Giorgio Valentini
CURRICULUM VITAE

CURRICULUM SINTETICO

Laurea in Biologia ed in Informatica, PhD in Informatica (Università di Genova). Professore Associato Confermato dal 2014, presso il Dipartimento di Informatica, Università degli Studi di Milano. Abilitazione Nazionale professore di I fascia conseguita nel 2017. Membro del Collegio di Dottorato di Informatica di UNIMI e responsabile scientifico per UNIMI del dottorato europeo in Genomics and Bioinformatics in collaborazione con Il Joint Research center dell' Unione Europea. Attività di ricerca: sviluppo ed applicazione di metodi di Intelligenza Artificiale a problemi bio-medici, con collaborazioni in atto con diversi centri di ricerca ed Università europee ed americane. E' autore di oltre 130 publicazioni scientifiche con peer-review in riviste, capitoli di libri e conferenze internazionali nell' ambito del Machine Learning, della Bioinformatica e delle Biologia Computazionale.

CURRICULUM ANALITICO

Titoli di studio e abilitazioni

- Abilitazione scientifica nazionale per le funzioni di professore di I fascia nel settore concorsuale 01/B1 Informatica -- valida dal 10/04/2017 al 10/04/2023 (art. 16, comma 1, Legge 240/10)
- Dottorato in Informatica, Università di Genova (2003). Titolo: "Ensemble methods based on bias-variance analysis". Relatore: Francesco Masulli.
- Laurea in Scienze dell'Informazione, Università degli Studi di Genova (1999) , voto: 110/110 e lode.
- Laurea in Scienze Biologiche, Università degli Studi di Genova (1981) , voto: 110/110 e lode.
- Maturità classica, Liceo Classico di Savona (1977), voto: 56/60.

Posizione attuale

Professore Associato (settore scientifico-disciplinare INF/01) presso il Dipartimento di Informatica, Università degli Studi di Milano dal 2010.
Esercito i compiti didattici nell'ambito del consiglio di coordinamento didattico di Informatica dello stesso Ateneo. Ho ricevuto la conferma in ruolo nel 2014 da parte della Facoltà e del Dipartimento di appartenenza.

Posizioni precedenti

- Ricercatore a tempo indeterminato presso il Dipartimento di Informatica, Università degli Studi di Milano (2005-2009)
- Ricercatore post-doc (assegnista di ricerca) presso il Dipartimento di Informatica, Università degli Studi di Milano (2003-2004)
- Dottorando in Informatica presso Universita' degli Studi di Genova (2000-2003)
- Ricercatore a contratto presso l'Istituto Nazionale di Fisica della Materia (1999), Genova
- Docente di ruolo in scienze naturali, biologia e chimica e presso diverse Scuole Superiori italiane (1983-1998)

Attività di ricerca

La mia attività di ricerca si situa fra il Machine Learning e la Bioinformatica, ed è motivata da problemi complessi nell'ambito della Biologia Molecolare e della Medicina. Modellando tali problemi come problemi di Machine Learning, sviluppo nuovi algoritmi di apprendimento automatico o adatto algoritmi esistenti per affrontare problemi rilevanti nell'ambito della Bioinformatica, con un interesse particolare, in ispecie negli ultimi anni, allo sviluppo di metodi di Machine Learning per la Medicina Genomica, e la Medicina di Precisione e Personalizzata.
Benchè nella mia attività di ricerca lo sviluppo di nuovi metodi di Machine Learning sia strettamente legato a problemi reali in ambito bio-medico, ho sviluppato anche linee di ricerca di Machine Learning "puro", soprattutto per la progettazione ed analisi di metodi di ensemble di learning machine.
Per tale ragione lo schema generale delle mie principali linee di ricerca si può schematicamente articolare nel modo seguente:

I. *Bioinformatica*
- A. *Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di Machine Learning supervisionato*.
  - A1. Metodi di Machine Learning per la Medicina Personalizzata
  - A2. Metodi di ensemble gerarchici per la predizione strutturata in ontologie biologiche
  - A3. Metodi di ensemble supervisionati per il supporto alla diagnosi bio-molecolare
- B. *Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di Machine Learning semi-supervisionato*.
  - B1. Metodi basati su funzioni di score kernelizzate per l'analisi di reti biomolecolari complesse
  - B2. Metodi basati su Reti di Hopfield parameterizzate e cost-sensitive per la predizione della funzione dei geni.
  - B3. Metodi scalabili per l'analisi di big biomolecular networks.
- C. *Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di Machine Learning non-supervisionato*.
  - C1. Metodi basati sull'analisi della stabilità per la valutazione dell'affidabilità dei cluster individuati in dati bio-molecolari complessi
  - C2. Metodi di ensemble clustering per la ricerca di pattern in dati bio-molecolari
- D. *Metodi per l'integrazione di big data in ambito biologico e medico*.
  - D1. Algoritmi per la combinazione massiva di reti biomolecolari
  - D2. Metodi di ensemble supervisionati per l'integrazione di dati omici

II. *Machine Learning*
- A. *Analisi e progettazione di metodi di ensemble*
  - A1. Metodi di hyper-ensemble per problemi di classificazione sbilanciati con big data
  - A2. Metodi di ensemble gerarchici multiclasse, multietichetta e multi-path
  - A3. Metodi di ensemble basati sulla scomposizione dell'errore in bias e varianza
  - A4. Metodi di ensemble supervisionati basati su proiezioni randomizzate
  - A5. Metodi di ensemble a codici a correzione d'errore per la classificazione multiclasse.
  - A6. Metodi di ensemble clustering
- B. *Progettazione ed implementazione di librerie software di Machine Learning*

Descrizione delle linee di ricerca
Di seguito sono sintenticamente descritte le linee di ricerca con riferimento alle pubblicazioni elencate in fondo al curriculum.

I. Bioinformatica

A. *Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di Machine Learning supervisionato*.

A1. Metodi di Machine Learning per la Medicina Personalizzata.
L'identificazione delle varianti genetiche associate a patologie umane rappresenta una delle sfide fondamentali della "Medicina Personalizzata e di Precisione", e richiede lo sviluppo di una nuova generazione di metodi di Machine Learning per selezionare le rare varianti potenzialmente "deleterie" (cioè causative o associate al rischio di malattia) nel mare di varianti "neutrali" che rappresentano la variabilità genetica "fisiologica" di ogni individuo. A tal fine ho sviluppato *hyperSMURF* (hyper-ensemble of SMOTE under-sampled random forests), un nuovo metodo che adotta strategie di apprendimento basate

su tecniche di resampling e tecniche di hyper-ensembling per affrontare il problema dell'elaborazione di big data genomici e del loro "sbilanciamento" che deriva direttamente dalla caratteristiche dei dati genomici stessi [R49, C77]. Tale metodo di Machine Learning, adattato al problema specifico, costituisce il "core" di Genomiser, una metodologia ed un tool software recentemente proposto nell'ambito di una collaborazione internazionale, che utilizza sia dati genotipici sia fenotipici per individuare varianti patologiche causative di malattie genetiche Mendeliane [R48].
HyperSMURF è un metodo generale per l'analisi di varianti genetiche e può essere applicato allo studio di diverse patologie, ma le sue performance dipendono significativamente dal tuning dei parametri di learning [C78]. Per questa ragione ho sviluppato nell'ambito del progetto *HyperGeV - Detection of Deleterious Genetic Variation through Hyper-ensemble Methods* una versione fortemente parallela dell'algoritmo per architetture High Performance Computing al fine sia di effettuare un tuning fine dei parametri, sia di analizzare big data in ambito benomico [paper sottoposto a GigaScience].
Sto inoltre sviluppando tecniche di learning imbalance-aware basate sul dimensionamento del mini-batch e su tecniche di campionamento imbalance-aware del mini-batch per la predizione di varianti genetiche patogeniche con reti neurali profonde [C84].

