

Attività di ricerca di Giorgio Valentini

L'attività di ricerca si articola in due aree principali, *bioinformatica* e *apprendimento automatico*.

Schema delle linee di ricerca

- I. Bioinformatica
 - A) Analisi, sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico non supervisionato
 - A1. *Metodi basati sull'analisi della stabilità per la valutazione dell'affidabilità dei cluster individuati in dati bio-molecolari complessi*
 - A2. *Metodi di ensemble clustering per la ricerca di pattern in dati bio-molecolari*
 - B) Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico supervisionato
 - B1. *Analisi e sviluppo di metodi di ensemble supervisionati per il supporto alla diagnosi bio-molecolare.*
 - B2. *Integrazione di metodi di feature extraction e feature selection e di dati bio-molecolari complessi per la classificazione supervisionata di geni co-espressi*
 - B3. *Classificazione funzionale di geni e proteine basata su ontologie*
 - B4. *Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche*
 - B5. *Metodi di apprendimento automatico per l'individuazione di noduli maligni in immagini radiografiche.*
 - C) Altre attività di ricerca in ambito bioinformatico
 - C1. *Modellazione biologicamente motivata dei profili di espressione.*
 - C2. *Metodi per la selezione dei geni "Gene Ontology driven".*
 - C3. *Analisi di dati di DNA microarray per lo studio della leucemia mieloide*
- II. Apprendimento automatico
 - A) Analisi e progettazione di metodi di ensemble
 - A1. *Metodi di ensemble a codici a correzione d'errore per la classificazione multiclasse.*
 - A2. *Metodi di ensemble basati sulla scomposizione dell'errore in bias e varianza*
 - A3. *Metodi di ensemble supervisionati basati su proiezioni randomizzate*
 - A4. *Metodi di ensemble gerarchici multiclasse, multietichetta e multi-path*
 - A5 *Metodi di ensemble clustering*
 - B) Sviluppo di librerie software di machine learning

Descrizione delle linee di ricerca

I. Bioinformatica

Le attività di ricerca in bioinformatica sono caratterizzate dallo sviluppo e dall'applicazione di metodi ed algoritmi di apprendimento automatico per l'estrazione di conoscenza biologica da dati bio-molecolari generati con bio-tecnologie high-throughput [9].

A) Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico non supervisionato.

In tale area si possono distinguere due linee tra loro collegate:

A1. *Metodi basati sull'analisi della stabilità per la valutazione dell'affidabilità dei cluster*

individuati in dati bio-molecolari complessi

A2. Metodi di ensemble clustering per la ricerca di pattern in dati bio-molecolari

A1. Metodi basati sull'analisi della stabilità per la valutazione dell'affidabilità dei cluster individuati in dati bio-molecolari complessi.

La validazione dei cluster individuati dagli algoritmi di clustering è un problema di grande rilevanza in ambito bioinformatico: la genomica e la proteomica presentano diversi problemi in cui è fondamentale valutare l'affidabilità delle strutture e dei pattern individuati in dati biomolecolari complessi.

L'attività di ricerca si è articolata nello sviluppo di algoritmi per l'analisi dell'affidabilità e selezione dell'ordine del modello per problemi non supervisionati [12,16,15,54,56,89], e di algoritmi per l'analisi dell'affidabilità dei singoli cluster [17,19,61], utilizzando un nuovo approccio basato sull'analisi della stabilità dei cluster ottenuti. Si sono inoltre sviluppati test statistici basati sulla distribuzione χ^2 [16,56] e sulla classica disuguaglianza di Bernstein [12,54] per la ricerca di strutture multiple in dati complessi.

I metodi sviluppati sono stati applicati alla validazione di sottoclassi patologiche caratterizzate a livello bio-molecolare ed alla ricerca di strutture multiple in dati biomolecolari complessi, utilizzando dati generati tramite bio-tecnologie high-throughput [12,16,18,17,90]. Attualmente sono in corso di studio nuovi metodi basati sull'analisi della stabilità per la ricerca non supervisionata e validazione statistica di clustering caratterizzati da un elevato numero di campioni e cluster, finalizzati alla ricerca e validazione non supervisionata di classi funzionali di geni [8,47].

A2. Metodi di ensemble clustering per la ricerca di pattern in dati bio-molecolari.

