ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 1, pp: (55-60), Month: April 2015 – September 2015, Available at: www.paperpublications.org

Supervised Multi Attribute Gene Manipulation For Cancer

Shenbagam.S[1], S.Brintha Rajakumari[2]

[1]PG Student, [2]Assistant Professor, Department of CSE, Bharath University, Chennai

Abstract: **Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.**

Keywords: **Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions, Supervised Multi Attribute Gene.**

1. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Web services have been promising in recent years and are by now one of the most popular techniques for building distributed systems. Service-oriented systems can be built efficiently by dynamically composing different web services, which are provided by other organizations. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

2. RELATED WORK

The cause of many cancers remains unknown. Apart from internal (genetic) causes, there are certain environmental and external factors too that participate in cancer formation within an organism, viz. environmental toxins, adulterated food intake, air pollution, and irregular lifestyle; (share as depicted in the Fig. 1). These can be categorized under epigenetics. Epigenetics is an un ignorable issue to be addressed by the biomedical community. Symptoms of cancer depend on the type and location of the cancer. For example, lung cancer can cause coughing, heavy breathing, chest pain, etc. Colon cancer often causes diarrhoea, constipation, dysentery, and blood in the stool.

2.1 MATHEMATICS IN NATURE: AN INTUITIVE CERTITUDE:

Mathematics is known to be an indispensable part all sciences. It forms the edifice of all existences being a formidable aegis that "holds" all parts together. One can have certain propensity towards it and more when going through

John A. Adam's texts Docile is the symmetry found in vegetation, anatomy of living creatures, shapes of heavenly bodies: planets being spherical and much so their orbits, to name a few. Mathematical modeling of any natural phenomena and its inherent dogma at core, which ensures it compositional as well as physical traits can be extremely expedient towards developing an understanding. A mathematical model is a feat if it fits the known data and makes accurate predictions for the future, as rendered in the fig.A snowball is defined to grow in size and attain an almost intermediary shape between a circle and a sphere as it rolls through ice. However, external factors like intensity of sunlight, heat produced due to friction/resistance dymanic surface area of the ball, etc., are the variates that contribute to the problem, profoundly (Adam, 2006). The author also registers that all mathematical models are flawed to some extent owing to the inappropriate presumptions made during their construction. The aesthetic of all natural and physical phenomena is brought to life once the mathematical undergird is realized."Mathematics is to nature as Sherlock Holmes is to evidence."With a compendium of suggestions to consult, it renders highly probabilistic scenario that the gene expression datawon't be an exception. Studies have shown that mathematical and statistical models can be built around it and they are seminal to identify biomarkers. The only apprehension is the uncertainity attached to it which is subjected to validation to establish the baseline parameters.

