1.1 Modelling Technique Selection

Data mining modelling techniques are used in metabolomics either in a hypothesis-driven or in a data-driven fashion. While hypothesis-driven objectives are generally motivated by the goals of the research investigation and the aims of its subsequent studies, data-driven objectives are performed based on the goals of data mining and knowledge discovery e.g. finding novel, but interesting results (Taylor et al., 2008). When selecting data mining techniques, it is crucial to understand data mining approaches, goals and tasks as well as the techniques they use to achieve their modelling objectives.

1.1.1 Data Mining Goals and Tasks

Data mining goals and tasks are related to the reason of applying data mining and knowledge discovery techniques and the purpose it seek to achieve. Discovery-oriented data mining aims find previously unknown phenomena in the data through prediction and description, while verification-oriented data mining aims to verify an existing or know phenomena implied in the data through description and hypothesis testing. Prediction goals of data mining can be achieved using regression, classification and rules induction tasks, while descriptive goals can be carried out using segmentation, association, dimensionality reduction, correlation, features extraction and analysis tasks. verification goals can be performed using descriptive and hypothesis testing tasks (Goodacre, 2006, Maimon and Rokach, 2005a). In order to achieve its goals and perform their tasks, data mining employs a wide spectrum of machine learning, statistical and pattern recognition techniques. Fig x.2 illustrates data mining goals and tasks.

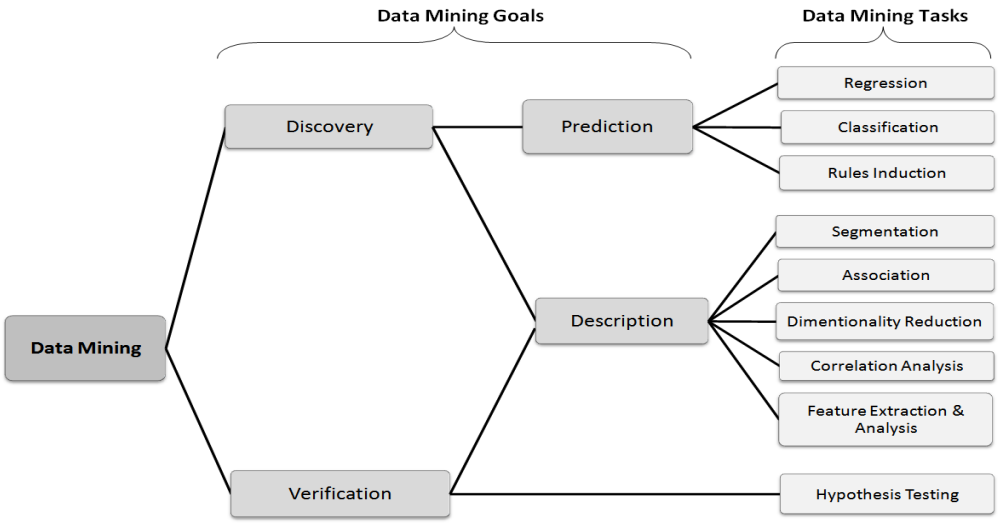

Fig. data mining goals and tasks

1.1.2 **Supervised vs. Unsupervised Learning**

Supervised methods learn through finding a model that represent association between inputs (X variables or predictors) which are typically the metadata of the study with the outcomes(Y variables or responses) which are typically the assay results e.g. classification, regression etc. *Unsupervised methods* learn from data through finding patterns or groups within the inputs (X variables) and is performed with no such guidance e.g. segmentation or data reduction. In metabolomics, the inputs represent the data set, while outcomes represent the traits or classes (Goodacre et al., 2004).

Table x.1 describes those tasks and provides examples of their modelling techniques showing whether it is performed in supervised or unsupervised fashion. This is usually performed either through prediction or description (Maimon and Rokach, 2005a, Maimon and Rokach, 2005b, Sumathi and Sivanandam, 2006) e.g. predicting biomarkers for a disease or classifying samples into healthy and diseased. Alternative techniques might be useful to see the results from different perspectives or to propagate new questions to be answered or even to seek explanations for results. On the other hand combining more than one technique might be useful to tackle the weakness or to enhance the selected technique.