La ricerca di pattern bio-molecolari in dati caratterizzati da elevata dimensionalità e bassa cardinalità (ad es: DNA microarray o dati spettrometrici relativi a proteine), ha portato alla progettazione e sviluppo di metodi di ensemble clustering specifici per tale tipologia di dati. In particolare si sono sviluppati metodi non supervisionati basati su proiezioni randomizzate per analizzare dati caratterizzati da elevata dimensionalità [57]. Tali metodi sono stati successivamente applicati all'analisi di dati di espressione genica [55]. Nell'ambito di una tesi di dottorato in corso di svolgimento al DSI, si sono inoltre sviluppati metodi di ensemble clustering basati su proiezioni randomizzate che utilizzano un approccio fuzzy sia per i base clustering costituenti l'ensemble, sia per combinare i clustering ottenuti sulle istanze multiple dei dati. Dall'algoritmo iniziale [52] si è sviluppato uno schema algoritmico più generale da cui sono derivabili diversi algoritmi di fuzzy ensemble clustering [50] e tale approccio è stato applicato all'analisi di dati di espressione genica per la ricerca di sottoclassi patologiche caratterizzate a livello bio-molecolare[10]. Una linea di ricerca in corso prevede l'integrazione di metodi per l'analisi della stabilità in algoritmi di ensemble clustering basati su proiezioni randomizzate.

B) Analisi, sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico supervisionato.

L'attività di ricerca si è concentrata soprattutto sullo sviluppo di metodi bioinformatici basati su ensemble di learning machine.

In particolare si possono distinguere cinque linee di ricerca principali:

B1. Analisi e sviluppo di metodi di ensemble supervisionati per il supporto alla diagnosi bio-molecolare.

B2. Integrazione di metodi di feature extraction e feature selection e di dati bio-molecolari complessi per la classificazione supervisionata di geni co-espressi

B3. Classificazione funzionale di geni e proteine basata su ontologie

B4. Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche

B5. Metodi di apprendimento automatico per l'individuazione di noduli maligni in immagini

radiografiche.

B1. Analisi e sviluppo di metodi di ensemble supervisionati per il supporto alla diagnosi bio-molecolare.

La classificazione di fenotipi patologici su base bio-molecolare richiede lo sviluppo di metodi specifici per le caratteristiche dei dati utilizzati. In tale contesto si sono esplorati diversi metodi di ensemble supervisionati, come i metodi basati su codici a correzione d'errore, sulla riduzione della dimensionalità dei dati attraverso proiezioni randomizzate e sull'analisi della complessità dei dati.

Metodi basati su codici a correzione di errore (ECOC: Error Correcting Output Coding) permettono di classificare fenotipi multipli per la diagnosi bio-molecolare, tramite la scomposizione di problemi di classificazione complessi in sottoproblemi dicotomici di minore complessità, contrastando il rumore che caratterizza alcune tipologie di dati bio-molecolari (ad es: dati di espressione genica) attraverso procedure automatiche di correzione d'errore. In particolare ensemble ECOC di reti neurali hanno consentito di ottenere predizioni allo stato dell'arte per il supporto alla diagnosi bio-molecolare di patologie tumorali [27,71,75].

Si sono esplorate anche altre possibilità, considerando la ridotta cardinalità dei dati a disposizione, attraverso l'applicazione di metodi basati su ricampionamento (bagging) per la diagnosi di tumori basata sull'analisi di dati di DNA microarray [24,68]. Una variante del bagging, basata sull'analisi bias-varianza, ha mostrato ottimi risultati [67].

Un'altra linea di ricerca estende i metodi di ensemble dei random subspace della Ho a metodi di ensemble più generali basati su proiezioni randomizzate che soddisfano il lemma di Johnson e Lindenstrauss: in tal modo la dimensionalità dei dati può essere ridotta senza introdurre distorsioni metriche rilevanti nei dati stessi. Tale approccio permette di affrontare il problema del curse of dimensionality che affligge i metodi per il supporto alla diagnosi bio-molecolare, basati su dati di espressione genica o spettrometria di massa, caratterizzati da ridotta cardinalità ed elevata dimensionalità [21,62,64].

Per migliorare l'accuratezza dei base learner dell'ensemble, una recente linea di ricerca associa alla generazione dei dati compressi tramite proiezioni randomizzate una loro selezione tramite un'analisi della complessità dei dati stessi: applicazioni alla diagnosi basata su dati di DNA microarray e SAGE ha mostrato risultati molto incoraggianti [43,49].