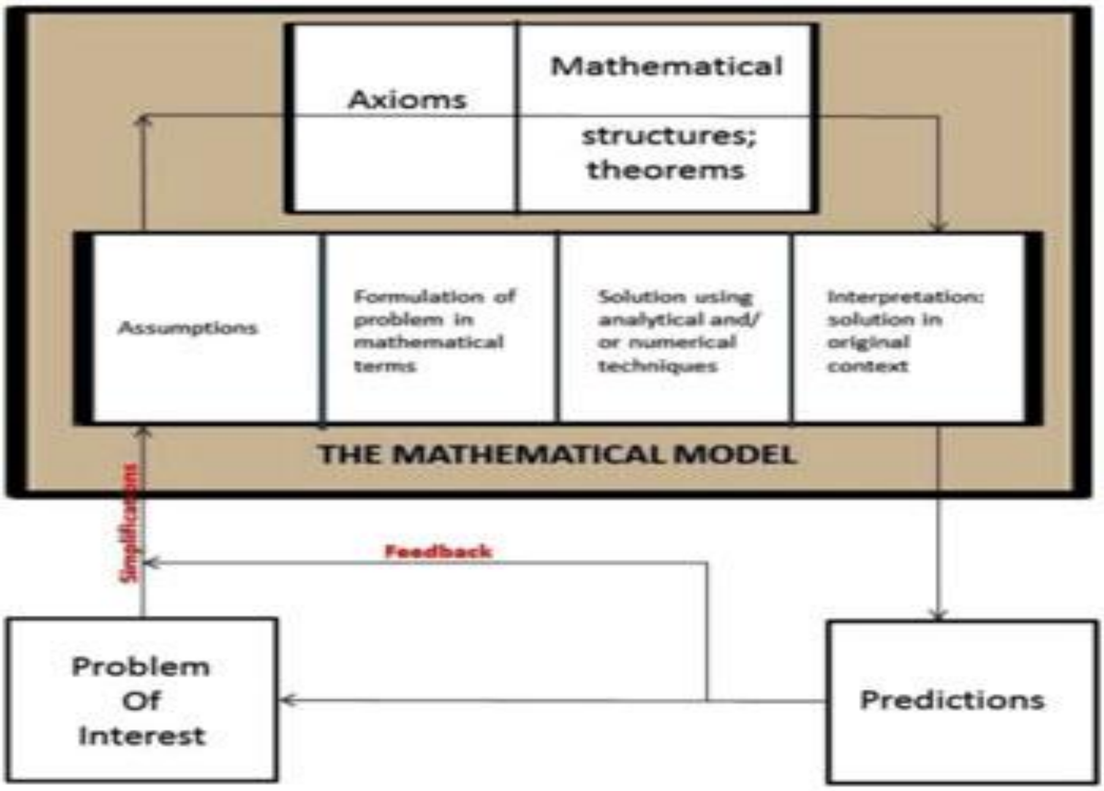

Axioms
Mathematical
structures; theorems
Assumptions
Formulation of problem in mathematical terms
Solution using analytical and/ or numerical techniques
Interpretation: solution in original context
THE MATHEMATICAL MODEL
Simplifications
Feedback
Problem Of Interest
Predictions

Fig. 2.1 MATHEMATICAL MODEL

2.2 EPIGENETICS:

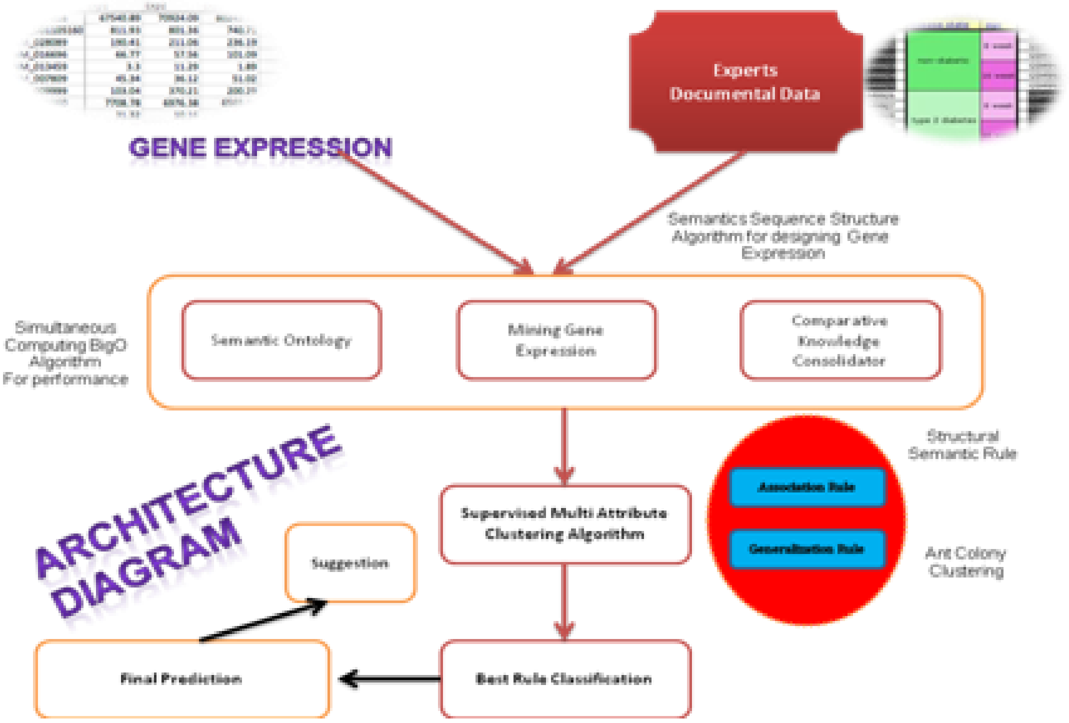
Literally speaking, "epi" stands for "on top of" and epigenetics is on top of genetics. Governing factors for the gene expression and protein building, explicit to the inherent DNA code, delineates epigenetics. The environment and our lifestyle can significantly direct our genetic behaviour and even that of our kids. The multi cellular organisms have optimally identical underlying code, yet they have incongruent phenotypes.

