EVALUATION OF CLUSTERING TECHNIQUES
IN DATA MINING TOOLS

Manying Qiu, Virginia State University, mqiu@vsu.edu
Steve Davis, Clemson University, davis@clemson.edu
Fidelis Ikem, Virginia State University, fikem@vsu.edu

ABSTRACT

Clustering divides a heterogeneous population into a number of more homogeneous subgroups or clusters to reflect the segments in a dataset such that patterns can be recognized. This research shows how a software evaluation framework may be adapted to evaluate commercial data mining tools for a specific user environment. We applied this adaptation to evaluate two major commercial data mining tools, SAS Enterprise Miner (EM) and IBM DB2 Intelligent Miner (IM), for use in a university environment.

Keywords: Data mining, data mining tools, clustering, software evaluation

INTRODUCTION

Segmentation as a marketing term simply means making different offers to different market segments—groups of people defined by some combination of demographic variables or by certain lifestyle indicators” (1). Businesses often use segmentation to identify groups of customers having similar buying behaviors and then develop marketing strategies best for target customers in each group to promote sales. However recognizing segments, for example, in a database of potential customers, is difficult for humans and generally requires support of clustering algorithms in a data mining tool. A common practice is to use clustering algorithms to assign similar records into appropriate clusters based on similarity of the variables (or attributes). Clustering divides a heterogeneous population into a number of more homogeneous segments. By analyzing the meaning of clusters formed, we can further understand the whole population.

Clustering is different from classification because clustering does not require predefined classes. The records are grouped based on self-similarity. It is up to the user to interpret the resulting clusters. Clustering is undirected knowledge discovery -- no target variable (or dependent variable) is defined. Clustering tools assign groups of records to the same cluster if they have something in common, making it easier to discover meaningful patterns from the dataset. Clustering often serves as a starting point for some supervised data mining techniques or modeling.

Because it is not clear how to choose a clustering software package, this paper adapts Collier et al’s framework (4) that considers four criteria: performance, functionality, usability and ancillary task support. However other criteria may be necessary for evaluation for certain user environments. We will suggest an additional criterion for an academic environment and apply the extended framework to evaluate clustering techniques in two representative commercial data mining tools, namely SAS Enterprise Miner and IBM DB2 Intelligent Miner. This example of a
structured evaluation process could be adopted by organizations to compare characteristics of clustering tools in order to select one that meets their needs.

OVERVIEW OF CLUSTERING

In information systems, a large database may contain too many variables, too many dimensions (i.e. attributes), and a structure too complex for even the best designed direct knowledge discovery mining tool to recognize any meaningful patterns from it. In many cases, the problem is not that no patterns exist, but that there are too many. When people try to make sense of complex scenarios, a natural tendency is to break the big picture into smaller pieces, each of which can be explained more simply. However, in many cases, we are not sure what variables and attributes differentiate the subgroups that constitute the population. Therefore, we need automatic cluster detection to help us learn more about the trees before we can see the forest.

Data Transformation

Usually data needs to be transformed before use in clustering, often because different attributes may be measured on different scales, e.g., miles and milligrams. In cases where the range of values differs widely from attribute to attribute, these differing attribute scales can contaminate clustering results and it is common practice to standardize the data so that all attributes are on the same scale. Another common practice is to discard variables that show little variation or have high correlation with other variables to reduce the dimensions of clustering.

Types of Clustering

Clustering techniques are heuristic in nature. Almost all techniques have a number of arbitrary parameters that can be “adjusted” to improve results.

