The Korean Communications
in Statistics Vol. 8, No. 2, 2001
pp. 531-541

A Comparison of Capabilities of Data Mining Tools

Yong-Seok Choi[1], Jong-Geoun Kim[2] and Jong-Hee Lee[3]

Abstract

In this study, we compare the capabilities of the data mining tools of the most updated version objectively and provide the useful information in which enterprises and universities choose them. In particular, we compare the SAS/Enterprise Miner 3.0, SPSS/Clementine 5.2 and IBM/Intelligent Miner 6.1 which are well known and easily gotten.

Keywords : Clustering detection, Data mining tools, Market basket analysis, Decision tree, Neural network, Mining techniques.

1. Introduction

Data mining tools are used widely to solve practical problems in engineering, science, and business. As the number of data mining software vendors increases, however, it has become more challenging to assess which of their rapidly-updated tools are most effective for a given application. The data mining tool market has become more crowded in recent years, with more than 50 commercial data mining tools, for example, listed at the KDNuggets web site (http://kdnuggets.com). Rapid introduction of new and upgraded tools is an exciting development, but does create difficulties for potential purchasers trying to assess the capabilities of off-the-shelf tools (Abbott et al., 1998).

There are those presented in Abbott et al.(1998) and Elder and Abbott(1998) and those released in Korea in comparisons of data mining tools. But the former is not suitable in current, because data mining tools they tested are out of date. And the latter is not objective, because they are made for their company's advertisement. Therefore, the objective comparison of rapidly-updated data mining tools is needed. In this paper, we choose three data mining tools as SAS/Enterprise Miner 3.0, SPSS/Clementine 5.2 and IBM/Intelligent Miner for data 6.1 that are integrated, easy to use, and particular popular in Korea. And they can be used under client-server environment.

[1] Associate Professor, Department of Statistics, Pusan National University, Pusan 609-735, Korea.
E-mail : yschoi@hyowon.pusan.ac.kr

[2] Associate Professor, Group-department of Information Communication, Pusan College of Information technology, Pusan 616-737, Korea.

[3] Department of Statistics, Pusan National University, Pusan 609-735, Korea.

From now, we specify SAS/Enterprise Miner 3.0 as Enterprise Miner, SPSS/Clementine 5.2 as Clementine and IBM/Intelligent Miner for data 6.1 as Intelligent Miner. These will be listed in alphabetical order. And in tables, we note that symbol ○ represents that each tool includes a list, and symbol × is the reverse of it.

In Section 2, we will compare three tools in explanation of main windows and nodes, and techniques(market basket analysis, clustering detection, decision tree, neural network). Finally, we provide some conclusions in Section 3.

2. Comparisons of Tools

In this section, we compare Clementine, Enterprise Miner and Intelligent Miner in constitutions of main windows and nodes, and in techniques. Particularly, technically, we compare concretely three tools in algorithms, controlling options and results.

2.1 Constitutions of Main Windows and Nodes

Firstly, we can compare in constitution of main windows of three tools. Main windows are composed of nodes area and work area generally. In main windows, we can click a node to put it on work area visually. Clementine and Enterprise Miner can link with nodes. And these streams can be understood very well. In visualization with stream, Clementine make streams very minutely.

Next, we can compare nodes's constitutions of three tools. Three tools all array nodes in order of using. But if we observe minutely, three tools are different in constitutions of modelling nodes. Intelligent Miner arrays modelling nodes according to purpose of modelling, but Clementine and Enterprise Miner don't array as Intelligent Miner.

2.2 Techniques

Data mining tools contain various techniques as market basket analysis, clustering detection, decision tree, neural network, and so on.

Here, we will compare algorithms, controlling options, and results of four techniques as market basket analysis, clustering detection, decision tree and neural network, since these four techniques are contained integrated data mining tools commonly.

2.2.1 Market Basket Analysis

Market basket analysis technique gives insight into the merchandise by telling us which products tend to be purchased together and which are most amenable to promotion. This information is actionable: It can suggest new store layout; it can determine which products to put on special: it can indicate when to issue coupons, and so on (Berry and Linoff, 1997, p. 124). In general, there are association rule and sequential pattern in market basket analysis algorithms. Association rule analyzes things that happen at the same time and sequential pattern analyzes things that happen sequentially.

Table 2.1 lists the algorithms implemented in the three tools evaluated. From this table, we can know that Enterprise Miner and Intelligent Miner include association rule and sequential pattern algorithms but Clementine includes only association rule algorithm.

Table 2.1 Algorithms of Market Basket Analysis

ALGORITHMS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Association Rule	○	○	○
Sequential Pattern	×	○	○

Table 2.2 shows five important options such as Minimum of Coverage, Minimum of Support, Minimum of Confidence, Maximum of Items, and Item Constraints in association rule.