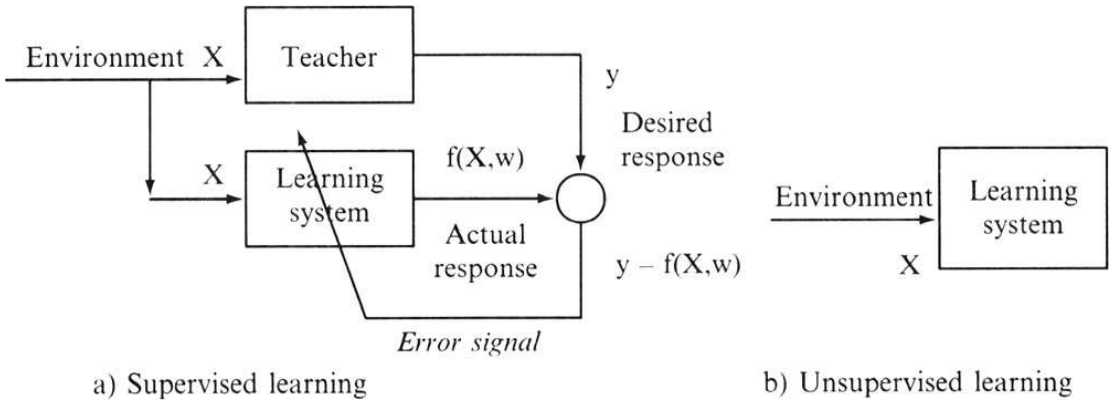

Fig. Supervised learning vs. unsupervised learning (Kantardzic, 2003)

Data Mining Task	Description	Data Mining Example Techniques	
		Supervised	Unsupervised
Regression	Build a model that use data to predict new continuous numerical data(Chatterjee and Hadi, 2006); Predict the dependent variable (response) from independent variables (Predictors) (Antoniewicz et al., 2006).	Multiple Linear Regression(MLR); Partial Least Square (PLS) (Bryan et al., 2008, Hayashi et al., 2009); Support Vector Machine (SVM) (Truong et al., 2004, Yao et al., 2004); Linear Regression (LR) (Sanchez et al., 2008, Zou et al., 2003); Regression Trees (Hollywood et al., 2006); Principle Component Regression (PCR) (Antoniewicz et al., 2006).	
Classification	Build a model that is capable of classifying data in order to predict new discreet or categorical data.	Artificial Neural Networks(ANN) (Goodacre et al., 2004), Decision Trees, Random Forest (Enot et al., 2008), Linear Discriminant Analysis (LDA), Discriminant Function Analysis (DFA) (Ye et al., 2004, Enot et al., 2008, Lindon et al., 2001),Support Vector Machine (SVM) (Yao et al., 2004), Soft Independent Modelling of Class Analogy (SIMCA) (Lindon et al., 2001),Genetic Programming (Brown et al., 2005), Genetic Algorithm (Johnson et al., 2003)	Kohonen Neural Networks Self Organising Map (SOM) Cluster Analysis Techniques (Steuer et al., 2007, Sumner et al., 2003).
Rules **Inductive**	Extract useful rules from data set based on significance.	Genetic Programming, Genetic Algorithm Classification and Regression Trees (CART), Inductive Logic Programming (Goodacre et al., 2004, Goodacre, 2007, Goodacre, 2005).	
Segmentation	Identify the natural grouping among the data set and classify the data accordingly.	Discriminant Function Analysis (DFA)(Ye et al., 2004, Enot et al., 2008, Lindon et al., 2001) Genetic Programming (Brown et al., 2005, Kell, 2002), Genetic Algorithm(Johnson et al., 2003).	Hierarchical Clustering Analysis (HCA) (Lindon et al., 2001, Sumner et al., 2003) , K-Means (Miroslava et al., 2009, Steuer, 2006) , fuzzy c-means (Li et al., 2009) Self Organising Map (SOM)(Steuer, 2006)
Association	Identify the relationships within the data set and the probability of their occurrence		Association Rules (Thakkar et al., 2007, Hipp et al., 2002, Agrawal et al., 1993, Gupta and Agrawal, 2009) , e.g. algorithms APRIORY(Osl et al., 2008), GRI, CARMA, etc.
Dimensionality Reduction	Create an optimised data set on which to base a model and eliminating non-informative features	Linear Discriminant Analysis (LDA) (Bryan et al., 2008), Partial Least Squares(PLS) (Yamamoto et al., 2009) , Discriminant Analysis (PLS-DA) (Kim et al., 2007, Bryan et al., 2008) Orthonornalized Partial Least Squares (OPLS) (Yamamoto et al., 2009).	Independent Component Analysis (ICA)(Scholz et al., 2004, Scholz and Selbig, 2006) Principle Component Analysis (PCA)(Yamamoto et al., 2009) Factor Analysis (FA)(Steuer, 2006).
Feature Extraction & Analysis	Gain insight into the rationale underlying class divisions(Bryan et al., 2008).	Partial Least Squares Discriminant Analysis (PLS-DA), Random Forest feature selection (Bryan et al., 2008).	
Correlation Analysis	Determine the association between the changes in the value of one variable with the changes in another variable.	Covariance Analysis (Zou et al., 2003, Mendes, 2002, Goodacre et al., 2007).	
Hypothesis Testing	Test assertion about the data set based on the concept of proof by contradiction	chi-test, z-test, f-test, Goodness of fit, Analysis of Variance (ANOVA) (Steuer, 2006), Multivariate analysis of variance (MANOVA)(Johnson et al., 2007)	