B2. Integrazione di metodi di feature extraction e feature selection e di dati bio-molecolari complessi per la classificazione supervisionata di geni co-espressi.

La predicibilità di classi di geni co-espressi dalle regioni regolatorie non codificanti del DNA è un problema aperto che può fornire indirettamente informazioni sui siti di legame dei fattori di trascrizione e sui motivi regolatori delle regioni non codificanti del DNA. Dal punto di vista dell'apprendimento automatico è un problema di classificazione e feature selection particolarmente rilevante per la complessità delle regioni regolatorie e per l'ambiguità dei motivi regolatori stessi, e per la necessità di integrare dati di sequenza (le regioni regolatorie non codificanti del DNA) e dati di espressione genica (per l'individuazione di classi di geni co-espressi).

A tal fine si sono sviluppati metodi che combinano algoritmi combinatori, algoritmi di feature selection ed algoritmi di classificazione per la classificazione di geni co-espressi basati sull'analisi delle regioni non codificanti del DNA, ottenendo, per il lievito, risultati in linea con lo stato dell'arte [11,88,89]. Tale linea di ricerca è ancora in fase di sviluppo e prevede inoltre l'applicazione dei metodi sviluppati ad altri organismi modello ed anche all'uomo per la ricerca dei motivi regolatori, dei siti di legame e dei fattori di trascrizione a livello dell'intero genoma.

Nell'ambito di questa linea di ricerca si sono sviluppati anche nuovi metodi per la predizione di classi di geni co-espressi utilizzando metodi per l'integrazione di dati biomolecolari eterogenei [39].

B3. *Classificazione funzionale di geni e proteine basata su ontologie*

Nel contesto delle problematiche legate alla predizione delle funzioni di geni/proteine con metodi computazionali, l'analisi dei grafi della Gene Ontology (GO) e degli alberi di FunCat del MIPS di Monaco, tramite cui sono strutturate le relazioni fra le classi funzionali di geni, sono di grande rilevanza in ambito bioinformatico. Inoltre per la classificazione strutturata delle classi funzionali dei geni è essenziale introdurre procedure automatiche per l'associazione dei geni alle classi funzionali e a diverse tipologie di dati bio-molecolari. A tal fine si sono sviluppati metodi ed algoritmi tramite cui è possibile selezionare classi funzionali di geni/proteine correlati a specifici problemi biologici, effettuare il pre-processing di dati bio-molecolari complessi e multi-view, e supportare lo sviluppo di metodi di classificazione gerarchica di geni basati sulle tassonomie della Gene Ontology e di FunCat [13].

La predizione della funzione dei geni è un problema di classificazione multiclasse e multietichetta complesso caratterizzato da una strutturazione gerarchica delle classi. Nel corso degli ultimi 2-3 anni si sono sviluppati metodi di classificazione gerarchici per la predizione delle classi funzionali di geni/proteine, basati su ensemble di learning machine strutturate ad albero. In particolare per FunCat si sono sviluppati metodi di ensemble basati sulla "true path rule" (TPR) che governa sia FunCat che la GO [2,38,42] e metodi bayesiani cost-sensitive per la riconciliazione probabilistica dell'output dei base learner [5,36]. Entrambi i metodi, benchè derivino da impostazioni teoriche ed euristiche differenti, hanno mostrato risultati comparabili, almeno quando un'unica sorgente di dati biomolecolari viene utilizzata per la classificazione dei geni a livello dell'intero genoma e dell'intera tassonomia FunCat [35]. L'estensione del metodo basato sulla "true path rule" alle tassonomie basate su DAG (i.e. la GO) è in corso di studio e sperimentazione.

In prospettiva, integrando le linee di ricerca sui metodi di ensemble gerarchici con quella per l'integrazione di sorgenti multiple di dati, si prevedono applicazioni alla predizione delle funzioni delle proteine di *S.cerevisiae*, *C. elegans*, *A. thaliana* e *M. musculus*. Risultati preliminari (non ancora pubblicati) ottenuti con *S.cerevisiae* (lievito) appaiono molto incoraggianti.

