

Giorgio Valentini - Research activity

The research activity may be set out in two main areas, *bioinformatics* and *machine learning*.

General scheme of the main research lines.

- I. Bioinformatics
 - A) Analysis, development and application of unsupervised machine learning methods in bioinformatics:
 - A1. *Stability-based methods for the assessment of the reliability of clusters discovered in complex bio-molecular data.*
 - A2. *Ensemble clustering methods for the analysis of patterns in bio-molecular data*
 - B) Analysis, development and application of supervised machine learning methods in bioinformatics:
 - B1. *Analysis and design of supervised ensemble methods to support bio-molecular diagnosis.*
 - B2. *Integration of complex bio-molecular data and integration of feature extraction and feature selection methods for the supervised classification of co-expressed genes.*
 - B3. *Ontology-based hierarchical classification of genes and proteins.*
 - B4. *Integration of multiple sources of bio-molecular data for gene function prediction.*
 - B5. *Machine learning methods for lung nodule detection in x-ray images*
 - C) Other bioinformatics research activities:
 - C1. *Biologically motivated modelling of gene expression profiles*
 - C2. *Ontology driven gene selection methods for the discovery of functional classes of genes related to a specific phenotype*
 - C3. *DNA microarray data analysis for the discovery of gene networks related to Human Acute Myeloid Leukaemia stem cells.*
- II. Machine learning
 - A) Analysis and design of ensembles of learning machines:
 - A1. *Error Correcting Output Coding ensemble methods for multiclass classification.*
 - A2. *Ensemble methods based on the bias-variance decomposition of the error.*
 - A3. *Supervised ensemble methods based on random projections*
 - A4. *Hierarchical ensembles for multi-class, multi-label and multi-path classification problems.*
 - A5. *Ensemble clustering methods*
 - B) Development of machine learning software libraries

Description of the main research lines.

Numbers in square brackets refer to publications listed at the end of this document. Cited papers are on-line available from: <http://homes.dsi.unimi.it/~valenti/pub.html>

I. Bioinformatics

Bioinformatics research activities of Giorgio Valentini are characterized by the development and application of machine learning methods and algorithms to extract biological knowledge from bio-molecular data generated through high throughput biotechnologies.

The main research lines in this area are described below.

A. Analysis, development and application of unsupervised machine learning methods in bioinformatics

This research line is articulated across two related directions:

A1. Stability-based methods for the assessment of the reliability of clusters discovered in complex bio-molecular data.

A2. Ensemble clustering methods for the analysis of patterns in bio-molecular data

A1. Stability-based methods for the assessment of the reliability of clusters discovered in complex bio-molecular data.

The validation of clusters discovered by clustering algorithms is a central problem in bioinformatics: in both genomics and proteomics several problems require assessing the reliability of structures and patterns discovered in complex biomolecular data.

The research activity is set out in the development of algorithms for the analysis of the clusters reliability and for the model order selection in an unsupervised setting of the problem [12,16,15,54,56,89], and in the development of algorithms to analyze the reliability of single clusters inside a clustering [17,19,61], using a novel approach based on the analysis of the stability of the obtained clusters.

New statistical tests based on the χ^2 distribution [16,56] and on the classical Bernstein inequality [12,54] have been developed, in order to search for multiple structures present at the same time in complex data.

The new methods have been applied to the analysis and validation of subclasses of pathologies characterized at bio-molecular level and to the discovery of multiple structures in complex bio-molecular data (e.g. hierarchical structures), using data generated through high-throughput biotechnologies [12,16,18,17,90].

We are now studying new stability-based methods for the discovery and statistical validation of clusterings characterized by a high number of clusters and examples, targeted to the unsupervised search and validation of functional classes of genes [8,47].

A2. Ensemble clustering methods for the analysis of patterns in bio-molecular data

The search for patterns in high dimensional data (e.g. DNA microarray or mass spectrometry data) require the development of clustering methods targeted to these type of biomolecular data.

In particular we developed unsupervised ensemble methods based on random projections to analyze data characterized by a high dimensionality [57]; these methods have been successively applied to the analysis of gene expression data [55]. In the context of a PhD thesis at DSI, new ensemble methods based on random projections, using a fuzzy approach for both the base clusterings of the ensemble and to combine the clusterings obtained from multiple instances of the projected data have been developed.

From an initial algorithm [52], a more general algorithmic scheme has been developed, from which different fuzzy ensemble clustering algorithms can be derived [50]. Some of these fuzzy ensemble algorithms have been applied to the analysis of gene expression data to discover subclasses of pathologies at bio-molecular level [10]. A novel on going research line tries to embed stability based algorithms within the unsupervised ensemble methods, in order to integrate the cluster validation process with the ensemble clustering algorithms.

B. Analysis, development and application of supervised machine learning methods in bioinformatics

The research activity focuses on the development of bioinformatics methods based on ensembles of

learning machines.

In particular 4 research lines can be distinguished:

B1. *Analysis and design of supervised ensemble methods to support bio-molecular diagnosis.*

B2. *Integration of complex bio-molecular data and integration of feature extraction and feature selection methods for the supervised classification of co-expressed genes.*

B3. *Ontology-based hierarchical classification of genes and proteins.*

B4. *Integration of heterogeneous biomolecular data for gene function prediction.*

B5. *Machine learning methods for lung nodule detection in x-ray images*

B1. Analysis and design of supervised ensemble methods to support bio-molecular diagnosis.