Coverage is a *Pr*(condition), Support is a *Pr*(condition and result) and Confidence is a *Pr*(result condition), because, in A ⇒ B, A can be put as a condition and B can be put as a result. And Lift shown in Table 2.3 is a *Pr*(condition and result) / *Pr*(condition) *Pr*(condition). Maximum of Items is a maximum number of items. In IBM Corp.(1999), Item Constraints is said to determine which rules are to be included in or excluded from the results.

From Table 2.2, we can know that Clementine includes Minimum Coverage and Maximum of Items, and Enterprise Miner includes Minimum of Support, Minimum of Confidence and Maximum of Items, and Intelligent Miner includes Minimum of Support, Minimum of Confidence, Maximum of Items and Item Constraints.

Table 2.2 Options of Association Rule

OPTIONS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Minimum of Coverage	○	×	×
Minimum of Support	×	○	○
Minimum of Confidence	×	○	○
Maximum of Items	○	○	○
Item Constraints	×	×	○

Table 2.3 lists Lift, Support, Confidence, Coverage, Textual Display and Visualization of Association in results of association rule. Textual Display represents association rules in English sentences. Visualization of association displays visually using Link analysis.

From this table, we can know that Enterprise Miner and Intelligent Miner include Lift, Support and Confidence but Clementine doesn't include Support and Lift. And both Clementine and Enterprise Miner include Visualization of Association, but Clementine show it more efficiently. Additionally, Intelligent Miner includes Textual Display.

Table 2.3 Results of Association Rule

RESULTS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Lift	×	○	○
Support	×	○	○
Confidence	○	○	○
Coverage	○	×	×
Textual Display	×	×	○
Visualization of Association	○	○	×

In market basket analysis, we can compare in algorithms, controlling options and results. we can compare in algorithms, and fundamental Lift, Support in options and results. Clementine has Visualization of Association to be understood easily, but Clementine is the most insufficient in three tools. And Enterprise Miner has foundations needed: Lift, Support, and Confidence. And Intelligent Miner has foundations needed and Textual Display.

2.2.2 Clustering Detection

Clustering detection is one of the few data mining activities that can properly be described as undirected knowledge discovery or unsupervised learning. And it is a finding subsets to have common characteristics of data records.

There are K-means, Kohonen map and demographic clustering algorithms in clustering detection at three tools.

Table 2.4 lists the algorithms implemented in the three tools evaluated. From this table, we can know that Clementine and Enterprise Miner include K-means and Kohonen map algorithms, but Intelligent Miner includes Kohonen map and demographic clustering algorithms.

Table 2.4 Algorithms of Clustering Detection

ALGORITHMS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
K-means	○	○	×
Kohonen Map	○	○	○
Demographic	×	×	○

Table2.5 shows six important options such as Standardization, Elimination of Outliers, Stopping Criterion, Missing Values, Number of Cluster and Categorical Data in K-means algorithm.

Standardization represents having standardizing methods of lengths of variables, Elimination of Outliers is very important and necessary in clustering detection, and Clementine and Enterprise Miner have each node for this. Stopping Criterion and Missing Values represents various methods to terminate the training process and to include or exclude missing values, respectively. Number of Cluster can assign number of clusters to use. Categorical Data represents handling methods of categorical data in K-means algorithm.

Since Intelligent Miner doesn't have K-means algorithm, it is excluded from table 2.5. In particular, Clementine includes fundamental Stopping Rule and Number of Cluster, and Enterprise Miner includes Clustering Criterion, Initial Seeds, Stopping Rule, Missing Values and Number of Cluster. Therefore, Enterprise Miner has the most various options in K-means.

Table 2.5 Options of K-means

OPTIONS	MINING TOOLS	
	Clementine	Enterprise Miner
Standardization	×	○
Elimination of Outliers	○	○
Stopping Criterion	○	○
Missing Values	×	○
Number of Cluster	△	○
Categorical Data	○	○

Table 2.6 shows Clusters, Distances and Variables in results of clustering detection. The Clusters means an existence of explanation inside cluster. The Distances means dissimilarity between clusters, and the Variables means degrees of the importance of variable. In particular,

Enterprise Miner and Intelligent Miner include all, but Clementine includes only Clusters and Distances.

Table 2.6 Results of Clustering Detection

RESULTS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Clusters	○	○	○
Distances	○	○	○
Variables	×	○	○

In clustering detection, we can compare algorithms, controlling options and results. From comparison in algorithms, Intelligent Miner doesn't have K-means, representative algorithms. From comparison in controlling options, Enterprise Miner has the most various controlling options of K-means. From comparison in results of Clustering Detection, since Enterprise Miner and Intelligent Miner have Clusters, Distances, and Variables, but Clementine has only Clusters and Distances. Conclusively, Since Intelligent Miner doesn't have K-means algorithm and Clementine is insufficient in controlling options and results.

