g is accessed through a web browser and the URL for this particular database is URL:http://athena.ict.ru.ac.za:5500/em/

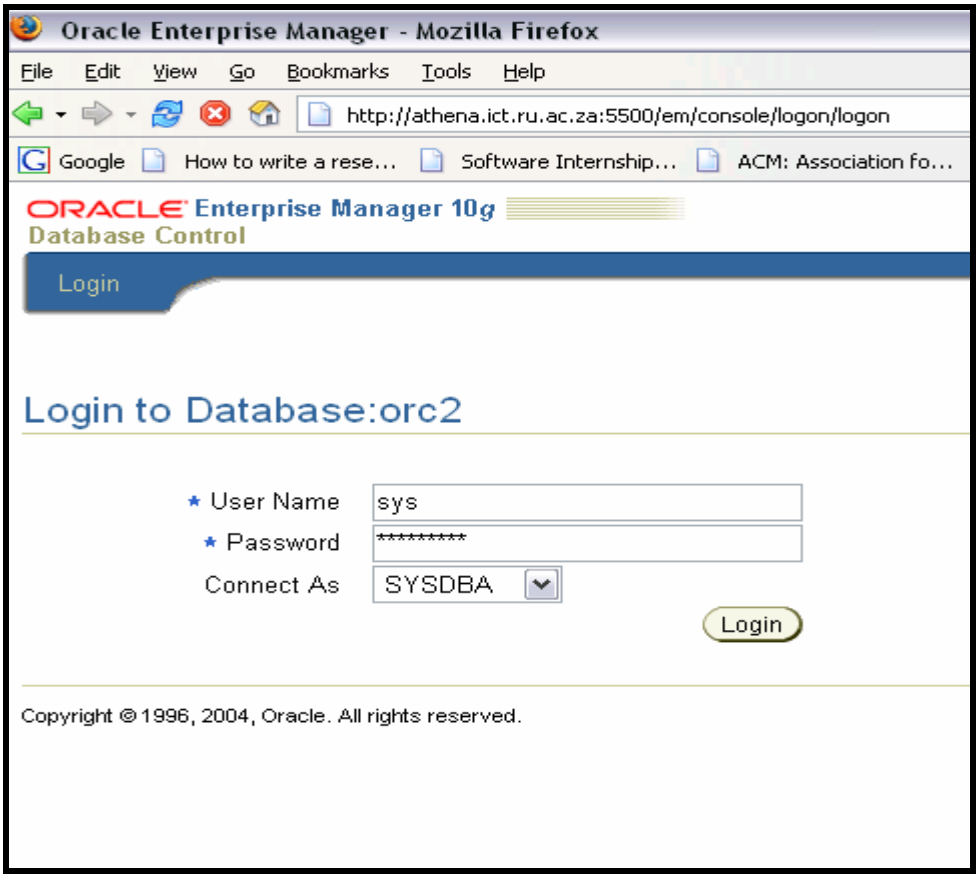


Figure 30: login screen to Oracle10 Enterprise Manager

Here enter the credentials, username-**sys**, and password-**oracle10g** then Connect as-SYSDBA (the system database administrator account) or the Data miner user account with Username-**odm_use** AND password-**datamine** then Connect As normal user.

When logged into the database, through the Enterprise Manager, you can view the tables available or even create tables to load new data (*go to Administration → Tables*).Any data sets that will be mined should be loaded into the **odm_use** schema in the Oracle database. The loading of data can be accomplished by using **sql*loader (sqlldr.exe).**

*Starting sql*loader*

Click on *Start* > *programs* > *command prompt* then cd (change directory) to the directory with the data you want to load into the database. Then type the command *sqlldr* to start the application. Note that sql*loader is an application for loading data into an Oracle database hence it can only be found on machine with the Oracle database installed.

To make a batch load of data sets into a table

Firstly create a control file and save it as a .ctl file in the same directory as the dataset file. The control file contains information on where the dataset file can be found and which database table the dataset is going to be loaded to. Most success in loading datasets was achieved by the use of .txt files with items in records separated by commas and each record on a new line.

Open command prompt, then change directory to the location where the .ctl file is located then type *sqlldr* (sql loader application) and *odm_use/datamine* (ODM account name & password) as follows to load the dataset into the database table:

- *C:\DATA\tsha_tsha>sqlldr odm_use/datamine*

This will prompt for the control file name after which it will load the data.

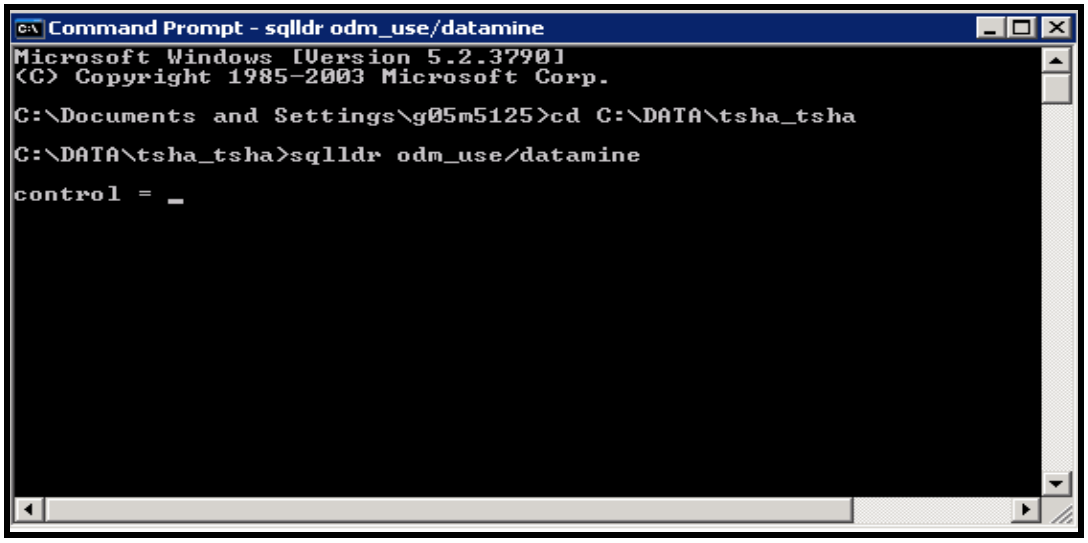


Figure 31: Example of how to load data

C.2 The Actual Data Mining

At this stage the reader should refer to the *Oracle10g Data Mining Tutorials* in the CD ROM that accompanies this research project. The tutorials provide a step by step mining methodology. The tutorials are in different stages, the building by a particular algorithm and applying of the model built. For this research the major algorithms used are the Clustering algorithms (O-Cluster and K-Means), however tutorials for all algorithms used will be provided in the CD ROM.

Appendix D:

Datasets

The dataset files used in this evaluation are available in the CD ROM that accompanies this project.

Spreadsheets

The spreadsheet files used for the evaluation and in the identification of predictors are also available in the CD ROM that accompanies the project.