D2.2.5 MineSet™

Cliff Brunk   Ron Kohavi

brunk@engr.sgi.com   ronnyk@engr.sgi.com

Data Mining and Visualization

Silicon Graphics, Inc.

2011 N. Shoreline Blvd, M/S 8U-850

Mountain View, CA, 94043

D2.2.5.1 Abstract

MineSet™ is a commercial data mining product from Silicon Graphics. It provides an interactive platform for data mining, integrating three powerful technologies: database and file access, analytical data mining engines, and data visualization. MineSet supports the knowledge discovery process from data access and preparation through iterative analysis and visualization to deployment. MineSet uses a client-server architecture for scalability and support of large data. The data access component provides a rich set of transformations that can be used to process stored data into forms appropriate for visualization and analytical mining. MineSet's 2D and 3D visualization capabilities allow direct data visualization for exploratory analysis. The analytical mining algorithms create models that can be viewed using visualization tools specialized for the learned models or deployed as part of a larger system. Third party vendors can interface to the MineSet tools for model deployment and for integration with other packages.

D2.2.5.2 Introduction

MineSet (Silicon Graphics 1998, Brunk, Kelly & Kohavi 1997) is a general purpose data analysis tool that provides database access, analytical data mining, and data visualization in a highly integrated environment that supports the knowledge discovery process (Fayyad, Piatetsky-Shapiro & Smyth 1996). In addition, MineSet is a platform for developing vertical applications that require analytical data mining and
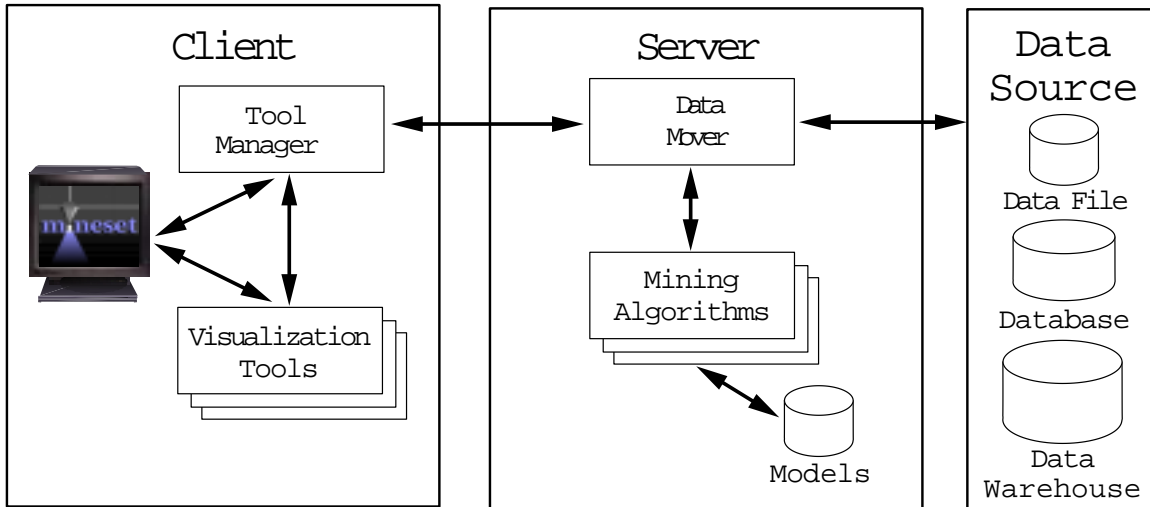
Figure D2.2.5.1: MineSet's three-tier architecture

visualization. MineSet is an evolving product. The following is a description of the 2.6 release.

We begin with an overview of the system architecture then describe the analytical algorithms, the visualization techniques, and the support for KDD process management. We conclude with a brief history of the project and commercial uses.