The biomolecular classification of pathologic phenotypes requires methods targeted to the characteristics of the biomolecular data. In this context we explored several supervised ensemble methods ranging from methods based on correcting codes, to methods based on dimensionality reduction by random projections and to data complexity-based ensemble methods.

Error Correcting Output Coding (ECOC) methods allow to classify multiple phenotypes through the decomposition of a complex multiclass classification problem in a set of dichotomic simpler problems, introducing also auto-correction capabilities by means of ECOC codes. This approach can also reduce the noise underlying data obtained from biotechnologies (e.g. gene expression data) by introducing automatic error correction procedures.

In particular, ECOC ensembles of neural networks permitted to obtain state-of-the-art predictions to support bio-molecular diagnosis of tumoral diseases [27,71,75].

We explored also other possibilities, considering the low cardinality of available data, using bagged ensembles of SVMs to support the diagnosis of malignancies based on DNA microarray data analysis [24,68]. A variant of bagging, based on bias-variance analysis, showed very interesting results [67].

Another research line extend Ho's random subspace ensembles to more general ensembles based on random projections obeying the Johnson-Lindestrauss lemma: in this way data dimensionality can be reduced without introducing relevant metric distortion into the data. This approach allows us to deal with the curse of dimensionality problem that plagues methods supporting bio-molecular diagnosis using high dimensional data such as gene expression or mass spectrometry data [21,62,64].

To improve the accuracy of the base learners of the ensemble, we started a new research line to join data generation of compress data through random projections with selection procedures of the base learners by an analysis of the complexity of the available data: applications to DNA microarray and SAGE data analysis for bio-medical diagnosis show preliminary very encouraging results [43,49].

B2. Integration of complex bio-molecular data and integration of feature extraction and feature selection methods for the supervised classification of co-expressed genes.

The predictability of classes of co-expressed genes of non coding regulatory DNA regions, is an interesting problem that can indirectly provide information on transcription factor binding sites and on regulatory motifs of non coding DNA regions.

From a machine learning standpoint, this classification and feature selection problem is particularly relevant for the complexity of the regulatory regions and for the ambiguity of the regulatory motifs, as well as for the need for integrating sequence data (non coding regulatory DNA regions) with gene expression data (to predict classes of co-expressed genes).

To this end we developed methods that combine combinatorial algorithms, feature selection and machine learning classification algorithms to predict classes of coexpressed genes, using non coding DNA sequence data. We obtained results comparable with state-of-the-art in the prediction

of classes of co-expressed genes in the yeast [11,88,89]. This ongoing research line includes the application of the developed methods to other model organisms (comprising humans) for the discovery of regulatory motifs and transcription factor binding sites at the level of the entire genome.

We developed also new methods for the predictions of co-expressed genes by integrating multiple sources of data through ensemble methods [39].

B3. *Ontology-based hierarchical classification of genes and proteins.*

In the context of the prediction of genes/proteins functions with computational methods, the analysis of graphs of the Gene Ontology (GO) and of the trees of FunCat (Functional Categories), by which the relationships between functional classes of genes are structured, is very relevant.

Moreover the structured classification of functional classes of genes requires automatic procedures to associate genes to functional classes and to the different types of biomolecular data.

To this end we developed methods and algorithms to select functional classes of genes/proteins to specific biological problems, to analyze and process the graphs of GO and FunCat, and to perform pre-processing of complex and multi-view bio-molecular data, in order to support the development of hierarchical classification methods based on the taxonomy of FunCat and on the Gene Ontology [13].

Gene function prediction is a complex multi-class and multi-label classification problem characterized by a hierarchical structure of the classes. We developed hierarchical classification methods for gene/protein function prediction based on tree-structured ensembles of learning machines. In particular we developed methods based on the “true path rule” (TPR) that governs both FunCat and the GO [2,38,42] and cost-sensitive bayesian methods for the probabilistic “reconciliation” of the probabilistic output of the base learners [5,36]. Both methods, despite the fact that comes from different theoretical and heuristic approaches, showed comparable results, at least when a single source of biomolecular data is used to classify genes at genome and ontology wide level with the FunCat taxonomy [35]. The extension of the PTR method to DAG-based taxonomies (i.e. the GO) is under development.

In perspective, by integrating the research lines on hierarchical ensemble methods with those for the integration of multiple sources of data, we foresee applications to the prediction of gene functions in yeast, *C. elegans*, *A. thaliana* e *M. musculus*. Preliminary results (not just published at this time) obtained with *S.cerevisiae* (yeast) are very encouraging.

B4. *Integration of heterogeneous biomolecular data for gene function prediction.*

Single sources of biomolecular data are usually predictive only for some functional classes, while can be uninformative for others, because each source of data can capture only a subset of the functional characteristics of genes and gene products.

For this reason the integration of different sources plays a central role in bioinformatics.

To this end we developed methods based on ensembles of learning machines [6,40,41,45], showing that also relatively simple methods such as weighted majority voting or decision template ensemble may achieve results comparable with those obtained with state-of-the-art methods [4,37]. Moreover we showed that ensemble methods can tolerate relatively high levels of noise in the data without significantly worsening their performances [1]. We studied also problems related to the biomolecular data base management using XML to integrate heterogeneous biological data [7,46].