3. SYSTEM MODEL

In this section, our system model is designed with two phases. In Phase 1 consists of analysis and design is considered and in Phase 2 the model is implemented and validated. The Initial analysis activity involves researching the problem in detail and evaluating various approaches to resolve the same. Design activity is carried out in two fold, high level and low level design. High level design involved architectural design of the framework and case study planning as it plays a vital role in validating the approach.

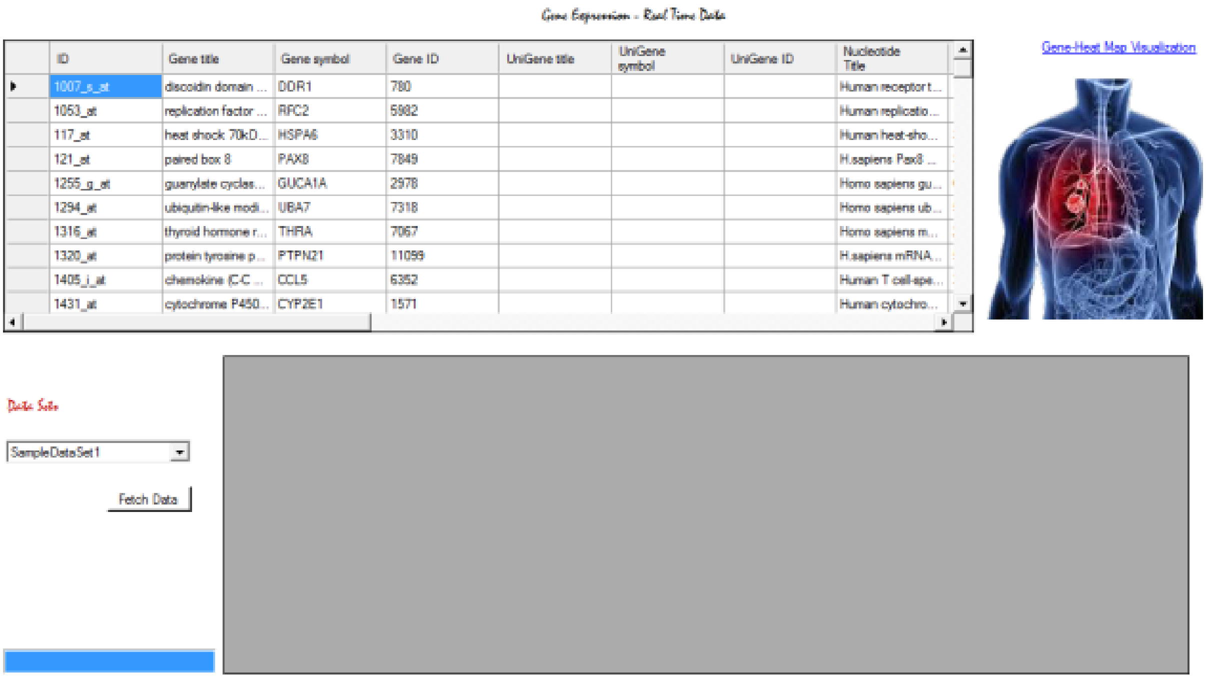
4. ARCHITECTURAL OVERVIEW

Predicting Cancer by analyzing gene and converting the gene expression is the proposed concept of our project, which leads to identifying and analyzing the cancer result set. Controlling Gene Activity From Gene to Functional Protein & Phenotype has also been analyzed in order to identify the cancer cells. In our proposed methodology the experts documental DNA data methylation (Gene expression segments) is a kind of binding site for proteins which make DNA inaccessible to be in alive state.



5. GENE HEAT MAP VISUALIZATION

The goal is to reduce the dimensionality of data to facilitate visualization and additional analysis. They are often used as a preliminary step to clustering of large data sets. MDS starts from a distance matrix between objects and finds the locations of these objects in a low dimensional space that best preserves the original distances. These techniques work on ratio-optimization principle. It's almost concomitant of clustering techniques for high dimensional data to be exploratory. Their strength is in providing rough maps and suggesting directions for further study. Also, clustering results are sensitive to a variety of user-specified inputs. The clustering of a large and complex set of objects can be planned in different ways depending on the goals.



Gene Expression - Real Time Data

ID	Gene title	Gene symbol	Gene ID	UniGene title	UniGene symbol	UniGene ID	Nucleotide Title
1007_s_at	discoidin domain ...	DDR1	780				Human receptor t...
1053_at	replication factor ...	RFC2	5982				Human replicatio...
117_at	heat shock 70kD...	HSPA6	3310				Human heat-sho...
121_at	paired box 8	PAX8	7849				H.sapiens Pax8 ...
1255_g_at	guanylate cyclas...	GUCA1A	2978				Homo sapiens gu...
1294_at	ubiquitin-like modi...	UBA7	7318				Homo sapiens ub...
1316_at	thyroid hormone r...	THRA	7067				Homo sapiens m...
1320_at	protein tyrosine p...	PTPN21	11099				H.sapiens mRNA...
1405_i_at	chemokine (C-C ...	CCL5	6352				Human T cell-spe...
1431_at	cytochrome P450...	CYP2E1	1571				Human cytochro...

Fig. 5.1 GENE HEAT MAP VISUALIZATION

6. CLUSTERING

These techniques can be used in microarray analysis to facilitate visual display (mostly preferred by biologists)and interpretation of experimental results and suggest the presence of subgroups of objects (genes or samples) that behave similarly. Often finds itself as the foremost step of data infiltration since it is vital to parry microarray data for noise elimination. Confusion marks with the trait of the genes to participate in multiple pathways that may or may not be coactive under all conditions, so a gene can find its place in multiple clusters or in none at all. Clustering can be sample-based and/or gene-based by character.A gene-based clustering shall abstract genes as objects and samples as features, while sample-based clustering would perceive vice-versa. A third category of clustering type also exists, subspace clustering. Subspace clustering is not “global” rather it aims to cluster genes based on their indulgence in any disease, being a part of one or more biological pathways.