Clustering techniques fall into the following broad categories (10):
1) **Hierarchical vs. partitional.** Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual instances at the bottom. Each intermediate level can be viewed as a combination of two clusters from the next lower level, or a split of a cluster from the next higher level into two. Partitional (or non-nested) techniques create a one-level partitioning of the data instances. After the user specifies the desired number of clusters, a partitional approach typically finds all clusters at once. This is in contrast to traditional hierarchical schemes, which bisect a cluster to get two clusters or merge two clusters to get one.
2) **Divisive vs. agglomerative.** Hierarchical clustering techniques proceed either from the top to the bottom or from the bottom to the top, i.e., clustering starts with one large cluster and splits it, or starts with clusters each containing a point and then merges them.
3) **Incremental vs. non-incremental.** Some clustering techniques work with one instance at a time and decide how to place it into an appropriate cluster, but most clustering techniques are non-incremental, using information about all the instances at once to form clusters.
Cluster Interpretation

The weakness of automatic cluster detection is lack of explanation. When you don’t know what you are looking for, you may not recognize it when you see it. However, it’s also possible that clusters generated by the automated clustering algorithms do not have any practical value at all. One way to gain insights about clusters is to use direct knowledge discovery mining tools to analyze each cluster separately. There are several approaches to understanding clusters. The most common approach is to build a decision tree with the cluster label as the target variable and use the decision tree to derive rules explaining how to assign new records to the correct cluster. Another approach is to use visualization to see how the clusters are affected by changes in the input variables.

DATA MINING SOFTWARE EVALUATION FRAMEWORK

We adapt the data mining software evaluation framework (4) to evaluate clustering techniques in data mining tools, using the following categories and associated evaluation criteria:

1. **Performance** – The ability to handle a variety of data sources in an efficient manner. Criteria: (1) software architecture; (2) heterogeneous data access.
2. **Functionality** – the inclusion of a variety of capabilities, techniques, and methodologies for data mining. Criteria: (1) algorithmic variety; (2) prescribed methodology.
3. **Usability** – accommodation of different levels and types of users without loss of functionality or usefulness. Criteria: (1) user types; (2) data visualization.
4. **Ancillary Task Support** -- allows the user to perform data cleansing, manipulation, transformation, visualization and other tasks that support data mining. Criteria: (1) data filtering; (2) deriving attributes.

Other criteria could also be important for our sample evaluation for an academic environment. For example, beyond this framework we will address differences in cost, which includes the extent of effort needed to interface with our DBMS and software.

SAS Enterprise Miner

SAS Institute’s Enterprise Miner (EM) addresses the data mining process through an intuitive point-and-click graphical user interface. EM offers seamless integration with data warehousing, reporting and the Web (2, 8).

**Software Architecture** EM sits on top of a large, bundled collection of SAS statistical products. EM is available in several configurations: standalone, workstation, or client/server. In the latter case, you can perform analysis on both the workstation and the server simultaneously (7). The client/server system integrates fully with other SAS software and allows applications to be deployed via the Internet (2).

**Heterogeneous Data Access** SAS software has built its reputation on the ability to access, manage and analyze data from various sources, and provides a range of drivers that allow full access to data stores. EM is able to access data in warehouse/data mart and over 50 different file structures. Supported data sources include all the major relational databases as well as non-SQL data sources, PC sources and ODBC compliant databases (2,8).
Algorithmic Variety  Enterprise Miner supports clustering and segmentation of databases, using k-means clustering, self-organizing maps and kohonen networks techniques. SAS Text Miner automatically groups documents into topical clusters (8). The results of clustering can be passed to other nodes such as Decision Tree Node for explanation of the clusters formed. It can also be passed as a group variable that enables the user to automatically construct separate models for each cluster.

Prescribed Methodology  EM provides a guiding, yet flexible, framework for conducting clustering by EMSMA (Explore, Modify, Sample, Model and Assess):

- Explore your data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
- Modify your data by creating, selecting, and transforming the variables.

Three additional capabilities are available when applying direct knowledge discovery algorithms to analyze the clusters formed:

- Sample your data by extracting a portion of a data set big enough to contain the significant information, yet small enough to manipulate quickly.
- Model your data by formalizing the patterns, from which predictions can be made.
- Assess your data by evaluating the usefulness and reliability of the findings.