A2. Metodi di ensemble gerarchici per la predizione strutturata in ontologie biologiche.
Concetti rilevanti nell'ambito della biologia molecolare (ad es: le funzioni dei geni e delle proteine) e della medicina (fenotipi anormali associati alle patologie umane) sono organizzati secondo ontologie gerarchiche strutturate come alberi (ad es: FunCat per la classificazione funzionale dei geni) o come grafi diretti aciclici (DAG) (ad es: la Gene Ontology (GO) per la classificazione dei geni e delle proteine e la HPO (Human Phenotype Ontology) per la classificazione dei fenotipi umani patologici).
In tale contesto ho sviluppati metodi di ensemble gerarchici [R39] basati sulla "true path rule" (TPR) [R29,C44,C48,C51] e metodi bayesiani cost-sensitive per la riconciliazione probabilistica dell'output dei base learner [R25,C50,C53], per la predizione strutturata delle funzione dei geni e delle proteine in ontologie strutturate ad albero. Ho quindi mostrato che la combinazione di metodi di ensemble gerarchici, strategie di learning cost-sensitive e l'integrazione di diverse tipologie di dati migliora significativamente le performance nella predizione delle funzioni dei geni a livello dell'intero genoma [R25,R32].
Nel contesto della predizione dei fenotipi umani anormali secondo la HPO, ho recentemente proposto nuovi metodi di ensemble gerarchici per la predizione strutturata basata su DAG che hanno conseguito risultati allo stato dell'arte [R50,C76,C79,C70,C71]. Recentemente nuovi metodi basati su algoritmi di isotonic regression combinati con l'algoritmo TPR hanno condotto a risultati allo stato dell'arte nell'ambito della predizione della funzione delle proteine [articolo in preparazione].

A3. 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 "omici" utilizzati, spesso caratterizzati da elevata dimensionalità. In tale contesto si sono esplorati diversi metodi di ensemble supervisionati, come i metodi basati su codici a correzione d'errore [R3,C11,C14], sulla riduzione della dimensionalità dei dati attraverso proiezioni randomizzate [R8,C22,C23], su metodi di bagging e sue varianti [R5,C18,C19] e sull'analisi della complessità dei dati [C37,C43].
Ho infine applicato metodi feature selection univariata ed SVM cost-sensitive all'analisi di immagini radiografiche (classificazione di noduli polmonari) con risultati comparabili con i migliori presenti in letteratura [R9,C25].

B. *Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di Machine Learning semi-supervisionato*.
Nell'ambito della Systems Biology e della Network Medicine ho sviluppato metodi di apprendimento semi-supervisionato basati su grafi, per lo studio dei sistemi biologici come entità complesse, in cui le funzioni biologiche derivano fondamentalmente dalle interrelazioni fra le parti che costituiscono il sistema.

B1. Metodi basati su funzioni di score kernelizzate per l'analisi di reti biomolecolari complesse.
Il problema della associazione di geni o proteine o più in generale di biomolecole ad una specifica proprietà biologica (ad es: problemi di predizione della funzione biomolecolare, o problemi di disease gene prioritization o drug repositioning) può essere modellato come un problema di node label ranking su un grafo. La maggioranza dei metodi proposti per l'analisi di grafi adotta strategie di apprendimento locali o globali per effettuare il ranking dei nodi o predire gli archi del grafo stesso. Per integrare strategie di learning locali e globali ho proposto metodi semi-supervisionati trasduttivi che tramite la kernelizzazione del grafo sono in grado di sfruttarne sia la topologia globale, sia di apprendere dalle caratteristiche locali di ogni nodo del grafo stesso [R46].

Tali metodi sono stati applicati con successo all'analisi di reti biomolecolari complesse per problemi di predizione della funzione della proteine, per problemi di ordinamento dei geni rispetto a patologie specifiche (disease gene prioritization), e per problemi di riposizionamento dei farmaci (ricerca di nuove indicazioni terapeutiche di farmaci originalmente progettati per scopi terapeutici differenti) [R31, R33, R35, R38, R41, C61].
Ho infine recentemente proposto nuovi metodi semi-supervisionati network-based basati su kernel per grafi che invece di analizzare lo "spazio dei biomarker", come usualmente avviene nell'ambito della Network Medicine, analizzano invece lo "spazio dei pazienti", tramite la costruzione di reti di pazienti basate sulla similarità dei loro profili biomolecolari (ad es: considerando i loro profili di espressione o genetici). Le "reti di pazienti" possono essere utilizzate sia a fini diagnostici o prognostici, sia per stratificare i pazienti stessi in sottotipi patologici, sia per individuare biomarker associati a patologie specifiche o alla risposta a farmaci [N21, paper sottoposto a PLOS Computational Biology].

B2. Metodi basati su Reti di Hopfield parameterizzate e cost-sensitive per la predizione della funzione dei geni.
Nel contesto della predizione della funzione delle proteine e dei geni secondo la GO, un problema rilevante è rappresentato dalla sbilanciamento nelle annotazioni. Infatti per la maggioranza delle classi GO solo un numero relativamente piccolo di geni annotati (esempi positivi) è disponibile. In questo contesto i metodi di apprendimento classici (comprese le Reti di Hopfield) tendono a predire sempre la classe di maggioranza. Per questa ragione ho progettato una nuova classe di Reti di Hopfield parameterizzate (COSNet) in grado di apprendere dai dati i propri parametri di learning, tenendo esplicitamente conto dello sbilanciamento fra esempi positivi e negativi [R36,R44,C58].
Una variante di COSNet, progettata per considerare esplicitamente "categorie" di neuroni note a priori nella rete è stata applicata con successo alla predizione della funzione delle proteine in un contesto multi-specie [R45]. Infine un metodo di integrazione di dati "imbalance-aware" accoppiato alle Reti di Hopfield parametriche ha ottenuto risultati allo stato dell'arte per la predizione delle funzioni delle proteine [R43,C67]

B3. Metodi scalabili per l'analisi di big biomolecular networks.
L'analisi di big data in reti biomolecolari di grandi dimensioni rappresenta un problema rilevante nell'ambito della biologia computazionale. In tale contesto ho sviluppato algoritmi "vertex-centric" ed utilizzato tecnologie come GraphChi basate sull'utilizzo efficiente della memoria secondaria per elaborare grafi di grandi dimensioni che non possono essere caricati in memoria primaria. L'obiettivo è di analizzare big data costruiti con dati "omici", con rilevanti applicazioni in Biologia Molecolare e Medicina, utilizzando semplici workstation stand-alone. Risultati sperimentali promettenti sono stati ottenuti nell'ambito della predizione multi-specie della funzione delle proteine [R42,C65,C74].
Un'altra linea di ricerca riguarda l'utilizzo della tecnologia GPU per l'implementazione massivamente parallela di algoritmi di node label prediction come COSNet per l'elaborazione efficiente di grafi di grandi dimensioni [R52], con applicazioni alla predizioni di classi GO, utilizzando la rete integrata multi-specie del database STRING [N20] che include milioni di proteine di diverse specie. E' in corso di sviluppo la costruzione di una rete costituita da decine di milioni di SNPs (single nucleotide polimorphisms) umane per la ricerca di SNPs potenzialmente deleterie o patologiche tramite la versione parallela [R52] di COSNet.