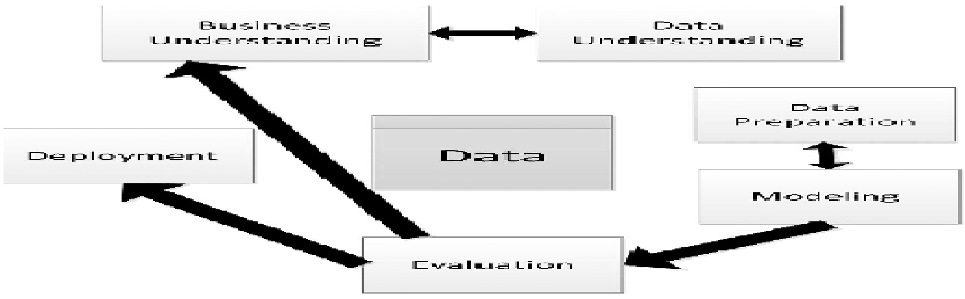


Fig. 6 SUBSPACE CLUSTERING

6.1 K-Means Clustering and Self Organized Maps (SOM):

It partitions objects into groups that have little variability within clusters and large variability across clusters. The user is required to specify the number K of clusters a priori. Estimation is iterative, starting with a random allocation of objects to clusters, re-allocating to minimize distance to the estimated “centroids” of the clusters, and stops when no further improvements can be made. Its implementation is easy and execution is faster. The time complexity was computed to be O (L_K_N), where L is the number of iterations in K clusters.

7. DATASET COLLECTION

In this module we describe dataset collection for multiple real world web services. The term "dataset" originated in the mainframe field. A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a variable. Each row corresponds to a member of the data set. It lists values for each of the variables. The data set may comprise data for one or more members, corresponding to the number of rows.

7.1 Given Input & Output Design:

Gene Heat map visualization.

Input: Fetch data and Dataset comparison.

Output: diagnosis for lung cancer.

8. TECHNIQUE USED

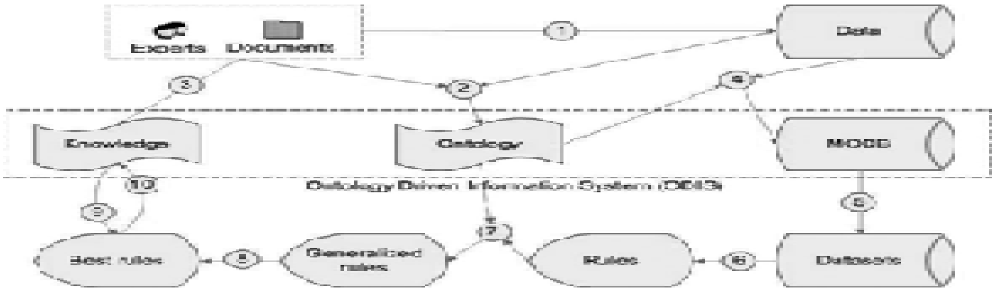
8.1 PRE-PROCESSING PHASE:

Genes, DNA clones, or expressed sequence tags [ESTs] usually constitute the DNA sequences that are scanned by microarray experiments, conditions contingent. They may include time series data of a biological process, e.g., life cycle of a yeast cell, or a collection of varied tissue samples, e.g., normal versus cancerous tissues. Study on promoter sequences can be staple for deriving transcription factors of an associated gene. Regulation of transcription is the most common form of gene control, and the activity of transcription factors allows genes to be specifically regulated during development and in different types of cells.

8.2 POST PROCESSING:

Since, the pre-processing phase aids in precipitating several groups, patterns, correlations of genes at the expression level basis, it becomes almost necessary to re-evaluate and formalize them in a phase called post-processing phase. During this phase, the domain experts analyze and match the extracted patterns to the business objectives and success criteria. The dogma of pattern management is heterogeneous pattern representation. Since the extracted patterns can be relevant as well as irrelevant; indexing them is a labor intensive task that involves marking and classifying them scrupulously.

Predictive Model Markup Language (PMML) and Common Warehouse model for Data Mining (CWM-DM) were designed for genetic data modelling, but they lacked the efficacy to handle and represent specific classes of patterns. As a solution, Rizzi et al. introduced Pattern Base Later, Kotsifakos et al. revised the PBMS architecture by enabling support for domain ontologies. After defining a data modelling system, it’s vital to design a mechanism to query and extract the required data. For the same, certain APIs namely, SQL/MM DM, Java Data Mining (JDM) API were standardized to handle data as well as the metadata entwining genetic correlational patterns.



9. CONCLUSION

A reliable lead and precise classification of tumors is essential for successful diagnosis and treatment of cancer. The microarray experiments may lead to a more complex understanding of the molecular version of the tumors. The ability to distuinguish between tumor classes using gene expression is a new approach to cancer classification.

