DS Datamining Co., LLC
8939 S. Sepulveda Blvd
Silicon Beach, CA 90045

https://sciprohawk.com/
david.stephens@dsscientific.com

DATA MINING FOR ENGINEERS AND SCIENTISTS

SUMMARY

This white paper is to introduce the concept of data mining for scientific and engineering applications. It's often observed in many fields that mining for data is performed on a regular basis with very little after thought in the discipline. So, what is data mining and how does it relate to engineering and scientific disciplines? Data mining is the practice of examining large databases to generate new information. This may be a very generalized definition, but it is very factual and performed by professionals in industry without any afterthought. In this white paper, we will introduce the SciPro Hawk™1 data mining software API, which is developed for engineering and scientific applications. The white paper explains the important role data mining offers in the analytical discovery process and improving productivity. This allows the engineering organizations to make informed decisions not on how much data the organization possesses, but in the insight that the data provides. High technology corporations and businesses from across every major industry are using data mining as a competitive enhancer to manage and eliminate risk, anticipate resource demands, increase response rates for marketing campaigns, and resolve technical challenges.

INTRODUCTION

Data mining is the process of discovering patterns in large data sets and is an interdisciplinary subfield of computer science. The purpose of datamining is to extract information from a data set and transform it into a meaningful representation for the end user to comprehend. Data mining involves several steps, aside from the raw analysis, which include databases, data management, pre-processing, modeling, post processing, visualizations, metrics, etc. Engineers and scientists utilize the term in a misleading way. The purpose is the extraction of patterns and knowledge from the data and not the extraction of the data itself.

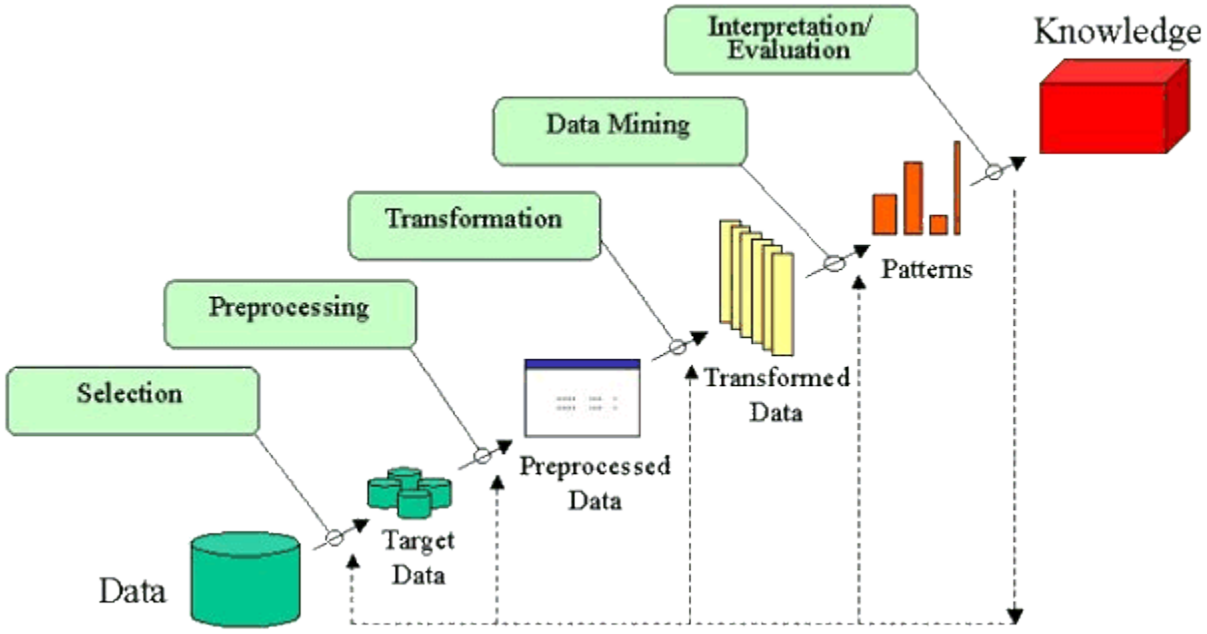
The extraction of patterns from data has occurred for many centuries including Bayes' theorem (1700s) and regression analysis (1800s). The advent of computer technology has dramatically increased data collection, storage, and manipulation capability. As data sets have grown in complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets. The main research began in the 1990's (1995) with the First International Conference on Data Mining and Knowledge Discovery (KDD-95) in Montreal under AAAI sponsorship.

The SciPro Hawk™1 data mining software bridge the gap between traditional data mining software and engineering tool for data analysis. The tool is integrated into excel providing the engineering community with a user-friendly API between the richness of the excel platform and the data set.

PROCESS

The process of data mining is separated into five steps:

1. Selection
2. Pre-processing
3. Transformation
4. Data mining
5. Interpretation/evaluation.



Before the data set can be utilized, the data must be pre-processed by assembling the raw data into a meaningful organized set. As data mining can only uncover patterns present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. The pre-processing allows the data to be pre-analyzed and cleaned of any noise or unwanted distributions/missing data. After the pre-processing of the data set has occurred the data mining portion can begin.