AGRAWAL, R., IMIELISKI, T. & SWAMI, A. Year. Mining association rules between sets of items in large databases. *In:* BUNEMAN, P. & JAJODIA, S., eds. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, May 25 - 28, 1993 1993 Washington, D.C.: ACM, 207-216.

ANTONIEWICZ, M. R., STEPHANOPOULOS, G. & KELLEHER, J. K. 2006. Evaluation of regression models in metabolic physiology: predicting fluxes from isotopic data without knowledge of the pathway *Metabolomics,* 2**,** 41-52.

BROWN, M., DUNN, W. B., ELLIS, D. I., GOODACRE, R., HANDL, J., KNOWLES, J. D., O'HAGAN, S., SPASIĆ, I. & KELL, D. B. 2005. A metabolome pipeline: from concept to data to knowledge. *Metabolomics,* 1**,** 39-51.

BRYAN, K., BRENNAN, L. & CUNNINGHAM, P. 2008. MetaFIND: A feature analysis tool for metabolomics data. *BMC Bioinformatics,* 9**,** 470.

CHATTERJEE, S. & HADI, A. S. 2006. *Regression analysis by example*, Hoboken, N.J., Wiley-Interscience.

ENOT, D. P., LIN, W., BECKMANN, M., PARKER, D., OVERY, D. P. & DRAPER, J. 2008. Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data. *Nat. Protocols,* 3**,** 446-470.

GOODACRE, R. 2005. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *J. Exp. Bot.,* 56**,** 245-254.

GOODACRE, R. 2006. Data Analsysis Standards in metabolomics. Mancheter: University of Mancheter.

GOODACRE, R. 2007. Metabolomics of a Superorganism. *Journal of Nutrition,* 137**,** 259-266.

GOODACRE, R., BROADHURST, D., SMILDE, A., KRISTAL, B., BAKER, J., BEGER, R., BESSANT, C., CONNOR, S., CAPUANI, G., CRAIG, A., EBBELS, T., KELL, D., MANETTI, C., NEWTON, J., PATERNOSTRO, G., SOMORJAI, R., SJÖSTRÖM, M., TRYGG, J. & WULFERT, F. 2007. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics,* 3**,** 231-241.

GOODACRE, R., VAIDYANATHAN, S., DUNN, W. B., HARRIGAN, G. G. & KELL, D. B. 2004. Metabolomics By Numbers: Acquiring Understanding Global Metabolite Data. *Trends in Biotechnology*, 22**,** 245-252.