C. *Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di Machine Learning non-supervisionato.*

C1. 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 [R14,R15,R18,C29,C34,N10], e di algoritmi per l'analisi dell'affidabilità dei singoli cluster [R11,R12,C24], utilizzando un nuovo approccio basato sull'analisi della stabilità dei cluster ottenuti. Si sono inoltre sviluppati test statistici basati sulla distribuzione chi^2 [R15,C29] e sulla classica disuguaglianza di Bernstein [R18,C34] 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, utilizzando dati generati tramite bio-tecnologie high-throughput [R12,R13,R15,R18,R22,C39,N9,N10].

C2. 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 [C28]. Tali metodi sono stati successivamente applicati all'analisi di dati di espressione genica [C31]. 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 [C33] si è sviluppato uno schema algoritmico più generale da cui sono derivabili diversi algoritmi di fuzzy ensemble clustering [C36] e tale approccio è stato applicato all'analisi di dati di espressione genica per la ricerca di sottoclassi patologiche caratterizzate a livello bio-molecolare [R20].

D. *Metodi per l'integrazione di big data in ambito biologico e medico*.

D1. Algoritmi per la combinazione massiva di reti biomolecolari.
Il problema della integrazione di reti biomolecolari costruite con diverse tipologie di dati "omici" costituisce un problema rilevante nell'ambito della Systems Biology e della Network Medicine.
La mia attività di ricerca ha contribuito a mostrare che l'integrazione con metodi non pesati e pesati di reti biomolecolari diverse riveste un ruolo essenziale per individuare i geni associati a oltre 700 patologie [R41]. Ho inoltre proposto diverse metodologie per la costruzione e l'integrazione di reti biomolecolari, con applicazioni rilevanti alla predizione delle funzione delle proteine, ed ai problemi di disease gene prioritization e associazione farmaco – molecola bersaglio [R33,R38,R43,C63,C66,C67,C75].

D2. Metodi di ensemble supervisionati per l'integrazione di dati omici.
La mia attività di ricerca ha mostrato che metodi di ensemble anche relativamente semplici come la votazione maggioranza o i decision template possono ottenere risultati comparabili con lo stato dell'arte nell'integrazione di dati "omici" [R24,R26]. Altri miei lavori hanno confermato l'efficacia dei metodi di ensemble per l'integrazione di dati biomolecolari complessi [C41,C45,C46,C47], mostrando anche che i metodi di ensemble sono in grado anche di tollerare livelli relativamente elevati di rumore nei dati, senza un significativo deterioramento delle prestazioni [R27]. Ho inoltre condotti studi sull'applicazione dei metodi di ensemble per l'integrazione di dati eterogenei per la predizione della localizzazione subcellulare delle proteine [C54,C57], e sull'utilizzo di XML per l'integrazione di dati biomolecolari eterogenei [R23,C28,C40].

II. *Machine Learning*
Benchè lo sviluppo di nuovi metodi di apprendimento automatico sia stato motivato soprattutto da problemi complessi nell'ambito della Biologia Molecolare e della Medicina, ho sviluppato anche linee di ricerca di Machine Learning "puro", soprattutto per la progettazione ed analisi di metodi di ensemble di learning machine.

A. *Analisi e progettazione di metodi di ensemble*

A1. Metodi di hyper-ensemble per problemi di classificazione sbilanciati con big data.
Molti problemi di classificazione rilevanti, non solo nell'ambito della Medicina Genomica, sono caratterizzati da un forte sbilanciamento degli esempi disponibili fra le classi. In questo contesto i metodi di Machine Learning classici tendono ad essere severamente biased verso la classe maggioritaria (classe negativa) e non riescono ad apprendere gli esempi della classe minoritaria (classe positiva). Il metodo hyperSMURF, originariamente progettato per problemi fortemente sbilanciati in ambito genomico [R49] è abbastanza generale da essere applicato in altri contesti caratterizzati da un forte sbilanciamento dei dati. L'algoritmo, basato su tecniche di hyper-ensembling e di ricampionamento dei dati, ha raggiunto risultati allo stato dell'arte nell'ambito della Medicina Genomica [R48].
La versione parallellizzata, parSMURF, è in grado di elaborare big data, e migliora ulteriormente le prestazioni di hyperSMURF, grazie al tuning automatico dei parameteri di learning dell'algoritmo e scala sia su architetture multi-core sia su cluster di workstation, a seconda del livello di complessità del problema affrontato [articolo sottoposto a GigaScience].

A2. 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 puo' appartenere a più classi) e multi-path (le classi sono strutturate secondo alberi o DAG). Gli algoritmi di ensemble gerarchici che ho sviluppato per questa tipologia di problemi [R25,R29,R32,R50] hanno comunque una valenza più ampia e possono essere applicati in altri contesti caratterizzate da tassonomie gerarchiche strutturate ad albero o a DAG. Tali algoritmi adottano una strategia di learning a due step: dapprima le singole classi vengono apprese in modo "flat" dai base learner, quindi le predizioni dei modelli addestrati vengono combinate sfruttando la gerarchia delle classi. Recentemente ho sviluppato un algoritmo per tassonomie strutturate a DAG basato sulla integrazione di tecniche di isotonic regression e di combinazione bottom-up (dalle classi più specifiche alle classi più generali) che garantisce la consistenza delle predizioni e migliora sistematicamente le predizioni degli algoritmi di apprendimento "flat" [articolo in preparazione].

A3. Metodi di ensemble basati sulla scomposizione dell'errore in bias e varianza.
In questa linea 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, ho analizzato le relazioni fra apprendimento e scomposizione dell'errore in bias e varianza nel caso delle Support Vector Machine [R7].
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 [R7,C16].
Sfruttando l'analisi bias-varianza delle SVM, ho 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 [C20]. Il metodo è stato applicato con successo alla classificazione di malattie tumorali su base bio-molecolare [C19]. 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 [R10,C21].

A4. 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, utilizzando SVM come base learner [R8,C22]. Un'estensione del modello, 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 [C23].

A5. 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 ho analizzato l'efficacia dei metodi ECOC per problemi multi-classe in ensemble di learning machine e learning machine singole [R4,C2] e successivamente ho analizzato sperimentalmente la dipendenza fra gli errori a livello dei singoli bit dei codeword ECOC tramite misure basate sulla mutua informazione [R6], al fine di comparare differenti tipologie di codifiche ECOC e diverse architetture di sistemi ad apprendimento automatico basati su ECOC [C5,C7,C8].
Oltre alle applicazioni in ambito bioinformatico [R3,C11,C12], gli ensemble ECOC (insieme con metodi di boosting) sono stati applicati con successo anche per problemi multi-classe con nasi elettronici [R1,C3,C6,C15].