B5. *Machine learning methods for lung nodule detection in x-ray images*

In this research activity, we applied classical machine learning methods (e.g. SVMs) to the classification of lung nodules in radiographic images. In particular, the application of univariate feature selection methods and cost-sensitive SVMs, to balance between positive and negative

examples, provided results comparable with best ones available in the literature [22,63].

C. Other bioinformatics research activities

Other bioinformatics research activities are not directly attributable to the previous research areas previously described:

C1. *Biologically motivated modelling of gene expression profiles*

C2. *Ontology driven gene selection methods for the discovery of functional classes of genes related to a specific phenotype*

C3. *DNA microarray data analysis for the discovery of gene networks related to Human Acute Myeloid Leukaemia stem cells.*

C1. *Biologically motivated modelling of gene expression profiles.*

The evaluation of gene selection methods is an open problem in bioinformatics. Indeed, in most cases the genes related to a given phenotype are not known in advance, and hence the estimate of the effectiveness of gene selection methods is difficult and only indirectly evaluated through classification results. For these reasons we developed a biologically motivated mathematical model to simulate gene expression data, based on the biological notions of expression profile and expression signature.

The mathematical modelling of these biological concepts have been realized by positive boolean functions and led to a model by which we can generate biologically plausible gene expression data, if we know in advance the expression signatures and the genes associated to a specific phenotype [3,14,58,60]. We are working to statistically validate the proposed model w.r.t. real gene expression data, and to better motivate both the biological background and the mathematical properties of the proposed model.

The main aim of this model consists in generating synthetic gene expression data to compare different gene selection methods, and preliminary results with a limited set of gene selection methods have been provided [44,91].

We plan to provide an extended experimental comparison between different gene selection and gene clustering methods, using “gold standard” data generated by the model.

C2. *Ontology driven gene selection methods for the discovery of functional classes of genes related to a specific phenotype*

The main idea behind this novel research line consists in the embedding of “a priori” biological knowledge in gene selection methods, exploiting the hierarchical structure of the Gene Ontology to select groups of genes related to a given pathology. In this context, the basic selection unit is not a single gene, but an entire class of genes. This approach from one hand reduces the complexity of the gene selection problem, on the other hand provides a direct biological interpretation of the selection results. In this context, classical algorithms from feature selection literature can be adapted to work on groups of functionally correlated genes, instead of on single genes, exploiting “a priori” biological knowledge on genes (e.g. membership to the same pathway or to the same functional class).

C3. *DNA microarray data analysis for the discovery of gene networks related to Human Acute Myeloid Leukaemia stem cells.*

This research activity is in the context of a collaboration between Università degli Studi di Milano

(Dept. of Biology and Genetics, Faculty of Medicine and Dept. of Computer Science, Faculty of Sciences) and the Niguarda Hospital of Milano. The main goal of the project consists in characterizing the gene networks activated in Human Acute Myeloid Leukaemia (AML) stem cells, in order to discover target genes for bio-molecular therapies at the earliest stage of development of AML.

From a more general standpoint we will apply to real gene expression data, obtained from the Affymetrix platform of the Niguarda Hospital, computational methods for the pre-processing and quality control of gene expression data, for the analysis and detection of differentially expressed genes, for the discovery of genes related to tumoral diseases in general and for the analysis of gene networks related to AML in particular, with relevant applications in the bio-medical field.

In this context we developed computational methods to estimate the reliability of clusters discovered by hierarchical clustering algorithms [8,47].

II. Machine learning

Research activity in machine learning, even if related to bioinformatics research, has its own research lines. They can be summarized in research activities related to the analysis and design of ensembles of learning machines, and the design, development and implementation of machine learning software libraries.

A. Analysis and design of ensembles of learning machines

In this research area we can distinguish between four main research lines for the analysis, design and development of different ensemble methods [32,33,34,69]:

- A1. Error Correcting Output Coding ensemble methods for multiclass classification.*
- A2. Ensemble methods based on the bias-variance decomposition of the error.*
- A3. Supervised ensemble methods based on random projections.*
- A4. Hierarchical ensembles for multi-class, multi-label and multi-path classification problems.*
- A5. Ensemble clustering methods.*

A1. Error Correcting Output Coding ensemble methods for multiclass classification.

Error Correcting Output Coding (ECOC) ensemble methods allow us to improve the reliability of predictions for multiclass classification problems, through the redundant coding of class labels realized by the decomposition of a multiclass classification problem in a set of dichotomic problems delivered to an ensemble of classifiers.

In this framework we analyzed the effectiveness of ECOC methods in ensembles and single learning machines [26,83], developing also new ECOC ensemble approaches [82]. We also evaluated the dependence between bit-level errors of ECOC codewords using an information theoretic approach [25], in order to compare different typologies of ECOC codes and different architectures of learning machines based on ECOC [77,78,81].

Besides applications in bioinformatics [27,31,74,75], ECOC ensembles (and boosting ensembles) have been successfully applied to multi-class classification problems with electronic noses [29,72,76].

A2. Ensemble methods based on the bias-variance decomposition of the error.

We used the bias-variance decomposition of the error to analyze the properties and the

characteristics of learning algorithms. On the basis of the Domingos theory that generalizes to the 0/1 loss the classical bias-variance analysis based on quadratic losses, we analyzed the relationships between learning processes and the bias-variance decomposition of the error in Support Vector Machines [23].

The characterization of the learning behaviour of SVMs in terms of the bias-variance decomposition of the error offers also a rationale for the development of new ensemble methods [23,70].