D2.2.5.3 Architecture

MineSet employs a three tiered architecture (Figure D2.2.5.1). The first tier is the client, which includes Tool Manager and the visualization tools. Tool Manager is the graphical user interface through which the user interacts with MineSet. The visualization tools are used to display data and models of data generated by the mining algorithms. After invoking a visual tool with Tool Manager, the user interacts directly with that tool and sends information from it to other tools via Tool Manager. The second tier is the server, which includes Data Mover and the analytical mining engine. Data Mover is the database access and data transformation component of MineSet. It extracts data from the source, transforms it, and orchestrates moving it from one MineSet component to another. The mining tools are used to generate models of the transformed data, which can be applied to new data or visualized. The third tier is the data source, which includes the storage subsystem that maintains the user's data. It can be either a file or a commercial database. The tiers are not tied to specific machines: all three can reside on a single hardware platform or three separate platforms. This architecture provides the flexibility needed to scale with

the size of the problem. Its allows large mining tasks to be performed on a powerful server machine, while smaller pilot projects can be performed on a desktop machine.

Knowledge discovery is a time consuming and iterative process involving modeling data then understanding and validating the model. Useful tools facilitate this process by generating models of the data quickly and allowing the user to interact with and understand those models. Because speed is of primary importance, MineSet's analytical mining algorithms operated on data in core memory and key components have been parallelized to further reduce execution time on multi processor machines. Although limited to core memory MineSet supports 64-bit addressing allowing access to large amounts of memory.

D2.2.5.4 Analytical Algorithms

MineSet uses $\mathcal{MLC}$++ [link to section D2.1.2] (Kohavi, Sommerfield & Dougherty 1997) as its analytical engine. The naive Bayes (Domingos & Pazzani 1997), decision tree (Quinlan 1993), option tree (Kohavi & Kunz 1997), k-means clustering (Dasarathy 1990), regression tree (Breiman, Friedman, Olshen & Stone 1984), decision table (Kohavi & Sommerfield 1998), association rule generation (Srikand & Agrawal 1995), and feature selection algorithms (Kohavi & John 1997) in $\mathcal{MLC}$++ have been made accessible through MineSet's Tool Manager. The emphasis has been on selecting algorithms that generate interpretable models that facilitate data understanding. Algorithms that create "black box" models, like neural networks, provide little insight into the data and have not yet been included in MineSet.

A plug-in API provides support for algorithms developed outside the $\mathcal{MLC}$++ framework. For instance, Ultimode has released a MineSet add-on plug-in called ACPro for clustering based on AutoClass (Cheeseman et al. 1988). This is extremely important because it is unrealistic to expect a single off-the-shelf tool to provide all the algorithms needed to analyze data in every problem domain. Instead MineSet provides the infrastructure common to the discovery process and allows the user to extend the tool as needed by plug-in algorithms specific to their task.

D2.2.5.5 Visualization

MineSet provides a rich set of visualization tools that enable users to interactively explore data and quickly discover new patterns, trends, and relationships. These

2D and 3D visualization capabilities allow direct data visualization for exploratory analysis, including tools for displaying high-dimensional data taking advantage of geographical and hierarchical information. In addition the visualization techniques have been specialized for displaying the models generated by the analytical mining algorithms. The algorithms help the user identify potentially interesting models of the data. The visual tools help make these models more understandable and allow the user to interact with the models to gain more insight into the model and the underlying data.

The human perception system can identify anomalies and patterns much faster in a representative landscape than in a spreadsheet. The visual tools utilize 3D landscapes that take advantage of a person's ability to navigate in space, track movement and compare objects of different sizes, colors, and shapes. In addition to visualization and navigation, the tools contain filtering and search facilities that allow users to quickly reduce the landscape to items of interest.

MineSet includes eight visualization tools. The *Statistics Visualizer* display basic statistics in histograms and box plots. The *Cluster Visualizer* extends the Statistics Visualizer to show the attribute by attribute differences between clusters identified by the clustering algorithms. The *Tree Visualizer* displays data hierarchically. Users can determine the hierarchy and map attributes to a histogram at each node. The *Map Visualizer* (Figure D2.2.5.3 left) displays data with a spatial component. A polygonal map must be provided and two attributes can be mapped to the polygon's height and color. Multiple maps may be linked together to show different attributes for the same spatial geography. The *Scatter Visualizer* displays scatter plots with up to eight dimensions: three axes, entity color, entity size, entity rotation,and two independent attributes shown through animation. It is also used to visualize the confidence and support of one-to-one association rules. The *Splat Visualizer* (Becker 1997) (Figure D2.2.5.2) extends the scatter plots when there are more than tens of thousands of records. It blurs the points using Gaussian smoothing. The *Decision Table Visualizer* (Kohavi & Sommerfield 1998) (Figure D2.2.5.3 right, shows the break down of class label according to attribute value. Initially, the two most predictive attributes are shown, but the user can show additionally informative attributes by clicking on the cakes of interest and drilling down. This provides visual OLAP (On-Line Analytical Processing) capability. The *Evidence Visualizer* (Becker, Kohavi & Sommerfield 1997) shows a graphic representation of the naive Bayes model [link to section C5.1.5] and allows the user to interact with the model by selecting known values, providing what-if analysis.