2.2.3 Decision Tree

Decision tree is a technique for classification and prediction. The attractiveness of tree-based methods is due in large part to the fact that decision tree represents rules. Rules can readily be expressed in English so that we humans can understand them.

There are many algorithms in decision tree, but we compare CHAID, CART, C4.5 or C5.0, ID3 and SPRINT in the Table 2.7.

CHAID(Chi Square Automatic Interaction Detection) used for classification, and provide a set of rules that can be applied to a new(unclassified) data set to predict which records will have a given outcome. CHAID segments a data set by using chi square test to create multi-way splits. CART(Classification and Regression Tree) is used for classification, and provides a set of rules that can be applied to a new(unclassified) data set to predict which records will have a given outcome. CART segments a data set by creating 2-way splits and requires less data preparation than CHAID (http://www.exclusiveore.com/index.html). C4.5 or C5.0 is the most recent available snapshot of the decision-tree algorithm. C4.5 or C5.0 produces trees with varying numbers of branches per node. ID3(Iterative Dichotomizer 3), the precursor to C4.5, uses a criterion called information gain to compare potential splits (Berry and Linoff, 1997, p. 261).

Table 2.7 lists the algorithms implemented in the three tools evaluated. From this table, we can know that Enterprise Miner includes representative algorithms as CHAID, CART and C4.5

algorithms, but Clementine only includes C5.0 and ID3 algorithms and Intelligent Miner only includes SPRINT. Therefore, Enterprise Miner has the most representative and various algorithms.

Table 2.7 Algorithms of Decision Tree

ALGORITHMS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
CHAID	×	○	×
CART	×	○	×
C4.5 or C5.0	○	○	×
ID3	○	×	×
SPRINT	×	×	○

Table 2.8 displays Misclassification Costs, Priors, Pruning Severity, Stopping Rule and Missing Value in options of decision tree.

Misclassification Costs allows the user to specify that some errors (or misclassifications) are more undesirable than others. The list of misclassification costs shows a list of possible misclassifications preceded by their associated costs. Priors is used to assign discrete target variable for prior probability. Pruning Severity is the removal of branch which has high risk that make classification error large and has inappropriate induction rule. Stopping Rule means having various rules that don't more split and end in a current node. Missing Value means that variables having large agreement surrogate missing values.

Firstly, Misclassification Costs and Priors put weights in data in advance, therefore they can have an influence on error rate. Enterprise Miner includes Misclassification Costs and Priors, and Clementine only includes Misclassification Costs, but Intelligent Miner doesn't include both Misclassification Costs and Priors. Therefore, Intelligent Miner is the insufficient in controlling error rate. In the remainders, all of the three tools are the same.

Table 2.8 Options of Decision Tree

OPTIONS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Misclassification Costs	○	○	×
Priors	×	○	×
Pruning Severity	○	○	○
Stopping Rule	○	○	○
Missing Value	○	○	○

Table 2.9 shows Tree View and Confusion Matrix in results of decision tree. From this table, we can know that Enterprise Miner and Intelligent Miner include both Tree View and Confusion Matrix. But Clementine doesn't include both, and only includes error rate and rule set. Therefore, Clementine is the insufficient in three tools.

Table 2.9 Results of Decision Tree

RESULTS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Tree View	×	○	○
Confusion Matrix	×	○	○

In Decision Tree, we can compare in algorithms, controlling options and results. From these comparisons in algorithms and controlling options, Enterprise Miner has the most various algorithms and controlling options. From comparisons in results, Enterprise Miner and Intelligent Miner have foundations needed as Tree View and Confusion Matrix, but Clementine doesn't have both.

2.2.4 Neural Network

Neural network is extremely simple model of the way that the nervous system operates. The basic unit is the "neuron", and these are typically organized into "layers".

We will compare two algorithms widely used in neural network in Table 2.10, though there are various algorithms in neural network.

It is Multilayer Perceptron(MLP) that most widely used for data analysis. MLP is composed of input layer, hidden layer and output layer. Yet Radial Basis Function(RBF) has one hidden layer and it's mathematical form is similar to that of MLP, but it differs in using RBF as combination function in hidden layer. we note that symbol △ in this table represents a partial existence of algorithm. Clementine, Enterprise Miner and Intelligent Miner include MLP and RBF, but Intelligent Miner only use RBF not in classification but in prediction.

Table 2.10 Algorithms of Neural Network

ALGORITHMS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
MLP	○	○	○
RBF	○	○	△

Table 2.11 displays Learning Rate, Momentum, Stopping Criterion, Normalized Input, Number of Hidden Layer and Set Random Seed in options of neural network.