From this perspective we proposed a new ensemble method, named Lobag (Low Bias Bagging), that estimates the bias of the base learner SVMs, then selects SVMs with the lowest bias and then it combines them by means of bootstrap aggregation. This approach reduces jointly both the bias and variance components of the error, and it may be interpreted as a “low-bias” variant of bagging [66]. This ensemble method has been successfully applied to the classification of tumoral diseases on bio-molecular basis [67].

The bias-variance analysis of the error has been successively extended to resampling-based ensembles, showing the relationships and explaining the different learning behaviour of bagging, random aggregation and lobag [20,65].

A3. Supervised ensemble methods based on random projections.

To support the diagnosis of tumoral diseases based on biomolecular data, we developed ensemble methods based on Ho’s random subspaces, using SVMs as base learner [21,64]

An extension of the Ho’s model, that adds a feature selection stage to eliminate the less relevant features, followed to the classical random subspace approach applied to the remaining features, showed results comparable with the best ones in the field of computer-aided diagnosis of tumors [62].

We are working on an extension of the random subspace ensemble method to more general random projection supervised ensemble methods, following an approach similar to those proposed for unsupervised problems [57,50,10].

A4. Hierarchical ensembles for multi-class, multi-label and multi-path classification problems.

The gene function classification problem stimulated the research and the development of multiclass (gene functional classes are hundreds or thousands too), multilabel (a gene may belong to multiple classes), multi-path (classes are structured according to trees or DAGs) classification algorithms.

These algorithms, originally developed to solve a bioinformatics problem, raise interesting problems also from a machine learning standpoint, when multiple interrelated hierarchically structured classification problems need to be considered at the same time [2,5].

A5. Ensemble clustering methods.

Unsupervised ensemble methods based on random projections, originally motivated by clustering problems in high dimensional spaces in bioinformatics, represent an unsupervised extension of Ho's random subspace methods [57].

In [61] we showed that random projections obtained by classical random subspace methods may induce significant distortions in high-dimensional gene expression data, while using random projections that obey the Johnson-Lindenstrauss lemma, we can generate with high probability lower-dimensional projected data, whose metric characteristics are similar to that of the data in the original space [18].

On the basis of this analysis we proposed clustering ensemble methods based on random projections [57] that have been applied to the analysis of DNA microarray data [55].

A fuzzy extension of the methods developed in [57] have been proposed in [50]: from the combination of different “crispization” techniques of the base fuzzy clustering, and different typologies of fuzzy aggregation of the base clusterings, we obtained an algorithmic scheme from which 9 different fuzzy ensemble clustering algorithms can be derived [50,53]. These algorithms have been applied to the analysis of gene expression data [10,52].

B. Development of machine learning software libraries

Research activity in machine learning have been always associated with the design and implementation of corresponding software libraries: in particular the new ensemble methods realized during the research activities, together with other methods published in literature, have been developed and implemented in a C++ library, *NEUROjects*, initially conceived for the software design of neural networks [28,85].

Symmetrically to the growing interest in bioinformatics research, we designed and implemented open source R libraries, available on-line, to analyze and process complex bio-molecular data.

In particular the *clusterv* library permits to analyze the reliability of single clusters in high dimensional biomolecular data [19]; the *mosclust* library allows us to select the “optimal” number of clusters and to discover multiple structures present at the same time in complex biomolecular data [15]; the *hcgene* library permits to analyze the direct acyclic graphs of the Gene Ontology and the trees of FunCat to support the hierarchical classification of genes and gene products [13].

We are also developing software libraries for the hierarchical classification e for the integration of heterogeneous data based on ensemble methods.

Publications

International journals with peer-review

- [1] M. Re, G. Valentini, Noise tolerance of Multiple Classifier Systems in data integration-based gene function prediction, *Journal of Integrative Bioinformatics*, 2010 (in press)
- [2] G. Valentini, True Path Rule hierarchical ensembles for genome-wide gene function prediction, *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 2010 (in press).
- [3] M. Muselli, A. Bertoni, M. Frasca, A. Beghini, F. Ruffino, and G. Valentini, A mathematical model for the validation of gene selection methods, *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 2010 (in press).
- [4] M. Re, G. Valentini, Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction, *Journal of Machine Learning Research*, W&C Proceedings, 2010 (in press).
- [5] N. Cesa-Bianchi, G. Valentini, Hierarchical cost-sensitive algorithms for genome-wide gene function prediction, *Journal of Machine Learning Research*, W&C Proceedings, 2010 (in press).
- [6] M. Re, G. Valentini, Integration of heterogeneous data sources for gene function prediction using Decision Templates and ensembles of learning machines, *Neurocomputing*, 2010 (in press)
- [7] M. Mesiti, E. Jimenez-Ruiz, I. Sanz, R. Berlanga-Llavori, P. Perlasca, G. Valentini and D. Manset, XML-Based Approaches for the Integration of Heterogeneous Bio-Molecular Data, *BMC Bioinformatics* 10:(S12)S7, 2009
- [8] R. Avogadri, M. Brioschi, F. Ferrazzi, M. Re, A. Beghini, and G. Valentini, A stability-based algorithm to validate hierarchical clusters of genes, *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(4), pp. 318-330, 2009
- [9] G. Valentini, R. Tagliaferri, F. Masulli, Computational Intelligence and Machine Learning in Bioinformatics, *Artificial Intelligence in Medicine* 45(2), pp. 91-96, 2009
- [10] R. Avogadri, G. Valentini, Fuzzy ensemble clustering based on random projections for DNA microarray data analysis, *Artificial Intelligence in Medicine* 45(2), pp. 173-183, 2009
- [11] G. Pavesi, G. Valentini, Classification of co-expressed genes from DNA regulatory regions, *Information Fusion* 10(3), pp. 233-241, 2009
- [12] A. Bertoni, G. Valentini, Discovering multi-level structures in bio-molecular data through the Bernstein inequality, *BMC Bioinformatics* vol.9, Suppl.2, 2008.
- [13] G. Valentini, N. Cesa-Bianchi, HCGene: a software tool to support the hierarchical classification of genes, *Bioinformatics*, 24(5), pp. 729-731, 2008
- [14] F. Ruffino, M. Muselli, G. Valentini, Gene expression modelling through positive Boolean functions, *International Journal of Approximate Reasoning*, 47(1), pp. 97-108, 2008.
- [15] G. Valentini, Mosclust: a software library for discovering significant structures in bio-molecular data, *Bioinformatics* 23(3):387-389, 2007.
- [16] A. Bertoni, G. Valentini, Model order selection for biomolecular data clustering, *BMC Bioinformatics*, vol.8, Suppl.2, 2007.
- [17] A. Bertoni, G. Valentini, Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses, *Artificial Intelligence in Medicine* 37(2):85-109 2006.
- [18] G. Valentini, F. Ruffino, Characterization of Lung tumor subtypes through gene expression cluster validity assessment, *RAIRO - Theoretical Informatics and Applications*, 40:163-176, 2006.