Additional capabilities shared by most visual tools include: mapping attributes to graphical attributes (color, height, shape); manipulating the scene using thumbwheels and mouse movements for rotation, translation, and zoom; data slicing and animating by manipulating sliders for two additional independent variables as shown in left; searching and filtering of data; drilling-down by pointing to elements in the scene; and sending records associated with selected visual elements to the Tool Manager for further analysis.

As with the analytical algorithms, emphasis has been placed on selecting visualization techniques that are relatively simple to interpret. Techniques which require extensive training to understand like parallel coordinates and grand tours have yet to be included in MineSet.

D2.2.5.5 KDD Process Management

MineSet is more than an ensemble of data access, transformation, analytical mining and visualization techniques connected by a common user interface. In addition to providing a consistent interface to all the tools, MineSet's Tool Manager provides a history mechanism that allows users to review and edit the steps performed in the current analysis, and change data sources and perform the same analysis on different data sets. Once satisfied with an analysis sequence it can be stored permanently and applied automatically to future data, or updated to account for changing future conditions.

D2.2.5.6 History

MineSet first released in early 1996 primarily as a visualization product. The importance of connecting to commercial databases was recognized early, with a native connection to Oracle® in MineSet 1.0, followed by connections to Sybase®, Informix®, and flat (ASCII) files in MineSet 1.02.

MineSet 1.1 integrated machine learning algorithms from $\mathcal{MLC}$++ (Kohavi et al. 1997), including decision trees, Naive-Bayes (evidence), column importance, and automatic (entropy-based) discretization. Support for unknown (null values) was added, as well as support for session management (save/restore), a batch mode, integration with the desktop environment (icon launching), and the ability to define new columns
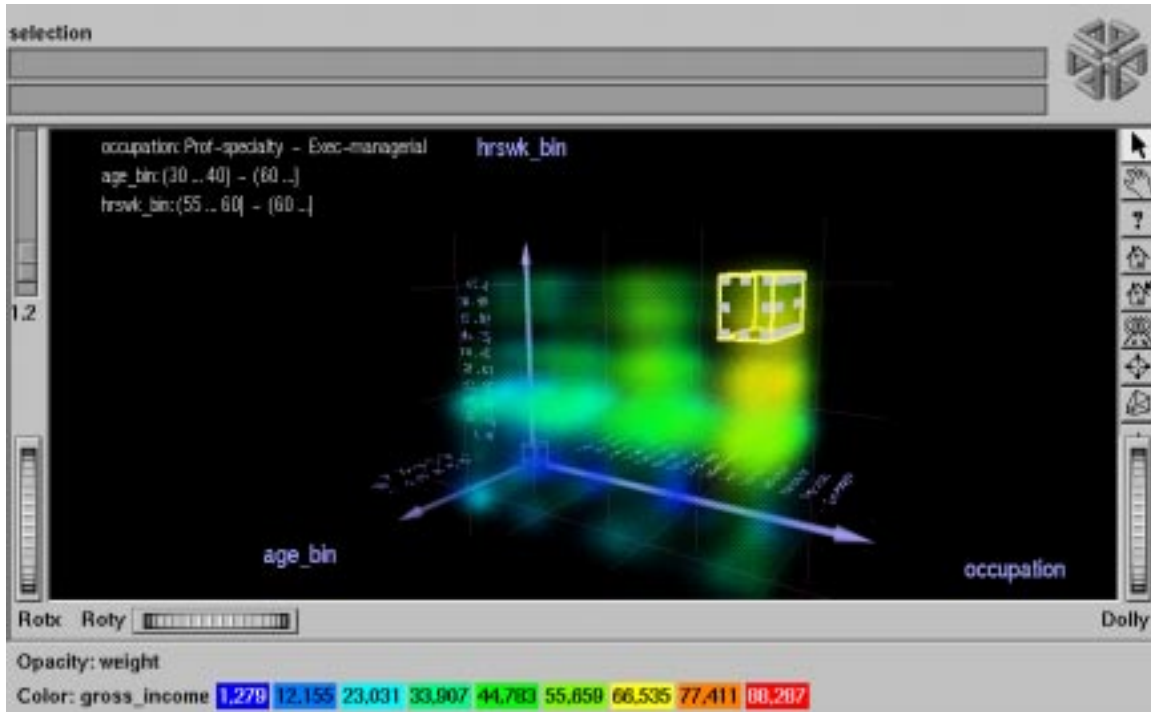
Figure D2.2.5.2: A Splat Visualizer view of census data on adults working in the US. The plot shows how gross income is affected by age, occupation, and the number of hours worked per week. The density of each splat represents the number of people, its color represents the average gross income. The selected cube represents people over 30, who work over 55 hours per week in a professional specialty or an executive managerial position.