A6. 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, rappresentano un'estensione non supervisionata dei metodi dei random subspace supervisionati [C28].

In [C24] ho 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 [R13]. In conformità a questa analisi sono stati proposti metodi di ensemble clustering basati su proiezioni randomizzate [C31] che sono stati applicati con successo all'analisi di dati di DNA microarray [C28].
Un'estensione fuzzy del metodo di ensemble sviluppato in [C31] è stato sviluppato in [C36]: 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 ho derivato diversi metodi di fuzzy ensemble clustering [C36]. Tali metodi sono stati applicati all'analisi di dati di espressione genica [R20,C32,C33].

B. *Progettazione ed implementazione di librerie software di Machine Learning.*
L'attività di ricerca sia in ambito Machine Learning, sia in ambito bioinformatico è stata sempre accompagnata da attività di progettazione ed implementazione di librerie software.
Diversi metodi di ensemble che ho sviluppato sono stati resi disponibili ed implementati nella libreria C++, *NEURObjects*, inizialmente concepita per la progettazione software di reti neurali [R2,C1].
In seguito, parallelamente all'incremento delle attività di ricerca in ambito bioinformatico, ho progettate ed implementate librerie R open source per l'analisi di dati bio-molecolari complessi. In particolare la libreria *clusterv* [R11] permette di analizzare l'affidabilità di singoli cluster in dati bio-molecolari di elevata dimensionalità, la libreria *mosclust* [R14] permette di determinare il numero "ottimale" di cluster e di individuare strutture multiple presenti in dati bio-molecolari complessi, mentre la libreria *hcgene* [R17] consente di analizzare i grafi diretti aciclici della Gene Ontology e gli alberi di FunCat per supportare la classificazione funzionale delle proteine. La libreria RANKS [R46] rende disponibili algoritmi basati su grafi per problemi di node label ranking e di classificazione. La libreria COSNet implementa i nuovi modelli di Reti di Hopfield parametriche [R44], HEMDAG [R50] i metodi di ensemble gerarchici, hyperSMURF [R49] i metodi di hyper-ensembling "imbalance-aware" per l'analisi di dati fortemente sbilanciati, ed è in via di rilascio la sua estensione parallela (parSMURF) in C++ per l'analisi di big data genomici [articolo in preparazione].

Indici di produttività scientifica

Di seguito vengono riportati gli indici di produttività estratti il 7/4/2019 dalla banca dati Scopus e da

Google Scholar.

Indici	Scopus	Google Scholar
Numero di citazioni	1642	3261
Indice H	23	32
Indice i10	50	75

Direzione o partecipazione alle attività di un gruppo di ricerca caratterizzato da collaborazioni a livello nazionale o internazionale

A). Direzione di AnacletoLab, Laboratorio di Biologia Computazionale e Bioinformatica del Dipartimento di Informatica, Universita' degli Studi di Milano
http://anacletolab.di.unimi.it/.

Il laboratorio raccoglie un gruppo di docenti/ricercatori, assegnisti e dottorandi del DI e collabora con diversi gruppi di ricerca nazionali ed internazionali nell'ambito della Biologia Computazionale, Biologia Molecolare e Medicina, tra cui:
a) Computational Biology Group della Charite - Universitatamedizin della Humboldt Universitat di Berlino,
b) il gruppo di Biologia Computazionale del Berlin Institute of Health,
c) il Computer Science dept. della Royal Holloway, University of London,
d) i Jackson Laboratory for Genomic Medicine, CT, USA,
e) Il Computer Science Dept dell' Aalto University, Helsiniki, Finland,
f) il Wellcome Trust Sanger Institute e l'European Bioinformatics Institute (EBI) di Hinxton, UK,
g) la Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA,
h) l'European Center for Living Technologies di Venezia,
i) Altri gruppi di ricerca nazionali sia in ambito medico (Istituto Nazionale Tumori, e Ospedale S.Raffaele di Milano, Istituto Nazionale di Genetica Medica) e informatico (Univ. di Salerno e Univ. di Cagliari).
(Periodo: dal 2010 ad oggi)

B) Partecipazione come responsabile del laboratorio AnacletoLab di UNIMI alla challenge internazionale CAFA2 (Critical Assessment of Functional Annotation) nell' ambito dello Special Interest Group "Protein Function Prediction" di ISCB (International Society of Computational Biology). Lo Special Interest Group riunisce i principali gruppi di ricerca internazionali per la predizione della funzione delle proteine con metodi computazionali (http://biofunctionprediction.org/).

In qualita' di responsabile del gruppo di ricerca AnacletoLab ho partecipato alla attivita' di ricerca dello Special Interest Group "Protein Function Prediction" che ha portato alla pubblicazione su Genome Biology (una delle principali rivista di biologia computazionale) di un lavoro collettivo che coinvolge l'attivita' di ricerca di oltre 50 gruppi di ricerca di tutto il mondo. In tale ambito il laboratorio si e' collocato nei primi 3 nella challenge CAFA2 per la predizione dei geni umani associati a fenotipi patologici. AnacletoLab ha partecipato alla nuova challenge CAFA3 per la predizione della funzione delle proteine dell'uomo e dei principali organismi modello ed alla predizione dei geni associati a fenotipi umani anormali. I risultati preliminari di tale challenge (la valutazione è ancora in corso di svolgimento) collocano nuovamente AnacletoLab tra i gruppi “top ranked” a livello internazionale per la predizione della funzione delle proteine.
(Periodo: dal 2013 ad oggi).

C) Partecipazione all'attivita' di un gruppo di ricerca internazionale per la ricerca di mutazioni associate a malattie genetiche in regioni non codificanti del genoma.
Il mio contributo consiste nello sviluppo di metodi di machine learning specifici per questo problema.
Il gruppo di ricerca internazionale include diversi gruppi di ricerca in Europa, Nord America ed Australia, tra cui la Queen Mary University of London; Genomics England, UK; il Wellcome Trust Sanger Institute, Hinxton, UK,; il Max Planck Institute for Molecular Genetics, Berlin; il Department of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh; la Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA; il Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, USA; il Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, Australia; la Charité-Universitätsmedizin Berlin.