B4. *Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche*

Singole sorgenti di dati biomolecolari sono in genere predittive solo per alcune classi funzionali, mentre possono risultare totalmente non informative per altre, poiché ogni sorgente di dati cattura solo alcune delle caratteristiche funzionali dei geni e dei prodotti genici. Per questa ragione l'integrazione di diverse sorgenti di dati è un problema centrale in bioinformatica. A questo fine si sono sviluppati approcci basati su ensemble di learning machine [6,40,41,45], mostrando che anche metodi relativamente semplici come la votazione maggioranza o i decision template possono ottenere risultati comparabili con lo stato dell'arte [4,37]. Si è inoltre mostrato che i metodi di ensemble sono in grado di tollerare anche relativamente elevati livelli di rumore nei dati, senza una significativo deterioramento delle prestazioni [1]. Si sono inoltre condotti studi sull'utilizzo di XML per l'integrazione di dati biomolecolari eterogenei [7,46].

B5. *Metodi di apprendimento automatico per l'individuazione di noduli maligni in immagini radiografiche.*

In tale attività di ricerca, si sono applicati metodi classici di apprendimento automatico (ad es. SVM) alla classificazione di noduli polmonari rilevati in immagini radiografiche. In particolare, l'applicazione congiunta di metodi univariati di feature selection e di SVM cost-sensitive per bilanciare lo squilibrio fra esempi positivi e negativi ha portato a risultati comparabili con i migliori presenti in letteratura [22,63].

C. Altre attività di ricerca in ambito bioinformatico.

Altre attività di ricerca in bioinformatica non sono direttamente ascrivibili alle due aree in precedenza descritte.

C1. Modellazione biologicamente motivata dei profili di espressione.

La valutazione dei metodi per la selezione dei geni è un problema aperto in ambito bioinformatico. Infatti, nella maggioranza dei casi non sono noti a priori i geni correlati ad un fenotipo specifico, per cui la valutazione dell'efficacia dei metodi di gene selection risulta difficile e mediata attraverso l'utilizzo di metodi di classificazione. Per queste ragioni si è sviluppato un modello matematico per la simulazione di dati di espressione genica, fondato sulle nozioni biologiche di profilo di espressione e di expression signature.

La modellazione matematica di tali concetti biologici è stata realizzata attraverso funzioni booleane positive ed ha condotto ad un modello tramite cui è possibile generare dati di espressione genica biologicamente plausibili, di cui sono note a priori le expression signature ed i geni correlati ad un fenotipo specifico [3,14,58,60]. I dati sintetici generati tramite il modello matematico sono stati utilizzati per comparare le prestazioni di alcuni metodi di gene selection [44,91]. Sono in fase di attuazione una comparazione sperimentale estesa di diverse tipologie di metodi di gene selection e di gene clustering che utilizzano i dati generati tramite il modello matematico come "gold standard".

C2. Metodi per la selezione dei geni "Gene Ontology driven".

Questa nuova linea di ricerca si situa nell'ambito di una collaborazione con il dipartimento di Bioinformatica del Centro de Investigacion Principe Felipe di Valencia. L'idea principale consiste nell'incorporare conoscenza biologica a priori nei metodi per la selezione dei geni, sfruttando la struttura gerarchica della Gene Ontology per individuare gruppi di geni correlati a patologie specifiche. In tale contesto, l'unità base da selezionare non è più il singolo gene ma un'intera classe funzionale di geni. Tale approccio da un lato riduce la complessità del problema di gene selection, dall'altro fornisce un'interpretazione biologica diretta dei risultati della selezione. In tale contesto algoritmi classici della letteratura della feature selection possono essere riadattati in modo da operare non sui singoli geni, ma su gruppi di geni funzionalmente correlati, sfruttando conoscenza biologica a priori sui geni stessi (ad es: appartenenza ad un medesimo pathway o ad una medesima classe funzionale).

C3. Analisi di dati di DNA microarray per lo studio della leucemia mieloide

Tale attività si colloca nell'ambito di un accordo-quadro tra l'Università degli Studi di Milano (Dipartimenti di Genetica di Medicina e DSI di Scienze MFN) e l'Ospedale di Niguarda, per l'elaborazione di dati di espressione genica relativi a pazienti affetti da patologie tumorali ed in particolare per lo studio della leucemia mieloide. Le attività di studio e ricerca permetteranno di applicare a dati reali, ottenuti tramite la piattaforma bio-tecnologica Affymetrix di Niguarda, metodi computazionali per il pre-processing e il controllo di qualità dei dati di espressione genica, per l'analisi e rilevazione di geni differenzialmente espressi, per la ricerca di geni correlati a patologie tumorali e per l'analisi di reti funzionali di geni, con applicazioni rilevanti in ambito bio-medico. In tale contesto si sono sviluppati metodi computazionali per la valutazione dell'affidabilità di cluster di geni individuati tramite algoritmi gerarchici [8,47].