Figure D2.2.5.3: Visualization of a decision table for the hypothyroid database (left). The figure shows the top-level view with two attributes: FTI and TSH. Users can see that several intersections are empty: high TSH values imply unknown FTI (probably not measured), and that most data is in the low range of TSH (below 6.3). High values for FTI (above 64.5) are negative hypothyroid with high probability (dark gray). The interesting intersections are for low FTI and high TSH. MineSet's map visualizer (right) showing refinancing costs, mapped to height, for every US county based on FIPS codes. Deviations from each state's average are colored from blue (zero deviation) to yellow (0.005) to red (0.01).

with expressions. MineSet 1.2 added web launching capabilities from machines that have MineSet installed.

MineSet 2.0 added drill-through, the Splat Visualizer, Statistics Visualizer, Record Viewer, binary file format, sampling, Option Trees, loss matrices, learning curves, probability estimates from classifiers, and backfitting of data.

MineSet 2.5 added boosting of classifiers, parallelization, clustering, regression trees, and decision tables. Support for multi-byte characters for internationalization and 64-bit support for large memory models was added in MineSet 2.6. Also added was a Java-based record viewer and a plug-in architecture for adding new data mining tools.

As of summer of 1998, the engineering effort in product development is estimated at over 55 person years, with the engineering team consisting of 18 people.

D2.2.5.7 Commercial Uses

MineSet has been used commercially since it released in 1996, but most customers are reluctant to publicize the exact uses as they perceive data mining as a competitive advantage. As of summer of 1998, there are several hundred commercial sites using MineSet and close to a thousand universities. In this section we mention a very restricted set of commercial uses. More information can be found in Adhikari (1998).

Incyte Pharmaceuticals (www.incyte.com) provides genomic technologies to aid in the understanding of the molecular basis of diseases. Incyte created the LifeSeq® 3D software to give scientists powerful visualization tools for sifting through the vast amounts of genomic data in the LifeSeq database (Incyte Pharmaceuticals 1997, Incyte Pharmaceuticals 1998). LifeSeq 3D is based on MineSet, displaying genomic information as interactive, multidimensional graphics, enabling scientists to easily navigate large data sets and uncover hidden relationships and important trends in gene expression.

Risk Monitors conducts statistical analyses of loan and mortgage data nationwide, building the models that mortgage servicers and banks rely on to calculate their underwriting risks. It analyzes 11 million loans nationwide dating back to 1989, and applies up to 200 variables to them during statistical analyses. Typical analysis was

previously done in group, or cohorts. With MineSet, Risk Monitors was able to work with loans at the individual level. More details are available at Goodarzi, Kohavi, Harmon & Senkut (1998).

Procter & Gamble Co.'s health-care division uses MineSet for clinical trials and efficacy tests of over-the-counter drugs. On average, the total cost to bring a pharmaceutical drug to market is approximately $500 million, making it important to effectively run clinical trials and process the data effectively. Using data mining, scientists try to find an active molecule in a drug, making sure it works, testing it for harmful side effects, and eventually testing it on humans. More information can be found at Stevens (1998).