La metodologia computazionale sviluppata per la ricerca di mutazioni associate a malattie genetiche Mendeliane in regioni non codificanti del genoma e' stata pubblicata sull' *American Journal of Human Genetics*, rivista leader in questo ambito. Tale metodologia, denominata *Genomiser*, rappresenta lo stato dell'arte a livello internazionale per la ricerca di mutazioni causative di malattie genetiche Mendeliane. Il metodo di Machine Learning, sviluppato specificamente da me in collaborazione con il gruppo di Biologia Computazionale della Charitè UniversitatzMedizin di Berlino, e che costituisce il core di Genomiser, è stato recentemente pubblicato su una rivista del gruppo *Nature*. Tale metodo e' abbastanza generale da essere utilizzato in altri contesti per la ricerca di varianti genetiche associate a malattie genetiche e tumorali, ed infatti e' in corso una collaborazione in tal senso con la School of Medicine dell' Università dello Utah per studiare i fattori ereditari alla base di diverse patologie tumorali.
(Periodo: dal 2014 ad oggi)

D) Collaborazione ad un progetto comune con il Berlin Institute of Health (BIH, che riunisce gruppi di ricerca della von Humboldt e della Freie Universität di Berlino e del Max Delbrück Center for Molecular Medicine) per lo studio sistematico delle variazioni genetiche nelle regione regolatorie del genoma umano e del loro impatto sulle patologie genetiche e tumorali. Il progetto richiede lo sviluppo di metodologie di Intelligenza Artificiale innovative per l'analisi di dati generati da nuove biotecnologie, come i Massive Parallel Reporter Assay (MPRA) per lo studio dell'effetto funzionale in vivo delle varianti genetiche, e per l'analisi e l'integrazione massiva di dati epigenomici recentemente resi disponibili dall' International Human Epigenome Consortium (IHEC). L'obiettivo è di decodificare il "codice genetico" alla base della regolazione genica nelle regioni non codificanti del genoma, e di studiarne le alterazioni per predire l'effetto patologico delle mutazioni genetiche alla base delle patologie tumorali e genetiche.
Il progetto e' attualmente finanziato dall DAAD tedesco e dal MIUR, ma la complessità del problema richiede la collaborazione con altri partner scientifici europei ed americani e finanziamenti rilevanti per la generazione ed il processing bioinformatico dei dati biotecnologici. Per questa ragione stiamo lavorando insieme con il BIH ed altri partner europei ed americani per la preparazione di un progetto di ricerca europeo finanziato nell'ambito di H2020.
(Periodo: dal 2017 ad oggi)

E) Collaborazione con i Jackson Lab for Genomic Medicine (CT, USA) e con il Wellcome Sanger Istitute – Hinxton (Cambridge, UK) per la ricerca di mutazioni di splicing negli Exonic Sequence Enhancer (ESE, motivi che promuovono lo splicing degli esoni) del genoma umano. Tali mutazioni, estremamente difficili da individuare perchè apparentemente "silenti", sono alla base di diverse patologie tumorali. L'idea su cui stiamo lavorando è di costruire predittori basati su tecniche di deep learning in grado di individuare tali mutazioni sulla base di un insieme di opportune feature genomiche ed epigenomiche e sulla base di ESE etichettati manualmente tramite analisi della letteratura medica. Dopo la pubblicazione dei primi risultati (che appaiono molto promettenti) prepareremo un'application per un grant dell' NIH (National Institute of Health degli Stati Uniti) per finanziare questa linea di ricerca.
(Periodo: dal 2018 ad oggi)

F) Collaborazione di ricerca con il Dipartimento di Urologia dell'Ospedale S.Raffaele (OSR) di Milano, regolato da un Confidentiality Agreement stipulato fra UNIMI e OSR, per il progetto "Male infertility as a proxy of the progerois syndrome", di cui sono responsabile scientifico per UNIMI (Andrea Salonia, capo unità dell' URI è il responsabile scientifico per OSR). Tale agreement ha lo scopo di favorire attività di ricerca comuni per la proposizione di un grant all' NIH.
(Periodo: dal 2019 ad oggi)

G) Nell' a.a. 2015/16 ho usufruito di un anno sabbatico durante cui ho svolto attività di ricerca presso l'European Center for Living Technologies di Venezia (gruppo di ricerca del Prof. Pelillo), la Charitè, Facoltà di Medicina dell' Università von Humboldt di Berlino (gruppo di ricerca di Biologia Computazionale del prof. Robinson), ed il Dipartimento di Intelligenza Artificiale dell' Universita' di Granada (prof. Blanco).

Responsabilità scientifica (Principal Investigator – PI) di progetti di ricerca internazionali e nazionali, ammessi al finanziamento sulla base di bandi competitivi che prevedano la revisione tra pari

- *Responsabile scientifico per UNIMI della EU Collaborative Doctoral Partnership in Genomics and Bioinformatics*, finanziata dalla Comissione Europea in collaborazione con il Joint Research Centre della UE (2018-2022). La nostra Università e' stata selezionata tra le prime 5 Università europee per l'area della Genomica e Bioinformatica. Il contratto di collaborazione di durata quinquennale è rinnovabile alla fine del quinquennio, previo accordo tra le parti e prevede il finanziamento da parte della Commissione Europea di dottorati presso il Dottorato di Informatica di UNIMI, finalizzati alla formazione di ricercatori nell'ambito della Genomica e della Bioinformatica in grado sia di svolgere ricerca innovative in tale ambito sia di supportare scientificamente la Commissione Europea nell'ambito delle politiche europee per la Genomica Medica e la Data Analytics in campo sanitario.

- *PI del progetto* "Developing machine learning methods for the prioritization of regulatory variants in human disease" in collaborazione con il Berlin Institute of Health, finanziato dal MIUR e dal DAAD (Germania) (2018-2019)

- *PI di "HyperGeV* : Detection of Deleterious Genetic Variation through Hyper-ensemble Methods " (2016-2018) , finanziato dal CINECA e dalla Regione Lombardia

- *PI di "HPC-SoMuC*: Development of Innovative HPC Methods for the Detection of Somatic Mutations in Cancer" (2017-2018), finanziato dal CINECA e dalla Regione Lombardia.

- *Responsabile dell' unita' UNIMI del progetto* "A composite predictive model of response to Fingolimod: integration of clinics, neuroimaging and genomics", finanziato dalla Fondazione Italiana Sclerosi Multipla. Il capofila (PI) del progetto è il Laboratorio di Genetica delle Malattie Neurologiche Complesse dell' Ospedale S.Raffaele di Milano. I biomarker genetici associati alla risposta al farmaco Fingolimod, individuati tramite i dati del S.Raffaele ed il modello predittivo sviluppato dalla unità UNIMI sono in corso di brevettazione (2016-2018).

- *Responsabile dell'unità UNIMI* (tramite subcontracting) del progetto *Finding-MS* nell'ambito del progetto europeo ERA-PerMed Joint Transnational Call (JTC) 2018 on "Research projects on Personalized Medicine – smart combination per pre-clinical and clinical research with data and ICT solutions". Il progetto e' in collaborazione con l'Ospedale S.Raffaele di Milano, il CNR – Istituto per le Tecnologie Bio-mediche, il Centre Hospitalier Universitaire de Toulouse e geneXplain, un' azienda bioinformatica tedesca (2019-2022)

Partecipazione a progetti di ricerca internazionali e nazionali, ammessi al finanziamento sulla base di bandi competitivi che prevedano la revisione tra pari