Data mining involves six common classes of tasks:

1. Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
2. Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
3. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
5. Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
6. Summarization – providing a more compact representation of the data set, including visualization and report generation.

Now we need to define what is data. Data is facts, numbers, images, and text, that can be processed by a computer. In general, data mining can be applied to any data set if the data is meaningful for the end user. An algorithm is then applied to the data set. The data mining pattern produced by the algorithm is then analyzed for any meaningful patterns.

The SciPro Hawk™1 can import four file types csv, txt, rtf, and log commonly utilized by engineers and scientist from a networked directory into individual workbooks in Excel and in one process delimit the data for pre-processing can be initiated.

DATA SYSTEMS

A data system is the creation and storage of the data. In engineering and scientific disciplines this is the servers that the data is store in and for the data to be analyzed properly. The access to the data is generally stored in servers, storage devices such as portable hard drives, and on local hard drives. The ability to access the data in a timely manner to process data distinguishes data mining and the data set from all other disciplines.

In this section we discuss how and where the data should be stored. The reason on how is to standardize the data set for example engineering hardware stores data in csv formats such as Agilent oscilloscopes and network analyzers. While other data sets are stored in text or log files. Industry standardizes the data set types to process the data in a timelier manner. The ability to handle multiple data sets adds extended flexibility to the data analysis tools set and allows the profession to work across disciplines.

FUNCTIONS

Data sets require to be placed in a standard tabulated form for statistical analysis to be performed. The basic statistical analysis will be discussed which include mean, standard deviation, mode, minimum, maximum, and uncertainty. These are the basic function widely utilized by engineers and scientist to analyze data sets.

In data mining measuring the center tendency of data is performed with the arithmetic mean. Let x1, x2, x3, etc. be a set of values like a voltage measurement. Then the mean is for the set of values

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

The mode which is the most frequently occurring number in a group of numbers. For example, the mode of 2, 3, 3, 5, 7, and 10 is 3. A mode value can be found simply by counting the number of times each value occurs.

The standard deviation is a measure of a data's dispersion. The variance of N observations, x1, x2, …., xN, for a numeric attribute X is

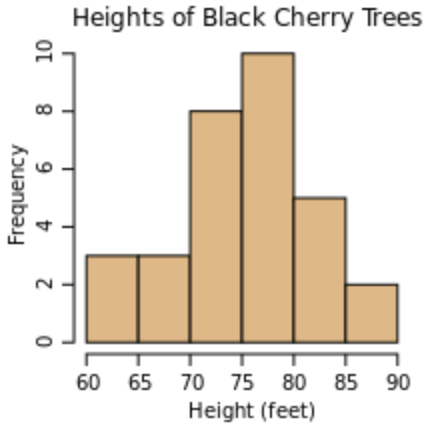
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2,$$

where ẋ is the mean value of the observations. The standard deviation, σ, of the observations is the square root of the variance, σ^2.

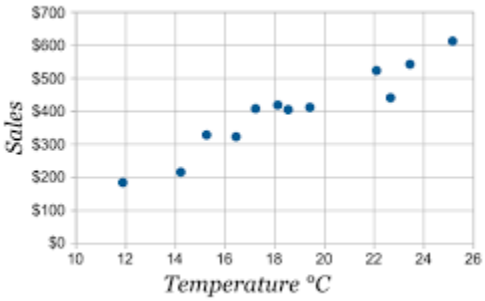
The minimum is the smallest value from a supplied set of numeric values and the maximum is the largest value from a supplied set of numeric values.

The uncertainty is amount of error associated with the data. In the context of the SciPro Hawk™1 the uncertainty is calculated by calculating the standard deviation and dividing by the mean U = σ/ ẋ.

Visualization of the data gives the data meaning. In data mining there are three basic graphing types linear, scatter, and histogram to display the data. Histogram plots is a summary of an accurate graphical representation of the distribution of numerical data, X. To construct a histogram, the range of values of X are divided into subsets referred to as bins. The range of each bin is the range and are equally distributed. For each subrange(bin), a bar is drawn with a height that represents the total count of items observed within the subrange.

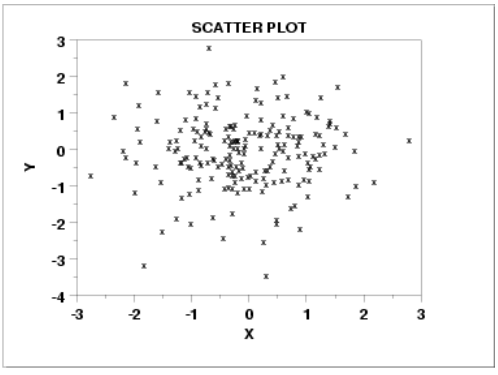


A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation.