Learning Rate controls how quickly the weights change. The best approach for the learning rate is to start big and decrease it slowly as the network is being trained. Momentum refers to the tendency of the weights inside each unit to change the "direction" they are heading in. Namely, each weight remembers if it has been getting bigger or smaller, and momentum tries to keep it going in the same direction (Berry and Linoff, 1997, pp. 304-305). Stopping Criterion represented an existence of various methods to terminate the training process. Normalized Input scales continuous and discrete numeric fields to a range of 0.0 to 1.0 and converts categorical data into 1/N . Number of Hidden Layer is an important decision. The more units, the more patterns the network can recognize. This would argue for a very large hidden layer. However, there is a drawback. The network might end up memorizing the training set instead of generalizing from it. Set Random Seed uses the given seed for the random number generator used to select internal training and test sets and initialize weights.

All of the three tools include fundamental options: Learning Rate, Momentum, and Number of Hidden Layer. But Clementine excludes Normalized Input and Intelligent Miner excludes Set Random Seed.

Table 2.11 Options of Neural Network

OPTIONS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Learning Rate	○	○	○
Momentum	○	○	○
Stopping Criterion	○	○	○
Normalize Input	×	○	○
Number of Hidden Layer	○	○	○
Set Random Seed	○	○	×

Table 2.12 displays Sensitivity Analysis, Confusion Matrix, and Statistics in results of neural network.

Sensitivity Analysis does not provide explicit rules, but it does indicate the relative importance of the inputs to the result of the network. Sensitivity Analysis uses the test set to determine how sensitive the output of the network to each input. Confusion Matrix is a table of showing relations with real categories of target variable and predicted categories by model. it show correct classified frequency and incorrect classified frequency. Statistics shows goodness-of-fit, new statistics and history statistics.

From Table 2.12, we can know that Clementine and Intelligent Miner include Sensitivity Analysis, and Intelligent Miner also includes Confusion Matrix, but Enterprise Miner includes only Statistics.

Table 2.12 Results of Neural Network

RESULTS	MINING TOOLS		
	Clementine	Enterprise Miner	Intelligent Miner
Sensitivity Analysis	○	×	○
Confusion Matrix	×	×	○
Statistics	×	○	×

In neural network, we can compare in algorithms, controlling options and results. From comparisons in algorithms, since Intelligent Miner only uses RBF in prediction and can't use in classification. Namely, Intelligent Miner has insufficient RBF algorithm. From comparisons in controlling options, all of the three tools include fundamental options, but Clementine excludes Normalized Input and Intelligent Miner excludes Set Random Seed. From comparisons in results, Intelligent Miner has fundamental results as Sensitivity Analysis and Confusion Matrix.

3. Conclusion

We compare three tools in constitutions of main windows and nodes and techniques. From comparison in constitutions of main windows and nodes, we can see that Clementine is the best of three tools in constitutions of main windows, because it can make streams most visually and Intelligent Miner is the best of three tools in constitutions of nodes, since it arrays modelling nodes according to purposes of modelling.

From comparisons in techniques, The results are as following: In market basket analysis, Clementine is the most insufficient in three tools, since it doesn't have sequential pattern in algorithms, and doesn't have fundamental Lift, Support in options and results and it has Input Data Horizontal Format. Intelligent Miner is the best in three tools, since it has fundamental options, results, additional option, and pivoting operator of input data format. In clustering detection, Intelligent Miner doesn't have K-means algorithm, and Clementine is insufficient in controlling options and results. Enterprise Miner is the best in three tools, since it has K-means, Kohonen map, and various controlling options and results. In decision tree, Enterprise Miner is the best in three tools, since it has various algorithms, controlling options and results. In neural network, Intelligent Miner has the most insufficient algorithms, and Enterprise Miner has the most various options, and Intelligent Miner has most various results of neural network

References

[1] Abbott, D. W., Matkovsky, I. P. and Elder IV, J. F.(1998)."An Evaluation of High-end Data Mining Tools for Fraud Detection", *SMC-98.*

[2] Berry, M. J. A. and Linoff, G.(1997). *Data Mining Techniques for Marketing, Sales, and Customer Support,* New York: John Wiley & Sons, Inc.

[3] Elder IV, J. F. and Abbott, D. W.(1998). "A Comparison of Leading Data Mining Tools", *KDD-98 Tutorial Slides.*

[4] http://www.exclusiveore.com/index.html.

[5] IBM Corp.(1999). *IBM DB2 Intelligent Miner for Data Using the Intelligent Miner for Data Version 6 Release 1.*

[6] Integral Solutions Ltd.(1998a). *Clementine Reference Manual Version 5.*

[7] Integral Solutions Ltd.(1998b). *Clementine User Guide Version 5.*

[8] SAS Institute Inc.(1999). *Enterprise Miner Reference.*