- Progetto PRIN "Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond", coordinato dall'Università di Pisa (2019-2021)
- Progetto "Discovering Patterns in Multi-Dimensional Data" (2016-2017) finanziato dall' Università degli Studi di Milano.
- Partecipazione all' unità milanese (responsabile N. Cesa-Bianchi) del Network of Excellence "Pattern Analysis, Statistical Modelling and Computational Learning 2 (PASCAL2)", 7th European Framework Programme ffinanziato dall'Unione Europea (2007-2013);
- Partecipazione all' unità milanese (responsabile N. Cesa-Bianchi) del Network of excellence PASCAL, 6th European Framework Programme (2004-2006)
- Progetto Computational methods for bio-medical pattern analysis finanziata da UNIMI (2011-2013)
- Progetto MIUR COFIN-PRIN Automata and formal languages: mathematical and applicative aspects (2010-11);
- Progetto PUR 2009: "Metodi automatici per l'analisi di pattern in ambito biomedico" finanziato dall' Università degli Studi di Milano;
- Progetto PUR 2008: "Modelli computazionali innovativi" finanziato dall' Università degli Studi di Milano;
- Progetto IEIIT-CNR 2007: Machine Learning Techniques for Modeling and Growing Up" (2007-2008);
- Progetto PUR 2006: "Modelli stocastici e quantistici per problemi computazionali e di bioinformatica" finanziato dall' Università degli Studi di Milano;
- Progetto PUR 2006: "Identificazione di profili trascrizionali nella Leucemia Mieloide Acuta mediante microarray su frazioni di staminali ematopoietiche" finanziato dall' Università degli Studi di Milano;
- Progetto MIUR COFIN-PRIN Formal Languages and automata: methods, models and applications (2003-2005)
- Progetto MIUR COFIN-PRIN Machine Learning Techniques for Bioinformatics: Analysis and Modeling of Functional and Structural Data of Gene Expression (2001-2002)
- From Bits to Information: Statistical Learning Technologies for Digital Information Management Search – USA, Progetto finanziato dalla National Scientific Foundation (NSF) (2002).

Partecipazione a comitati editoriali di riviste internazionali

- Membro dell' Editorial board di *Scientific Reports, Nature* (IF 2017: 4.122), indicizzata in Scopus ed ISI Web of Science (dal 1 giugno 2018).
- Membro dell' editorial board di *Advances in Bioinformatics, Hindawi* (IF 2017: 1.8), indicizzata in Scopus (dal 2014)
- Guest Editor di *Artificial Intelligence in Medicine, Elsevier* (IF: 2.879) indicizzata in Scopus ed ISI Web of Science per le Special issue "Computational Intelligence and Machine Learning in Bioinformatics" (2008-2009).
- Sono stato inoltre membro dell' Editorial board di altre riviste minori di area Bioinformatica e Machine Learning dal 2008 al 2017.

Attività di reviewer per riviste internazionali

Ho svolto e svolgo attività di reviewer per le principali riviste internazionali nell' ambito della Bioinformatica e del Machine Learning, tra cui:

- Journal of Machine Learning Research
- Machine Learning
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Pattern Recognition
- IEEE Transactions on Neural Networks and Learning Systems
- IEEE Transactions on Knowledge and Data Engineering
- Artificial Intelligence
- Computational Intelligence,
- Neurocomputing
- Neural Networks
- IEEE Transactions on Cybernetics,
- GigaScience
- Bioinformatics
- PLoS Computational Biology
- BMC Bioinformatics
- Briefings in Bioinformatics
- IEEE ACM Transactions on Computational Biology and Bioinformatics
- Artificial Intelligence in Medicine
- Journal of Bioinformatics and Computational Biology
- PLoS One

Organizzazione Conferenze e Workshop

- Co-organizzatore del workshop Soft computing methods for characterizing diseases from omics data – CIBB 2019, Bergamo 3-4 settembre 2019
- Chair di CIBB 2018 - Computational Intelligence methods for Bioinformatics and Biostatistics, Lisboa (Portogallo) (Caparica 6-8 Settembre 2018)
- Chair del Workshop BigTargets all' European Conference on Machine Learning (ECML), Porto (Portogallo), 2015;
- Chair del Fourth Italian Workshop on Machine Learning and Data Mining (MLDM 2015) – XIV Conference of AIIA – Pisa, 2015;
- Chair di SUEMA 2010, Third International Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications at ECML – European Conference on Machine Learning 2010, Barcelona, Spagna;

- Wokshop Learning from Multiple Sources with Applications to Robotics at NIPS 2009, Whistler, Canada, 2009;
- Chair di SUEMA 2008, European Conference on Artifical intelligence (ECAI) 2008 Patras, Grecia;
- Chair di CIBB 2007, Fourth International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics, Portofino, Italia
- Chair di SUEMA 2007, International Workshop on Supervised and Unsupervised Ensemble Methods and Their Application (nell'ambito di IbPRIA2007) a Girona (Spagna).

Partecipazione al comitato scientifico di programma di conferenze internazionali.

Membro del Comitato Scientifico di Programma di oltre 50 conferenze internazionali nell'ambito del machine learning e della biologia computazionale dal 2007 ad oggi, tra cui:

- European Conference in Computational Biology
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- International Joint Conference on Artificial Intelligence (IJCAI) - Machine Learning Track
- International Symposium on Foundations and Applications of Big Data Analytics
- Automatic Function Prediction - Critical Assessment of Functional Annotation experiment (nell'ambito di ISMB – Intelligent Systems for Molecular Biology)
- ICANN - International Conference on Artificial Neural Networks
- SIAM International Conference on Data Mining
- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
- S+SSPR IAPR Joint International Workshops on Statistical + Structural and Syntactic Pattern Recognition
- International Conference on Pattern Recognition (ICPR)
- International Conference on Multiple Classifier Systems (MCS)

Relatore invitato a congressi e convegni internazionale e nazionali

- Invited talk allo Statistical Workshop "Statistical approaches and validation in clustering: mixture models and nonparametric methods", organizzato dal Dipartimento di Statistica dell' Universita' di Caen (Francia), dal titolo "Stability-based methods for cluster validation".
  (29-06-2007)
- Invited talk a CIBB 2008, FIFTH INTERNATIONAL MEETING ON COMPUTATIONAL INTELLIGENCE METHODS FOR BIOINFORMATICS AND BIOSTATISTICS
  IIASS - Vietri sul Mare, Salerno (Italy) dal titolo: "Unsupervised stability-based ensembles to discover reliable structures in complex bio-molecular data".
  (04-10-2008)
- Invited talk al Third Italian Workshop on Machine Learning and Data Mining - XIII AI*IA Symposium on Artificial Intelligence, Pisa December 2014,
  "Analysis of bio-molecular networks through semi-supervised graph-based learning methods".
  http://aiia2014.di.unipi.it/mldm/index
  (10-12-2014)
- Invited talk al Fifth Italian Workshop on Machine Learning and Data Mining - XV AI*IA Symposium on Artificial Intelligence, Genova, Novembre 2016,
  "A hyper-ensemble approach for the genome-wide prediction of disease and trait-associated genetic variants".
  (28-11-2016)
- Invited speaker al Workshop Interdisciplinary Aspects of Biomolecular Modelling
  "Machine Learning approaches for Modelling Complex Biomolecular Systems"
  (26-6-2019)
- Invited talk all' Annual Workshop in Bioinformatics and Genomics of the Catalan Society of Biology and the Bioinformatics Barcelona association,

"Machine Learning for Computational Biology and Precision Medicine"
(17-12-2019)

Relatore invitato presso Università e Centri di ricerca italiani ed europei.