D2.2.5.8 Conclusion

MineSet provides a set of scalable analytical mining algorithms for identifying interesting patterns in data. MineSet also provides a rich selection of visualization techniques that help make these patterns understandable. But, the primary feature that differentiates MineSet from other KDD tools is the integrated environment in which these algorithms and techniques are combined.

References

Adhikari, R. (1998), 'Data mining muscle', *Information Week* pp. 65–67.
http://www.informationweek.com/695/95iudat.htm.

Becker, B. (1997), Volume rendering for relational data, *in* 'Proceedings of Information Visualization', IEEE Computer Society, pp. 87–90.

Becker, B., Kohavi, R. & Sommerfield, D. (1997), Visualizing the simple bayesian classifier, *in* 'KDD Workshop on Issues in the Integration of Data Mining and Data Visualization'.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth International Group.

Brunk, C., Kelly, J. & Kohavi, R. (1997), MineSet: an integrated system for data mining, *in* D. Heckerman, H. Mannila, D. Pregibon & R. Uthurusamy, eds, 'Proceedings of the third international conference on Knowledge Discovery and

Data Mining', AAAI Press, pp. 135–138.
`http://mineset.sgi.com`.

Cheeseman et al. (1988), AutoClass: a Bayesian classification system, *in* 'Proceedings of the Fifth International Conference on Machine Learning', Morgan Kaufmann, pp. 54–64. Also appears in Readings in *Machine Learning* by Shavlik and Dietterich.

Dasarathy, B. V. (1990), *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press, Los Alamitos, California.

Domingos, P. & Pazzani, M. (1997), 'Beyond independence: Conditions for the optimality of the simple Bayesian classifier', *Machine Learning* **29**(2/3), 103–130.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'The KDD process for extracting useful knowledge from volumes of data', *Communications of the ACM* **39**(11), 27–34.

Goodarzi, A., Kohavi, R., Harmon, R. & Senkut, A. (1998), Loan prepayment modeling, *in* T. H. Hann & G. Nakhaeizadeh, eds, 'KDD Workshop on Data Mining in Finance', pp. 62–69.

Incyte Pharmaceuticals (1997), Incyte releases LifeTools 3D.
`http://www.incyte.com/news/1997/PR9712-LT3D.html`.

Incyte Pharmaceuticals (1998), LifeSeq 3D: Data mining and visualization software.
`http://www.incyte.com/products/lifeseq/lifeseq3d.html`.

Kohavi, R. & John, G. H. (1997), 'Wrappers for feature subset selection', *Artificial Intelligence* **97**(1-2), 273–324.
`http://robotics.stanford.edu/users/ronnyk`.

Kohavi, R. & Kunz, C. (1997), Option decision trees with majority votes, *in* D. Fisher, ed., 'Machine Learning: Proceedings of the Fourteenth International Conference', Morgan Kaufmann Publishers, Inc., pp. 161–169.
`http://robotics.stanford.edu/users/ronnyk`.

Kohavi, R. & Sommerfield, D. (1998), Targeting business users with decision table classifiers, *in* R. Agrawal, P. Stolorz & G. Piatetsky-Shapiro, eds, 'Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining', AAAI Press, pp. 249–253.

Kohavi, R., Sommerfield, D. & Dougherty, J. (1997), 'Data mining using $\mathcal{MLC}$++: A machine learning library in C++', *International Journal on Artificial Intelligence Tools* **6**(4), 537–566.
http://www.sgi.com/Technology/mlc.

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

Silicon Graphics (1998), *MineSet User's Guide*, Silicon Graphics, Inc. http://mineset.sgi.com.

Srikand, R. & Agrawal, R. (1995), Mining generalized association rules, *in* 'Proceedings of the 21st International Conference on Very Large Databases'.

Stevens, D. (1998), 'Mineset's data visualization enhances clinical studies at procter & gamble', *DM Review* **8**(7), 135.
http://www.dmreview.com/issues/1998/jul/reviews/jul98_135.htm.