II. Apprendimento Automatico

L'attività di ricerca in apprendimento automatico da un lato è collegata alla bioinformatica e ad alle applicazioni nell'ambito della biologia molecolare, dall'altro ha una sua propria autonomia disciplinare nell'ambito dell'analisi, ricerca e sviluppo di nuovi metodi di apprendimento supervisionato e non supervisionato. In particolare, l'analisi e progettazione di metodi di ensemble ed il design e l'implementazione di librerie software di apprendimento automatico, rappresentano una linea di ricerca attiva sin dai primi anni dell'attività di dottorato e che prosegue tuttora in

parallelo con la ricerca in bioinformatica.

A) Analisi e progettazione di metodi di ensemble

In tale sottoarea si possono distinguere cinque linee principali di ricerca per l'analisi, progettazione e sviluppo di diverse tipologie di metodi di ensemble di learning machine [32,33,34,69]:

A1. Metodi di ensemble a codici a correzione d'errore per la classificazione multiclasse.

A2. Metodi di ensemble basati sulla scomposizione dell'errore in bias e varianza

A3. Metodi di ensemble supervisionati basati su proiezioni randomizzate

A4. Metodi di ensemble gerarchici multiclasse, multi-etichetta e multi-path

A5. Metodi di ensemble clustering

A1. Metodi di ensemble a codici a correzione d'errore per la classificazione multiclasse.

I metodi di ensemble ECOC (Error Correcting Output Coding), consentono di migliorare l'affidabilità della predizione per problemi di classificazione multiclasse attraverso la codifica ridondante delle etichette delle classi realizzata attraverso la scomposizione di un problema multiclasse in una serie di problemi dicotomici risolti da un ensemble di classificatori.

In tale contesto si è analizzata l'efficacia dei metodi ECOC per problemi multi-classe in ensemble di learning machine e learning machine singole [26,83] e si è analizzata sperimentalmente la dipendenza fra gli errori a livello dei singoli bit dei codeword ECOC tramite misure basate sulla mutua informazione [25], al fine di comparare differenti tipologie di codifiche ECOC e diverse architetture di sistemi ad apprendimento automatico basati su ECOC [77,78,81].

Oltre alle applicazioni in ambito bioinformatico [27,31,74,75], gli ensemble ECOC (insieme con metodi di boosting) sono stati applicati con successo anche per problemi multi-classe con nasi elettronici [29,72,76]

A2. Metodi di ensemble basati sulla scomposizione dell'errore in bias e varianza.

In questa attività di ricerca la scomposizione dell'errore in bias e varianza è utilizzata come strumento per analizzare la proprietà e le caratteristiche degli algoritmi di apprendimento.

Sulla base della teoria di Domingos, che generalizza alla funzione di perdita 0/1 l'analisi classica basata sulla funzione di perdita quadratica, si sono analizzate le relazioni fra apprendimento e scomposizione dell'errore in bias e varianza nel caso delle Support Vector Machine [23].

La caratterizzazione dell'apprendimento delle SVM in termini della scomposizione dell'errore in bias e varianza offre inoltre una base razionale per lo sviluppo di nuovi metodi di ensemble [23,70]. Sfruttando l'analisi bias-varianza delle SVM, si è proposto un nuovo algoritmo di ensemble, denominato Lobag (Low Bias Bagging), che stima il bias delle SVM, seleziona le SVM con minor bias e quindi le combina attraverso meccanismi di aggregazione basati su bootstrap. Tale approccio riduce congiuntamente il bias e la varianza dell'errore, e può essere interpretato come una variante "low-bias" del bagging [66]. Il metodo è stato applicato con successo alla classificazione di malattie tumorali su base bio-molecolare [67].

L'analisi bias-varianza dell'errore è stata successivamente estesa a metodi di ensemble basati su ricampionamento, mostrando le relazioni e le differenze nei meccanismi di apprendimento dei metodi di bagging, aggregazione random e lobag [20,65].