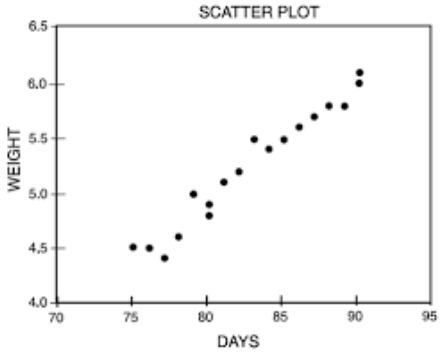


A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes, hence their correlation. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

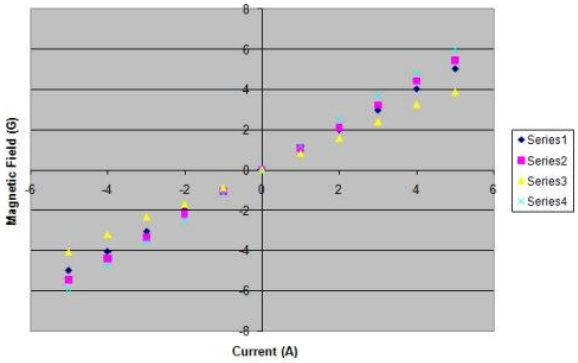
Low Correlation Scatter Plot



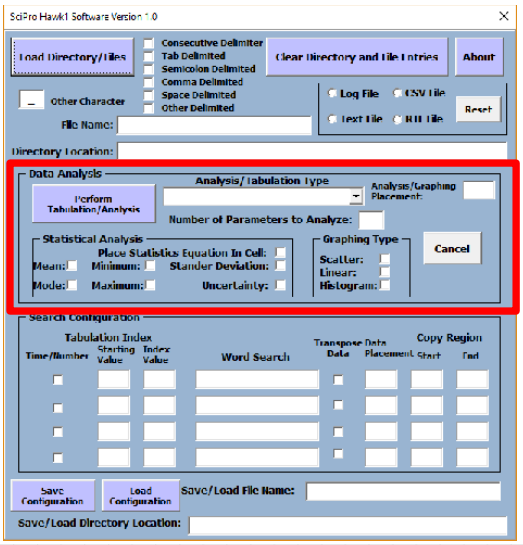
High Correlation Scatter Plot

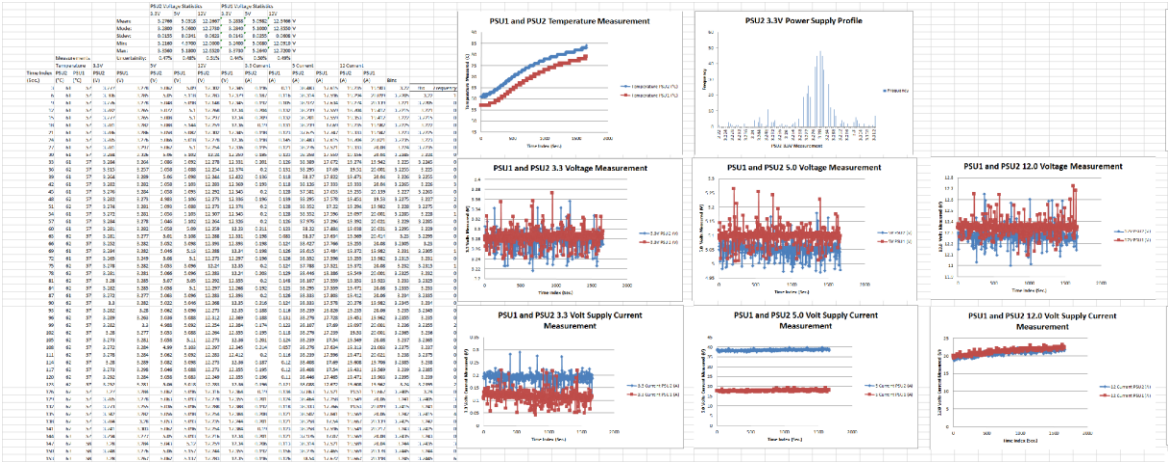


A linear plot is similar to a scatter plot, but line charts are used to display trends over time. A series of points, discrete or continuous, as in forming a curve or surface, each of which represents a value of a given function.



The SciPro Hawk™1 engineering analysis tool offers all the basic analysis types including all three graphing methods. The tool can display all the analysis and graphs in a worksheet through the user GUI interface provided for the user.





CONCLUSION

Data mining for engineers and scientists is a growing industry. As data sets become more complex and networked allowing engineers to have access to the data for analysis and design debugging becomes even more paramount for product time to market. Engineers and scientist often struggle in finding ways to import data from scientific instruments to a more standardized and usable form. The SciPro Hawk™1 software which is integrated into excel will give engineers and scientists the flexibility to standardize the data with excel as the main foundation due to excels wide distribution.

REFERENCES

1. https://en.wikipedia.org/wiki/Data_mining.
2. http://storm.cis.fordham.edu/~yli/documents/CISC4631Spring17/Chapter1_Intr.pdf
3. http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
4. Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, 3rd Edition.
5. https://en.wikipedia.org/wiki/Scatter_plot.