- [19] G.Valentini, Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data, *Bioinformatics* 22(3):369-370, 2006.
- [20] G.Valentini, An experimental bias-variance analysis of SVM ensembles based on resampling techniques, *IEEE Transactions on Systems, Man and Cybernetics*, Part B vol.35(6) pp. 1252-1271, 2005.
- [21] Bertoni, R. Folgieri, G. Valentini, Bio-molecular cancer prediction with random subspace ensembles of Support Vector Machines, *Neurocomputing* vol. 63C pp. 535-539, 2005.
- [22] P. Campadelli, E. Casiraghi, G.Valentini, Support Vector Machines for candidate nodules classification, *Neurocomputing*, vol.68 pp. 281-289, 2005.
- [23] G. Valentini, T. G. Dietterich, Bias-variance analysis of Support Vector Machines for the development of SVM-based ensemble methods, *Journal of Machine Learning Research*, 5(Jul) pp. 725--775, MIT Press, 2004.
- [24] G. Valentini, M. Muselli and F. Ruffino, Cancer recognition with bagged ensembles of Support Vector Machines, *Neurocomputing* vol. 56 pp. 461-466, 2004.
- [25] F. Masulli, G. Valentini, An experimental analysis of the dependence among codeword bit errors in ECOC learning machines. *Neurocomputing* vol. 57 pp. 189-214, 2004.
- [26] F. Masulli, G. Valentini, Effectiveness of output coding decomposition schemes in ensemble and monolithic learning machines. *Pattern Analysis and Applications* vol. 6 pp. 285-300, 2003.
- [27] G. Valentini, Gene expression data analysis of human lymphoma using Support Vector Machines and Output Coding ensembles, *Artificial Intelligence in Medicine* 26(3) pp 283-306, 2002.
- [28] G. Valentini, F. Masulli, NEUROObjects: an object-oriented library for neural network development, *Neurocomputing* 48(1-4) pp. 623-646 , 2002.
- [29] M. Pardo, G. Sberveglieri, A.Taroni, F. Masulli, G. Valentini Decompositive classification models for electronic noses, *Anal. Chim. Acta* (446) pp. 223-232, 2001.

National journals with peer-review

- [30] F.Ruffino, G.Valentini, M. Muselli, Valutazione di metodi di gene selection per l'analisi di dati con DNA microarray, *Automazione e Strumentazione*, LIII (10) pp. 106-119, 2005
- [31] G. Valentini, Gene expression-based prediction of malignancies, *AIIA Notizie* XV(4) pp. 34-38, 2002.

Books

- [32] O. Okun, G. Valentini (eds.), Applications of Supervised and Unsupervised Ensemble Methods, *Studies in Computational Intelligence*, vol. 245 © Springer, 2009.
- [33] O. Okun, G. Valentini (eds.), Proceedings of the the Second Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications (SUEMA), European Conference on Artificial Intelligence, University of Patras, Greece, ISBN: 978-960-89282-2-0, 2008.
- [34] O. Okun, G. Valentini (eds.), Supervised and Unsupervised Ensemble Methods and their Applications, *Studies in Computational Intelligence*, vol. 126, Springer, 2008.