Ho svolto diversi seminari su argomenti di Machine learning e Bioinformatica in diverse unità e centri di ricerca italiani e stranieri dal 2005 ad oggi, tra cui:

- Department of Computer Science della Royal Holloway, University of London;
- Department of Computer Science, Aalto University, Helsinki;
- Department of Computer Science, Aristotle University of Thessaloniki;
- Charitè Universitatzmedizin Berlin
- European Center for Living Technologies, Venezia;
- Departamento de Sistemas Informáticos y Computación Universitat Politècnica de València;
- Computational Genomics Department del Centro de Investigacion Prencipe Felipe, Valencia;
- School in Bioinformatics, University of Brno (Czech Republic);
- Computer Science and Artificial Intelligence Department, University of Granada;
- Dipartimento di Informatica, Università di Pisa;
- Dipartimento di Matematica e Informatica dell' Università di Palermo;
- Dipartimento di Ingegneria ed Elettronica (DIEE) Universita' di Cagliari;
- IIASS Istituto Internazionale Alti Studi Scientifici "E. R. Caianiello di Vietri (Salerno);
- Dipartimento di Informatica dell' Università degli Studi di Salerno;
- IST - Istituto Nazionale per la Ricerca sul Cancro, Genova.
- INT – Istituto Nazionale Tumori, Milano
- Dipartimento di Biologia e Genetica per le Scienze Mediche, Università degli Studi di Milano

Premi e riconoscimenti internazionali.

- L'articolo M. Notaro, M. Schubach, P.N. Robinson, G. Valentini. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble method, BMC Bioinformatics, vol. 18 (1), 2017 è stato premiato come uno dei 5 migliori articoli dell'anno dall' International Medical Informatics Association (IMIA) per la sezione "Knowledge Representation and Management".

Brevetti
E' in fase di brevettazione una signature di 21 SNP (Singe Nucleotide Polymorphism) per la predizione della risposta di pazienti affetti da sclerosi multipla al farmaco Fingolimod, realizzata tramite un algoritmo ed ed un software da me sviluppato ed applicato all'analisi di dati genomici forniti dal dipartimento di Neurologia dell'Ospedale San Raffaele di Milano.

Attività di valutazione nell'ambito di procedure di selezione competitive nazionali e internazionali.

- Research expert per la Commissione Europea (number: EX2014D182522) per la valutazione di proposte, progetti e programmi sottoposti a valutazione per finanziamenti europei (dal 2014)
- Revisore iscritto a REPRISE (albo degli esperti scientifici istituito presso il MIUR) per la sezione "ricerca di base": Settori scientifico-disciplinari: Informatica (INF/01) Settori ERC: Computer Science and Informatics: Informatics and Information Systems, copmuter science, scientific computing and intelligent systems (PE6), Bioinformatics, biocomputing, and DNA and molecular computation (PE6_13), Diagnostic tools (e.g. genetic, imaging) (LS7_2), Computational biology (LS2_11) (dal 2018).
- Reviewer per proposal sottoposte alla Netherlands Organisation for Scientific Research (NWO):

- Reviewer del progetto “Collaborative experimentation for data mining” (2011-2012)
- Reviewer del progetto “Context-aware protein function prediction” (2012/13)

- Reviewer per proposal sottoposte alla Research Foundation – Flanders (Fonds Wetenschappelijk Onderzoek - Vlaanderen, FWO):
  - Reviewer del progetto “COCO-MULT: Constrained Constructive Machine Learning for Multiple Targets “ (2018)
- Reviewer per progetti SIR (Scientific Independence of young Researchers) finanziati dal MIUR (2014/15)
- Reviewer per progetti del programma "Bloodwise" per lo studio dei tumori ematologici, finanziato da 14M Genomics, una company per la diagnostica spin-out della Wellcome Trust Sanger Institute di Hinxton, UK.
  (dal 2012 al 2016)

Formale attribuzione di incarichi di insegnamento o di ricerca (fellowship) presso qualificati atenei e istituti di ricerca esteri o sovranazionali

- *Research fellowship* - Computer Science Dept. - Oregon State University - USA: sviluppo di metodi di ensemble basati sull'analisi bias-varianza.
  (dal 01-03-2001 al 30-04-2001)
- *Visiting researcher* - Centro de Investigacion Principe Felipe, Valencia, Spain (2009)
  (dal 14-09-2009 al 15-10-2009)
- *Visiting professor* al Dipartimento di Informatica dell' Aristotle University of Thessaloniki per attivita' di ricerca comuni nell'ambito di metodi di machine learning per problemi di predizione multi-target caratterizzati da elevata dimensionalita' sia nello spazio di input, sia nello spazio di output.
  (dal 29-09-2014 al 28-10-2014)
- *Visiting professor* al Dipartimento di Computer Science della Royal Holloway, University of London per attivita' di ricerca comuni (sviluppo di metodi per l'integrazione e l'analisi di reti biomolecolari complesse) e per lo svolgimento di seminari di supporto alla didattica del master e del dottorato in Computer Science della Royal Holloway.
  (marzo 2015)
- *Fellowship per visiting professor* presso il Computer Science Department della Aalto Univeristy (Helsinki) nell'ambito del progetto "Machine Learning for Metagenomics", finanziato dall' Aalto Science Foundation di Helsinki (AScI Visiting Fellow Programme 2014-15).
  (dal 25-05-2015 al 25-06-2015)
- *Visiting professor* all' European Center for Living Technologies (ECLT). Durante la visita si sono rafforzate le attivita' di ricerca comune nell'ambito della game theory e dei metodi semi-supervisionati network-based per l'analisi di reti bio-molecolari.
  (dal 01-02-2016 al 28-02-2016)
- *Fellowship per visiting professor* presso il gruppo di Biologia computazionale della Charité - Universitätsmedizin Berlin nell'ambito del Funding programme Research Stays for University Academics and Scientists finanziato dal DAAD (German Academic Exchange Service) per il progetto "Hierarchical ensemble methods for structured predictions in Biological Ontologies".
  (dal 01-03-2016 al 1-06-2016)
- *Visiting professor* al Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada. Durante la visita si sono sviluppate attivita' di ricerca comuni per lo sviluppo di algoritmi graph-based e la loro applicazione a problemi di ranking dei geni associati a patologie specifiche e a problemi di "riposizionamento" (uso terapeutico alternativo) di farmaci.
  (dal 07-06-2016 al 28-06-2016)

Responsabilità di studi e ricerche scientifiche affidati da istituzioni pubbliche o private

- Incarico di ricerca per lo sviluppo e modellazione di metodi di machine learning con applicazioni in biologia computazionale, come membro affiliato al CNR nell'ambito del progetto "Machine Learning Techniques for Modeling and Growing Up", affidato dallo IEIIT-CNR (2007-2008)

- Contratto di consulenza con la ditta SIGEDA di Milano per lo sviluppo di metodi di ensemble basati su codici a correzione d'errore per la classificazione multiclasse di dati provenienti da un sistema olfattivo artificiale (naso elettronico)
  (2007)
- Responsabilita' di un progetto di ricerca per la predizione con metodi di machine learning di fenotipi tumorali quantitativi in organismi modello e per lo studio della suscettibilita' alla cancerogenesi polmonare ed epatica, affidato dall'unita' di ricerca di Epidemiologia genetica e Farmacogenomica dell'Istituto Nazionale Tumori (INT) di Milano.
  (dal 01-04-2015 al 1-10-2017)
- Incarico della Fondazione Centro S.Raffaele di Milano per lo sviluppo di un modello predittivo di risposta al trattamento con farmaci di II livello per pazienti affetti da Sclerosi Multipla.
  (dal giugno 2016 al giugno 2018)

Progettazione e sviluppo di librerie software rese disponibili alla comunità scientifica in repository pubbliche.