GUPTA, R. K. & AGRAWAL, D. P. 2009. Improving the Performance of Association Rule Mining Algorithms by Filtering Insignificant Transactions Dynamically. *Asian Journal of Information Management,* 3**,** 7-17.

HAYASHI, S., AKIYAMA, S., TAMARU, Y., TAKEDA, Y., FUJIWARA, T., INOUE, K., KOBAYASHI, A., MAEGAWA, S. & FUKUSAKI, E. 2009. A novel application of metabolomics in vertebrate development. *Biochemical and Biophysical Research Communications,* 386**,** 268-272.

HIPP, J., GÜNTZER, U. & NAKHAEIZADEH, G. 2002. Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. *In:* PERNER, P. (ed.) *Advances in Data Mining.* Berlin/Heidelberg: Springer.

HOLLYWOOD, K., BRISON, D. R. & GOODACRE, R. 2006. Metabolomics: Current technologies and future trends. *Proteomics,* 6**,** 4716-4723.

JOHNSON, H., LLOYD, A., MUR, L., SMITH, A. & CAUSTON, D. 2007. The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. *Metabolomics,* 3**,** 517-530.

JOHNSON, H. E., BROADHURST, D., GOODACRE, R. & SMITH, A. R. 2003. Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry,* 62**,** 919-928.

KANTARDZIC, M. 2003. *Data mining : concepts, models, methods, and algorithms,* Hoboken, NJ, Wiley-Interscience : IEEE Press.

KELL, D. B. 2002. Genotype-phenotype mapping: genes as computer programs. *Trends in Genetics,* 18**,** 555-559.

KIM, Y., PARK, I. & LEE, D. Year. Integrated Data Mining Strategy for Effective Metabolomic Data Analysis. *In:* XIANG-SUN ZHANG, L. C., LING-YUN WU AND YONG WANG, ed. Optimization and Systems Biology, The First International Symposium, OSB'07, 8-10/8/2007 2007 Beijing, China. ORSC & APORC, 45-51.

LI, X., LU, X., TIAN, J., GAO, P., KONG, H. & XU, G. 2009. Application of Fuzzy c-Means Clustering in Data Analysis of Metabolomics. *Analytical Chemistry,* 81**,** 4468-4475.

LINDON, J. C., HOLMES, E. & NICHOLSON, J. K. 2001. Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy,* 39**,** 1-40.

MAIMON, O. & ROKACH, L. 2005a. *Data Mining and Knowledge Discovery Handbook,* New York, NY, Springer.

MAIMON, O. & ROKACH, L. 2005b. *Decomposition methodology for knowledge discovery and data mining: theory and applications,* NJ/London, World Scientific.

MENDES, P. 2002. Emerging bioinformatics for the metabolome. *Brief Bioinform,* 3**,** 134-145.

MIROSLAVA, CCARON, UPERLOVI, CACUTE, -CULF, BELACEL, N., CULF, A. S., CHUTE, I. C., OUELLETTE, R. J., BURTON, I. W., KARAKACH, T. K. & WALTER, J. A. 2009. NMR metabolic analysis of samples using fuzzy K-means clustering. *Magnetic Resonance in Chemistry,* 47**,** S96-S104.

OSL, M., DREISEITL, S., PFEIFER, B., WEINBERGER, K., KLOCKER, H., BARTSCH, G., SCHAFER, G., TILG, B., GRABER, A. & BAUMGARTNER, C. 2008. A new rule-based algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry. *Bioinformatics,* 24**,** 2908-2914.

SANCHEZ, D. H., REDESTIG, H., KRAMER, U., UDVARDI, M. K. & KOPKA, J. 2008. Metabolome-ionome-biomass interactions: What can we learn about salt stress by multiparallel phenotyping? *Plant Signal Behav,* 3**,** 598-600.

SCHOLZ, M., GATZEK, S., STERLING, A., FIEHN, O. & SELBIG, J. 2004. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics,* 20**,** 2447-2454.