Proceeding of international conferences and book chapters with peer-review

- [35] M. Re, G. Valentini, An experimental comparison of Hierarchical Bayes and True Path Rule ensembles for protein function prediction, 9th International Workshop on Multiple Classifier Systems MCS 2010, *Lecture Notes in Computer Science*, Springer (in press)
- [36] N. Cesa-Bianchi, G. Valentini, Hierarchical cost-sensitive algorithms for genome-wide gene function prediction, *Machine Learning in Systems Biology, Proceedings of the Third international workshop*, Ljubljana, Slovenia, pp. 25-34, 2009.
- [37] M. Re, G. Valentini, Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction, *Machine Learning in Systems Biology, Proceedings of the Third international workshop*, Ljubljana, Slovenia, pp. 95-104, 2009.
- [38] G. Valentini, M. Re, Weighted True Path Rule: a multilabel hierarchical algorithm for gene function prediction, *MLD-ECML 2009, 1st International Workshop on learning from Multi-Label Data*, Bled, Slovenia, pp. 133-146, 2009.
- [39] M. Re, G. Valentini, Predicting gene expression from heterogeneous data, *CIBB 2009, The Sixth International Conference on Bioinformatics and Biostatistics*, Genova, Italy, 2009.
- [40] M. Re, G. Valentini, Comparing early and late data fusion methods for gene function prediction, Neural Nets WIRN09 - Proceedings of the 19th Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy, 2009, *Frontiers in Artificial Intelligence and Applications* vol. 204, pp. 197-207, IOS Press, 2009.
- [41] M. Re, G. Valentini, Ensemble based Data Fusion for Gene Function Prediction, In: (J. Kittler, J. Benediktsson, F. Roli, Eds.) Eighth International Workshop on Multiple Classifier Systems MCS 2009, *Lecture Notes in Computer Science*, vol.5519 pp.448-457, Springer 2009.
- [42] G. Valentini, True Path Rule Hierarchical Ensembles, In: (J. Kittler, J. Benediktsson, F. Roli, Eds.) Eighth International Workshop on Multiple Classifier Systems MCS 2009, *Lecture Notes in Computer Science*, vol.5519 pp.232-241, Springer 2009.
- [43] O. Okun, G. Valentini, H. Priisalu, Exploring the link between bolstered classification error and dataset complexity for gene expression based cancer classification, In T. Maeda, ed., *New Signal Processing Research*, Nova Publishers, 2009.
- [44] A. Bertoni, G. Valentini, Unsupervised stability-based ensembles to discover reliable structures in complex bio-molecular data, in: Proc. CIBB 2008, The Fifth International Conference on Bioinformatics and Biostatistics, *Lecture Notes in Computer Science*, vol. 5488 pp. 25-43, Springer, 2009.
- [45] M. Re, G. Valentini, Prediction of gene function using ensembles of SVMs and heterogeneous data sources, in: Applications of supervised and unsupervised ensemble methods, *Computational Intelligence Series*, Springer, 2009.
- [46] M. Mesiti, E. J. Ruiz, I. Sanz, R. Berlanga, G. Valentini, P Perlasca, D. Manset, Data Integration and Opportunities in Biological XML Data Management, in: E. Pardede (editor): Open and Novel Issues in XML Database Applications: Future Directions and Advanced Technologies, Information Science, pp. 263-286, 2009.
- [47] R. Avogadri, M. Brioschi, F. Ruffino, F. Ferrazzi, A. Beghini and G. Valentini, An algorithm to assess the reliability of hierarchical clusters in gene expression data, *KES2008, 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, workshop on

Unsupervised Clustering for Exploratory Data Analysis, *Lecture Notes in Computer Science*, Springer, 2008 (in press)

[48] M. Mesiti, E. J. Ruiz, I. Sanz, R. Berlanga, G. Valentini, P. Perlasca, D. Manset, XML-based approaches for the integration of heterogeneous bio-molecular data, *NETTAB 2008 workshop on: "Bioinformatics Methods for Biomedical Complex System Applications"*, 2008 .

[49] O. Okun, G. Valentini, Dataset Complexity Can Help to Generate Accurate Ensembles of K-Nearest Neighbors, Proc. of the *International Joint Conference on Neural Networks (IJCNN2008)* as part of 2008 *IEEE World Congress on Computational Intelligence (WCCI2008)*, Hong Kong, 2008.

[50] R. Avogadri, G. Valentini, Ensemble Clustering with a Fuzzy Approach, in: "Supervised and Unsupervised Ensemble Methods and their Applications", *Studies in Computational Intelligence*, vol. 126, Springer, 2008.

[51] R. Tagliaferri, A. Bertoni, F. Iorio, G. Miele, F. Napolitano, G. Raiconi and G. Valentini, A Review on clustering and visualization methodologies for Genomic data analysis (extended abstract). *Workshop on Computational Intelligence approaches for the analysis of Bioinformatics data, IJCNN 2007*, Orlando, USA, 2007.

[52] R. Avogadri, G. Valentini, Fuzzy ensemble clustering for DNA microarray data analysis, *CIBB 2007, The Fourth International Conference on Bioinformatics and Biostatistics, Lecture Notes in Computer Science*, vol. 4578, pp.537-543, 2007.

[53] R. Avogadri, G. Valentini, An unsupervised fuzzy ensemble algorithmic scheme for gene expression data analysis, *NETTAB 2007 workshop on a Semantic Web for Bioinformatics*, Pisa, Italy, 2007.

[54] A. Bertoni, G. Valentini, Discovering Significant Structures in Clustered Bio-molecular Data Through the Bernstein Inequality Knowledge-Based Intelligent Information and Engineering Systems, 11th International Conference, KES 2007, *Lecture Notes in Computer Science*, vol. 4694 pp. 886-891, 2007.

[55] A. Bertoni, G. Valentini, Randomized Embedding Cluster Ensembles for gene expression data analysis, *SETIT 2007 - IEEE International Conf. on Sciences of Electronic, Technologies of Information and Telecommunications*, Hammamet, Tunisia, 2007.