The user builds a process flow diagram by dragging and dropping nodes in the workspace.

User Types  EM is designed for a combination of beginning, intermediate, and advanced users. Analysis and modeling capabilities are provided for business users and professional users having statistical insight and business knowledge.

Data Visualization  EM supports visual analysis and reporting. Graphics including 3D rotating charts are used to view the desired data and the clustering results. If users use direct knowledge discovery data mining tools to analyze the clusters formed, they can watch a model being built in real time through a window, and can stop the operation any time they wish.

Data Filtering  The Filter Outliers node supports data filtering, for example, identifying and removing outliers from data sets. Users can eliminate rare values in class variables and/or extreme values in interval variables.

Deriving Attributes  A user can create new variables from existing ones in the Transform Variables node. EM provides default functions, such as squares, inverse, exponential, standardized, square root and logarithms, but allows users to input their own computational formulae and build expressions.

IBM DB2 Intelligent Miner

Intelligent Miner (IM) enables users to apply clustering for market segmentation, store profiling and reveal buying behavior. The business analysis results are applicable in real-time, and the results of data mining can be driven back into operational applications (6). Academic users can download for free IM and its underlying DB2 database from the IBM Web site.

Software Architecture  IBM normally implements data mining in an organization as part of a data warehousing architecture. The main processing component, the Mining Kernel, is a client/server system (3).

Heterogeneous Data Access  IM is not specific to IBM data warehouses or databases. It can process information stored in DB2 or flat files on AS/400, AIX or MVS. Clustering operations can also be performed directly on Oracle or Terradata databases accessible by DataJoiner.
addition, a high-speed extract is provided to import into DB2 Universal Database data from Oracle, Sybase, or DB2 for OS/390 databases (3,9).

**Prescribed Methodology** IBM provides minimum user guidance. There is no single set way of using IM. Different users utilize the tool set differently, using the operations alone or in combination, that best meet the needs of the business. However, a custom interface to select pre-defined subsets of the function is designed for business analysts.

**Algorithmic Variety** IBM offers clustering that generates simple groupings of entities which share common properties. IM text mining features provide a means of clustering, discovering predominant themes in a collection of documents (9).

**User Types** The users include “expert” mining analysts and business analysts. IBM has provided a set of technologies that allow developers who have an understanding of the issues facing clustering to build their own solutions, based on the algorithms and functions that IBM provides. Initially, they are still assisted by IBM consultants.

**Data Visualization** The main client window has an Explorer-like view, showing a directory of the data available by type. After the data has been mined, IM provides a range of viewing options including charts and graphs, histograms and tree diagrams. A cluster map provides an overview of the categories. Color coding and navigation aids guide users to the desired information.

**Data Filtering** IBM has developed techniques that allow data to be filtered and reduced, for example, filtering variables out of input records, filtering records out of the input database and filtering records using a value set.

**Deriving Attributes** To prepare data for mining, you may derive attributes by applying formulas to existing fields. For example, you may aggregate records using built in functions for average, minimum, maximum, sum, and count (3,5).

**COMPARATIVE ANALYSIS**

Let’s assume that we are University professors looking for a clustering tool for our teaching, research and consulting activities. We teach graduate courses on data mining in Information Systems Department. We shall evaluate using the aforementioned framework (4) and an additional criterion. The criteria within each category are assigned weights appropriate for our environment such that the total of these weights within each category equals 1.00 or 100%. Likewise each category is assigned a weight. Next the tools can be scored for comparison. SAS Enterprise Miner is used as a reference. Enterprise Miner receives a score of 3 for each criterion, and the other tool is rated against Enterprise Miner for each criterion using the following scale:

<table>
<thead>
<tr>
<th>Relative Performance</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse than Enterprise Miner</td>
<td>1</td>
</tr>
<tr>
<td>Worse than Enterprise Miner</td>
<td>2</td>
</tr>
<tr>
<td>Same as Enterprise Miner</td>
<td>3</td>
</tr>
<tr>
<td>Better than Enterprise Miner</td>
<td>4</td>
</tr>
<tr>
<td>Much better than Enterprise Miner</td>
<td>5</td>
</tr>
</tbody>
</table>