A microarray dataset contains the numerous groups of co-expressed genes. A typical strategy for the biologist is to start from genes which are known to be closely related to a biological function and to browse the preliminary rough clustering result, to focus on a small subset of those genes. Thus biologist follow explanatory strategies by manual knowledge. So, on experimenting with these 'superficial' data and applying for various data mining techniques to them, the data has to be concise and close to accurate to obtain the results of cancer.

REFERENCES

[1] Data and Statistics. World Health Organization, Geneva, Switzerland,2010.

[2] PubMedHealth- U.S. Nat. Library Med., (2009).[Online].Available:http://www. ncbi.nlm.nih.gov/pubmedhealth/ PMH0002267/

[3] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," J. Amer. Statist. Assoc., vol. 97, no. 457, pp. 77–87, Mar 2009.

[4] G.-M. Elizabeth and P. Giovanni, (2008, Dec.). "Clustering and classification methods for gene expression data analysis." Johns Hopkins Univ., Dept. of Biostatist. Working Papers. Working Paper 70. [Online]. Available: http://biostats.bepress. com / jhubiostat/paper70//

[5] E. Shay, (2007, Jan.). "Microarray cluster analysis and applications"[Online]. Available: http:// www.science.co.il/enuka/Essays//Microarray-Review.pdf.

[6] M. B. Eisen, T. P. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc.Nat. Acad. Sci. USA, vol. 95, no. 25, pp. 14863–14868, Dec. 2007.

[7] S. Tavazoie, D. Hughes, M. J. Campbell, R. J. Cho, and G. M.Church, "Systematic determination of genetic network architecture," Nature Genetics, vol. 22, pp. 281–285, 2006.

[8] T. Kohonen, Self-Organising Maps. Berlin, Germany: Springer-Verlag, 2005.

[9] N. Pasquier, C. Pasquier, L. Brisson, and M. Collard, (2005)."Mining gene expression data using domain knowledge," Int. J.Softw. Informat, vol. 2, no. 2, pp. 215–231, [Online] Available: http://www.ijsi.org/1673-7288/2/215//

[10] N. Revathy and R. Amalraj, "Accurate cancer classification using expressions few genes," Int. J. Comput. Appl., vol. 14, no. 4,pp. 19–22, Jan. 2005.

[11] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, (2005)"RankGene: Identification of diagnostic genes based on expression data," Bioinformatics, vol. 19, no. 12, pp. 1578–1579, [Online] Avaialble: http://bioinformatics.oxfordjournals.org/content/19/

[12] K. Raza and A. Mishra, "A novel anticlustering filtering algorithm for the prediction of genes as a drug target," Amer. J. Biomed. Eng.vol. 2, no. 5, pp. 206–211, 2004.

[13] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," IEEE Trans. Knowl. Data Eng., vol. 16,no. 11, pp. 1370–1386, Nov. 2004.

[14] D. A. Roff and R. Preziosi, "The estimation of the genetic correlation:The use of the jackknife," Heredity, vol. 73, pp. 544–548, 2004.

[15] T. Scharl and F. Leisch, "Jackknife distances for clustering timecourse expression data," in Proc. ASA Biometrics, 2005, p. 8.

[16] K. M Williams, "Statistical Methods for analysing microarray data: Detection of differentially expressed genes" Inst. Signal Process.,Tampere Univ. Technol. Tampere, Finland, Dep. Biology,Univ. York, York, U.K., 2004.

[17] B. Collard, "An ontology driven data mining process" Inst.TELECOM, TELECOM Bretagne, CNRS FRE 3167 LAB-STICC,Technopole Brest-Iroise, France & Univ. Nice Sophia Antipolis,France, 2003.

[18] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data,"Quaestiones Geographicae, vol. 30, no. 2, pp. 87–93, 2003.

[19] B. Collard, "How to semantically enhance a data mining process?"Lecture Notes Bus. Inform. Process., vol. 13, pp. 103–116, 2003.