[56] A. Bertoni, G. Valentini, Model order selection for clustered bio-molecular data, In: *Probabilistic Modelling and Machine Learning in Structural and Systems Biology*, J. Rousu, S. Kaski and E. Ukkonen (Eds.), Tuusula, Finland, 17-18 June, pp. 85-90, Helsinki University Printing House, 2006.

[57] A. Bertoni, G. Valentini, Ensembles Based on Random Projections to Improve the Accuracy of Clustering Algorithms, Neural Nets, *WIRN 2005, Lecture Notes in Computer Science*, vol. 3931, pp. 31-37, 2006.

[58] F. Ruffino, M. Muselli, G. Valentini, Modelling gene expression data via positive Boolean functions, *NETTAB 2006 workshop on Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics*, S. Margherita di Pula 10-13 July, Italy, 2006.

[59] B. Apolloni, G. Valentini, A. Brega, BICA and Random Subspace ensembles for DNA microarray-based diagnosis, *CIBB 2006 - International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics* In Proc. of 7th International FLINS Conference on Applied Artificial Intelligence pp. 623-631, World Scientific, 2006.

[60] F. Ruffino, M. Muselli, G. Valentini Biological specifications for a synthetic gene expression data generation model, In: I. Bloch, A. Petrosino, A. Tettamanzi (Eds.) *WILF 2005, Lecture Notes in*

Artificial Intelligence vol. 38, pp. 277-283, 2006.

[61] A. Bertoni, G. Valentini, Random projections for assessing gene expression cluster stability, *IJCNN '05. Proceedings IEEE International Joint Conference on Neural Networks*, vol. 1 pp. 149-154, 2005.

[62] A. Bertoni, R. Folgieri, G. Valentini, Feature selection combined with random subspace ensemble for gene expression based diagnosis of malignancies, In: (B. Apolloni, M. Marinaro and R. Tagliaferri, eds) *Biological and Artificial Intelligence Environments*, pp. 29-36, Springer, 2005.

[63] P. Campadelli, E. Casiraghi, G. Valentini, Lung nodules detection and classification, *ICIP 05, The IEEE International Conference on Image Processing*, Genova, Italy, 2005.

[64] A. Bertoni, R. Folgieri, G. Valentini, Random subspace ensembles for the bio-molecular diagnosis of tumors, *Models and Metaphors from Biology to Bioinformatics Tools*, NETTAB 2004.

[65] G. Valentini, Random aggregated and bagged ensembles of SVMs: an empirical bias-variance analysis, In: (F. Roli, J. Kittler, T. Windeatt Eds.) *Fifth International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, vol. 3077, pp. 263-272, 2004

[66] G. Valentini, T.G. Dietterich, Low Bias Bagged Support Vector Machines, *The Twentieth International Conference on Machine Learning, ICML 2003*, Washington D.C. USA, pp. 752-759, AAAI Press, 2003.

[67] G. Valentini, An application of Low Bias Bagged SVMs to the classification of heterogeneous malignant tissues, Pre-WIRN workshop on Bioinformatics and Biostatistic, *Lecture Notes in Computer Science*, vol. 2859, pp. 316-321, 2003.

[68] G. Valentini, M. Muselli and F. Ruffino, Bagged Ensembles of SVMs for Gene Expression Data Analysis, *IJCNN2003, Proc. of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Portland, USA, pp. 1844-1849, IEEE, 2003.

[69] G. Valentini, F. Masulli, Ensembles of learning machines. In R. Tagliaferri and M. Marinaro, editors, *Neural Nets WIRN Vietri-2002, Lecture Notes in Computer Sciences*, vol. 2486, pp. 3-19, 2002.

[70] G. Valentini, T.G. Dietterich, Bias-Variance Analysis and Ensembles of SVM. In J. Kittler and F. Roli (Eds) *Third International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science* vol. 2364, pp. 222-231, 2002.

[71] G. Valentini, Supervised gene expression data analysis using Support Vector Machines and Multi-Layer Perceptrons, *Sixth International Conference on Knowledge-Based Intelligent Information & Engineering Systems KES'2002, special session Machine Learning in Bioinformatics*, 2002

[72] F. Masulli, M. Pardo, G. Sberveglieri, G. Valentini, Boosting and Classification of Electronic Nose Data, *Third International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science* vol. 2364, pp. 262-271, 2002.

[73] F. Ruffino, M. Muselli and G. Valentini, Feature Selection and Bagging Improve Malignancy Prediction based on Gene Expression Data. *Understanding the Genome: Scientific Progress and Microarray Technology*, Genova, Italy, 2002.

[74] G. Valentini, Identifying different types of human lymphomas by SVM and ensembles of learning machines using DNA microarray data, *ISMB 2001, 9th International Conference on Intelligent Systems and Molecular Biology* (Poster section), Copenhagen, Denmark, 2001.

[75] G. Valentini, Classification of human malignancies by machine learning methods using DNA microarray gene expression data, *Proceedings of the Fourth International Conference Neural*

Networks and Expert Systems in Medicine and HealthCare, Milos island, Greece, pp. 399-408, 2001.