SCHOLZ, M. & SELBIG, J. 2006. Visualization and Analysis of Molecular Data. *In:* WECKWERT, W. (ed.) *Metabolomics: Methods and Protocols.* Totowa, NJ: Humana Press.

STEUER, R. 2006. Review: On the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform,* 7**,** 151-158.

STEUER, R., MORGENTHAL, K., WECKWERTH, W. & SELBIG, J. 2007. A Gentle Guide to the Analysis of Metabolomic Data. *In:* WECKWERTH, W. (ed.) *Metabolomics: Methods and Protocols.* Totowa, NJ.

SUMATHI, S. & SIVANANDAM, S. N. 2006. Data Mining Tasks, Techniques, and Applications. *In:* SUMATHI, S. (ed.) *Introduction to Data Mining and its Applications.* Berlin/NY: Springer.

SUMNER, L. W., MENDES, P. & DIXON, R. A. 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry,* 62**,** 817-836.

TAYLOR, C. F., FIELD, D., SANSONE, S., AERTS, J., APWEILER, R., ASHBURNER, M., BALL, C. A., BINZ, P.-A., BOGUE, M., BOOTH, T., BRAZMA, A., BRINKMAN, R. R., MICHAEL CLARK, A., DEUTSCH, E. W., FIEHN, O., FOSTEL, J., GHAZAL, P., GIBSON, F., GRAY, T., GRIMES, G., HANCOCK, J. M., HARDY, N. W., HERMJAKOB, H., JULIAN, R. K., KANE, M., KETTNER, C., KINSINGER, C., KOLKER, E., KUIPER, M., NOVERE, N. L., LEEBENS-MACK, J., LEWIS, S. E., LORD, P., MALLON, A.-M., MARTHANDAN, N., MASUYA, H., MCNALLY, R., MEHRLE, A., MORRISON, N., ORCHARD, S., QUACKENBUSH, J., REECY, J. M., ROBERTSON, D. G., ROCCA-SERRA, P., RODRIGUEZ, H., ROSENFELDER, H., SANTOYO-LOPEZ, J., SCHEUERMANN, R. H., SCHOBER, D., SMITH, B., SNAPE, J., STOECKERT, C. J., TIPTON, K., STERK, P., UNTERGASSER, A., VANDESOMPELE, J. & WIEMANN, S. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology,* 26**,** 889-896.

THAKKAR, D., RUIZ, C. & RYDER, E. F. Year. Hypothesis-Driven Specialization of Gene Expression Association Rules. *In:* HU, X., MANDOIU, I., OBRADOVIC, Z. & XIA, J., eds. Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine, 2-4 November 2007 2007 Fremont, CA. Los Alamitos, CA: IEEE Computer Society, 48-55.

TRUONG, Y., LIN, X. & BEECHER, C. Year. Learning a complex metabolomic dataset using random forests and support vector machines. *In:* KOHAVI, R., GEHRKE, J., DUMOUCHEL, W. & GHOSH, J., eds. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22 - 25, 2004 2004 Seattle, WA. New York, NY: ACM, 835 - 840.

YAMAMOTO, H., YAMAJI, H., ABE, Y., HARADA, K., WALUYO, D., FUKUSAKI, E., KONDO, A., OHNO, H. & FUKUDA, H. 2009. Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. *Chemometrics and Intelligent Laboratory Systems,* 98**,** 136-142.

YAO, X. J., PANAYE, A., DOUCET, J. P., ZHANG, R. S., CHEN, H. F., LIU, M. C., HU, Z. D. & FAN, B. T. 2004. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *Journal of Chemical Information and Computer Sciences,* 44**,** 1257-1266.

YE, J., JANARDAN, R., LI, Q. & PARK, H. Year. Feature extraction via generalized uncorrelated linear discriminant analysis. *In:* The twenty-first international conference on Machine learning, 04 - 08 July 2004 2004 Banff, Alberta. Los Alamitos, CA: ACM, New York, NY, 895-902.

ZOU, K. H., TUNCALI, K. & SILVERMAN, S. G. 2003. Correlation and simple linear regression. *Radiology,* 227**,** 622.