A3. Metodi di ensemble supervisionati basati su proiezioni randomizzate

In relazione alla diagnosi di patologie tumorali basata su dati bio-molecolari, sono stati sviluppati metodi di ensemble basati sui random subspace della Ho, utilizzando SVM come base learner [21,64]. Un'estensione del modello della Ho, che prevede uno stage di feature selection per eliminare le feature meno rilevanti per la classificazione, seguito dall'applicazione del metodo dei random subspace sulle feature rimanenti, ha mostrato risultati competitivi con i metodi di ensemble allo

stato dell'arte pubblicati in letteratura [62]. E' in corso di studio un'estensione del metodo della Ho basato su proiezioni randomizzate, analogamente a quanto già realizzato in ambito non supervisionato [57,50,10].

A4. *Metodi di ensemble gerarchici multiclasse, multietichetta e multi-path*

Il problema della classificazione funzionale dei geni ha stimolato la ricerca e lo sviluppo di algoritmi di classificazione multiclasse (le classi funzionali dei geni sono dell'ordine delle centinaia o migliaia), multietichetta (un gene può appartenere a più classi) e multi-path (le classi sono strutturate secondo alberi o DAG). Gli algoritmi, benchè sviluppati per rispondere a un problema bioinformatico, hanno una valenza più ampia, e pongono problematiche interessanti anche dal punto di vista dell'apprendimento automatico [2,5].

A5. *Metodi di ensemble clustering*

I metodi di ensemble non supervisionati basati su proiezioni randomizzate, motivati da problemi di clustering in spazi di elevata dimensionalità e ridotta cardinalità che caratterizzano diversi problemi in ambito bioinformatico, sono un'estensione non supervisionata dei metodi dei random subspace della Ho [57].

In [61] si è mostrato come le proiezioni casuali indotte dal metodo dei random subspace possano produrre distorsioni significative nei dati di espressione genica, mentre utilizzando proiezioni randomizzate che obbediscono al lemma di Johnson e Lindenstrauss è possibile generare con elevata probabilità dati di ridotta dimensionalità le cui caratteristiche metriche sono simili a quelli dello spazio originale [18].

In conformità a questa analisi sono stati proposti metodi di ensemble clustering basati su proiezioni randomizzate [57] che sono stati applicati con successo all'analisi di dati di DNA microarray [55].

Un'estensione fuzzy del metodo di ensemble sviluppato in [57] è stato sviluppato in [50]: dalla combinazione di diversi meccanismi di "crispizzazione" dei fuzzy clustering di base e di diverse tipologie di aggregazione fuzzy dei clustering di base è stato proposto uno schema algoritmico da cui derivano 9 diversi metodi di fuzzy ensemble clustering [50,53]. Tali metodi sono stati applicati all'analisi di dati di espressione genica [10,52].

B) *Sviluppo di librerie software di machine learning*

L'attività di ricerca nell'ambito dei metodi di apprendimento automatico è stata sempre accompagnata da attività di progettazione ed implementazione di librerie software: in particolare i nuovi metodi di ensemble realizzati durante le attività di ricerca, insieme con altri metodi classici pubblicati in letteratura sono stati implementati e resi disponibili in un libreria C++, *NEUROjects*, inizialmente concepita per la progettazione software di reti neurali [28,85].

In seguito, parallelamente all'incremento delle attività di ricerca in ambito bioinformatico, si sono progettate ed implementate librerie R open source, disponibili on line, per l'analisi di dati bio-molecolari complessi. In particolare la libreria *clusterv* [19] permette di analizzare l'affidabilità di singoli cluster in dati bio-molecolari di elevata dimensionalità, la libreria *mosclust* [15] permette di determinare il numero "ottimale" di cluster e di individuare strutture multiple presenti in dati bio-molecolari complessi, mentre la libreria *hcgene* [13] consente di analizzare i grafi diretti aciclici della Gene Ontology e gli alberi di FunCat per supportare la classificazione funzionale delle proteine. Sono in corso di sviluppo librerie software per la classificazione gerarchica e per l'integrazione di dati eterogenei basate su metodi di ensemble.

Publicazioni

Le pubblicazioni elencate sono scaricabili on line da: <http://homes.dsi.unimi.it/~valenti/pub.html>

Riviste internazionali con 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.

Riviste nazionali con 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.

Libri

- [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.

Atti di conferenze internazionali e capitoli di libri con 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, W e ighted 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.

Atti di conferenze nazionali

- [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.