[76] F. Masulli, G. Valentini, M. Pardo, G. Sberveglieri Classification of sensor array data by Output Coding decomposition methods. *Proc of the International Workshop MATCHEMS 2001*, pp. 169-172, Brescia, Italy, 2001

[77] F. Masulli, G. Valentini, Quantitative evaluation of dependence among outputs in ECOC classifiers using mutual information based measures, *Proceedings of the International Joint Conference on Neural Networks IJCNN'01*, K. Marko and P. Webos (eds.), vol.2, IEEE, Piscataway, NJ, USA, pp. 784-789, 2001.

[78] F. Masulli and G. Valentini, Dependence among Codeword Bit Errors in ECOC Learning Machines: an Experimental Analysis, In: J.Kittler and F.Roli (eds.) Proceedings of the Second International Workshop Multiple Classifier Systems MCS 2001, Cambridge, UK, *Lecture Notes in Computer Science* vol. 2096, pp. 158-167, 2001

[79] M. Pardo, G. Sberveglieri, G. Valentini, D. Della Casa, F.Masulli, Boosting applied to electronic nose data, LFTNC-SC 2001 - 2001 NATO ARW on Limits and Future Trends of Neural Computing, 2001.

[80] M. Pardo, G. Sberveglieri, D. Della Casa, F.Masulli, G. Valentini, Multiple classifiers for electronic nose data, *8th International Symposium on Olfaction and Electronic Noses*, Washington D.C., USA, 2001

[81] F. Masulli, G. Valentini, Comparing Decomposition Methods for Classification, *KES'2000, Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, Brighton, UK, IEEE, Piscataway, NJ, USA, pp. 788-791, 2000.

[82] F. Masulli, G. Valentini, Parallel Non Linear Dichotomizers, *IJCNN2000, The IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Como, Italy, vol.2, pp. 29-33, 2000.

[83] F. Masulli, G. Valentini, Effectiveness of error correcting output codes in multiclass learning problems, In: J.Kittler and F.Roli (eds.) Proceedings of the First International Workshop Multiple Classifier Systems MCS 2000, Cagliari, Italy, *Lecture Notes in Computer Science* vol.1857, pp.107-116, 2000.

[84] M. Pardo, G. Sberveglieri, G. Valentini, F. Masulli, Decompositive classification models for electronic noses. *7th International Symposium on Chemometrics in Analytical Chemistry (CAC)*, Antwerp, 2000.

[85] G. Valentini, F. Masulli, NEUROObjects, a set of library classes for neural networks development, *Proceedings of the third International ICSC Symposia on Intelligent Industrial Automation (IIA'99) and Soft Computing (SOCO'99)*, ICSC Academic Press, Millet, Canada, 1999, pp. 184-190.

Proceedings of national conferences

[86] M. Re, G.Valentini, Data fusion based gene function prediction using ensemble methods, *BITS 2009, Bioinformatics Italian Society Meeting*, Genova, Italy, 2009.

[87] N. Cesa-Bianchi, G. Valentini, Genome-wide hierarchical classification of gene function, *BITS 2009, Bioinformatics Italian Society Meeting*, Genova, Italy, 2009.

[88] R. Avogadri, A. Bertoni, G. Valentini, An integrated algorithmic procedure for the assessment and discovery of clusters in DNA microarray data, *BITS 2009, Bioinformatics Italian Society*

Meeting, Genova, Italy, 2009

[89] G.Valentini, Statistical methods for the assessment of clusters discovered in bio-molecular data, *Proc. of the 6th SIB National Congress, Statistics in Life and Environment Sciences*, Pisa, Italy, 2007.

[90] A.Bertoni, G.Valentini, A statistical test based on the Bernstein inequality to discover multi-level structures in bio-molecular data, *BITS 2007, Bioinformatics Italian Society Meeting*, Napoli, Italy, 2007.

[91] G.Pavesi, G.Valentini, Classification of co-expressed genes from DNA regulatory regions *BITS 2007, Bioinformatics Italian Society Meeting*, Napoli, Italy, 2007.

[92] G. Pavesi , G. Valentini, G. Mauri, G. Pesole, Motif Based Classification of Coregulated Genes, *BITS 2006, Bioinformatics Italian Society Meeting*, Bologna Italy, 2006.

[93] Bertoni, R. Folgieri, F. Ruffino, G. Valentini, Assessment of clusters reliability for high dimensional genomic data, *BITS 2005, Bioinformatics Italian Society Meeting*, Milano Italy, 2005

[94] F. Ruffino, G. Valentini, M.Muselli, Evaluation of gene selection methods through artificial and real-world data concerning DNA microarray experiments, *BITS 2005, Bioinformatics Italian Society Meeting*, Milano Italy, 2005

[95] M. Muselli, F. Ruffino, and G. Valentini, An Artificial Model for Validating Gene Selection Methods, *BITS 2004, Bioinformatics Italian Society Meeting*, Padova, Italy, 2004

[96] F. Ruffino, G. Valentini, and M. Muselli, Metodi di Bagging e di selezione delle variabili per l' analisi dei dati di DNA microarray, *SIS 2003*.

[97] G. Valentini, Metodi di apprendimento automatico supervisionato per il riconoscimento di linfomi tramite DNA microarray, *Atti III Convegno Federazione Italiana Scienze della Vita - FISV 2001*", Riva del Garda (TN), 2001.

[98] M. Pardo, G. Benussi, G. Sberveglieri, G. Valentini, F. Masulli and M. Riani, Application of parallel non-linear dichotomizers to electronic noses, *INFMeeting 2000*, Genova, 2000.