The three authors compared tools and agreed upon the following ratings. For an actual evaluation we would invite faculty members who are interested in clustering to participate in the rating. We would average the ratings to fill in the following table.
### Decision Tree Tool Evaluation Scoring by University Professors

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Weight</th>
<th>Enterprise Miner</th>
<th>Intelligent Miner</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance (0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Architecture</td>
<td>0.5</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Heterogeneous data access</td>
<td>0.5</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>Category Score</strong></td>
<td></td>
<td>3.0</td>
<td>2.0</td>
</tr>
<tr>
<td><strong>Functionality (0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithmic Variety</td>
<td>0.5</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Prescribed Methodology</td>
<td>0.5</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>Category Score</strong></td>
<td></td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td><strong>Usability (0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Types</td>
<td>0.6</td>
<td>3</td>
<td>1.8</td>
</tr>
<tr>
<td>Data Visualization</td>
<td>0.4</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Category Score</strong></td>
<td></td>
<td>3.0</td>
<td>2.4</td>
</tr>
<tr>
<td><strong>Ancillary Task Support (0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Filtering</td>
<td>0.6</td>
<td>3</td>
<td>1.8</td>
</tr>
<tr>
<td>Deriving Attributes</td>
<td>0.4</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Category Score</strong></td>
<td></td>
<td>3.0</td>
<td>2.6</td>
</tr>
<tr>
<td><strong>Cost (0.2)</strong></td>
<td>1.0</td>
<td>3</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Weighted Average</strong></td>
<td></td>
<td>3.0</td>
<td>2.3</td>
</tr>
</tbody>
</table>

**Software Architecture** IM gets 2 for this criterion because it only uses client-server architecture. Although we have implemented computer networking, stand-alone software architecture certainly is convenient to students or professors who want to do analysis on their own stand alone computers. EM is preferred.

**Heterogeneous data access** Although IM can extract data from other sources it is mainly oriented on extracting data from DB2, so it gets 2 for this criterion. EM can extract data from a variety of DBMS conveniently.

**Algorithmic Variety** Both EM and IM support clustering techniques. Both provide various direct knowledge discovery techniques for interpreting the clustering results. IM gets 3 for this criterion.

**Prescribed Methodology** EM’s SEMMA (sample, explore, modify, model and assess) provides a methodology that clarifies the clustering process. IM is not as good as EM in this respect. IM gets 2 for this criterion.

**User Types** EM can be utilized by beginning, intermediate and advanced users, so it will be good for students who have never used EM. Intelligent Miner is not suited for beginners, so it gets 2 for this criterion.
**Data Visualization** IM gets the same rating as EM.

**Data Filtering** Although they are implemented in differently, both clustering tools have about the same capability and get 3 for this criterion.

**Deriving Attributes** IM doesn’t offer as many default functions to derive attributes as Enterprise Miner does, so it gets 2 for this criterion.

**Cost** There is not much difference in license prices for academic users. Because we have been using other software in the SAS packages for teaching, research and service activities we may get a lower marginal price for EM. Therefore, we assign 2 to IM.

Enterprise Miner won with its reference score of 3.0, whereas the weighted average score for Intelligent Miner was 2.3

## CONCLUSION

Modern data mining software provides clustering tools that divide a heterogeneous, complex data population into more homogeneous, meaningful subgroups such that patterns can be discovered easier. Direct knowledge discovery tools in the data mining software can be used to explain the resulting clusters. We adapted a data mining software evaluation framework (2) and demonstrated it by evaluating two different data mining tools for use by university professors. This framework appeared appropriate and useful. It could be employed by others having different environments by augmenting it as necessary with environment-specific criteria and adjusting the weights.

## REFERENCES