- HEMDAG: Hierarchical Ensemble Methods for Directed Acyclic Graphs https://cran.r-project.org/web/packages/HEMDAG
- HyperSMURF: Hyper-Ensemble Smote Undersampled Random Forests (software library for supervised prediction with highly imbalanced big data).
  https://cran.r-project.org/web/packages/hyperSMURF
- RANKS: Ranking of Nodes with Kernelized Score Functions (software library for node label learning in graphs).
  https://cran.r-project.org/web/packages/RANKS
- COSNET: Cost Sensitive Network for node label prediction on graphs with highly unbalanced labels.
  http://www.bioconductor.org/packages/devel/bioc/html/COSNet.html;
- HCGene: an R package to support the hierarchical classification of genes.
  https://homes.di.unimi.it/valentini/SW/hcgene;
- Clusterv: an R package for cluster validation.
  https://homes.di.unimi.it/valentini/SW/clusterv;
- Mosclust: an R package for the discovery of significant structures in bio-molecular data.
  https://homes.di.unimi.it/valentini/SW/mosclust;
- NEURObjects a set of C++ library classes for neural networks development.
  http://homes.di.unimi.it/valentini/SW/NEURObjects;
- PerfMeas: an R package implementing different performance measure for classification and ranking tasks.
  http://cran.r-project.org/web/packages/PerfMeas;
- NetPreProc an R package that implements preprocessing and normalization methods for network-structured data.
  http://cran.r-project.org/web/packages/NetPreProc;
- Bionetdata an R data package that includes several examples of chemical and biological data networks.
  http://cran.r-project.org/web/packages/bionetdata.

Partecipazione a Società ed Associazioni scientifiche nazionale ed internazionali

Sono membro di ISCB - International Society of Computational Biology, di BITS - Società Italiana di Bioinformatica, di INNS (International Neural Network Society), del Data Mining and Big Data Analytics Technique Committee (DMTC) – IEEE Computational Intelligence Society (IEEE-CIS), del comitato

scientifico del working group on Machine Learning and Data Mining, all'interno di AI*IA, Italian Association for Artificial Intelligence. Membro del laboratorio AIIS del CINI

Attività didattica

- L'attività didattica si è articolata sia attraverso i corsi istituzionali, svolti all'Università degli Studi di Milano sia attraverso i corsi di dottorato, master e corsi tenuti all'Università di Milano ed in altri atenei soprattutto nell'ambito della Bioinformatica e dell'Apprendimento Automatico.
- In particolare ho tenuto corsi per la laurea triennale e magistrale in Informatica, per la laurea in Comunicazione digitale per la laurea magistrale in Biotecnologie Molecolari e Bioinformatica e per la laurea in Biotecnologie industriali ed ambientali, nonché coris di Informatica di base per diverse lauree triennali di Medicina. Ho anche tenuto e tengo tuttora corsi in inglese.
- Attualmente sono responsabile anche del corso di Informatica Generale per la laurea in Scienze e tecnologie per lo studio e la conservazione dei beni culturali e dei supporti dell'informazione (il Dipartimento di Informatica è dipartimento associato a tale corso di laurea).
- Partecipo al *collegio dei docenti del Dottorato in Informatica* dell'Università degli Studi di Milano dal 2008. In tale contesto sono responsabile del corso di *Machine Learning for Genomic Medicine*.
- Dall' anno accademico 2018/19 sono membro del comitato coordinatore per il Dipartimento di Informatica del *master in Bioinformatics and Functional Genomics* dell'Università degli Studi di Milano, svolto in collaborazione con l'Istituto Nazionale di Genetica Medica (INGM) ed il Policlinico dell'Università degli Studi di Milano. Nell'a.a. 2017/18 ho gia' svolto alcune lezioni (Machine Learning for Personalized and Precision Medicine) nell'ambito del medesimo master.
- *Sono responsabile scientifico per UNIMI della EU Collaborative Doctoral Partnership in Genomics and Bioinformatics*, finanziata dalla Comissione Europea in collaborazione con il Joint Research Centre (JRC) della UE. Il piano quinquennale del dottorato collaborativo fra UNIMI e JRC è finalizzato alla formazione di ricercatori nell'ambito della Genomica e della Bioinformatica in grado sia di svolgere ricerca innovative in tale ambito sia di supportare scientificamente le decisioni politiche della Commissione per la Genomica Medica e la Data Analytics in campo sanitario.

La seguente tabella riassume gli incarichi didattici svolti in UNIMI a partire dall' a.a. 2004/05 ad oggi:

a.a	corso	corso di laurea	Liv.
18/19	Machine Learning for Genomic Medicine	Corso di dottorato in Informatica	D
18/19	Bioinformatica	Informatica	M
18/19	Informatica Generale	Scienze e tecnologie per lo studio e la conservazione dei beni culturali e dei supporti dell'informazione (il Dipartimento di Informatica è associato a tale corso di laurea)	T
18/19	Machine Learning for Personalized and Precision Medicine	master in Bioinformatics and Functional Genomics UNIMI	Master
17/18	Bioinformatica	Informatica	M
17/18	Informazione Multimediale	Comunicazione Digitale	T
17/18	Informatica Generale	Scienze e tecnologie per lo studio e la conservazione dei beni culturali e dei supporti dell'informazione	T
16/17	Bioinformatica	Informatica	M
16/17	Bioinformatics Methods (in inglese)	Molecular Biotechnologies and Bioinformatics	M
16/17	Bioinformatica e Biostatistica	Biotecnologie Industriali ed Ambientali	T
14/15	Bioinformatica	Informatica	M
14/15	Metodi Bioinformatici	Biotecnologie Biomolecolari e Bioinformatica	M
14/15	Informatica Avanzata	Biotecnologie Industriali ed Ambientali	T
13/14	Bioinformatica	Informatica	M
13/14	Metodi Bioinformatici	Biotecnologie Biomolecolari e Bioinformatica	M

13/14	Informatica Avanzata	Biotecnologie Industriali ed Ambientali	T
12/13	Bioinformatica	Informatica	M
12/13	Metodi Bioinformatici	Biotecnologie Biomolecolari e Bioinformatica	M
12/13	Informatica Avanzata	Biotecnologie Industriali ed Ambientali	T
11/12	Bioinformatica	Informatica	M
11/12	Metodi Bioinformatici	Biotecnologie Biomolecolari e Bioinformatica	M
11/12	Informatica Avanzata	Biotecnologie Industriali ed Ambientali	T
10/11	Bioinformatica	Informatica	M
10/11	Metodi Bioinformatici	Biotecnologie Biomolecolari e Bioinformatica	M
10/11	Informatica Avanzata	Biotecnologie Industriali ed Ambientali	T
09/10	Bioinformatica	Informatica	M
09/10	Metodi Bioinformatici	Biotecnologie Biomolecolari e Bioinformatica	M
08/09	Linguaggi di programmazione per la bioinformatica	Genomica Funzionale e Bioinformatica	M
08/09	Bioinformatica	Informatica	M
08/09	Informatica applicata ai processi biotecnologici	Biotecnologie Industriale e Ambientali	T
07/08	Linguaggi di programmazione per la bioinformatica	Genomica Funzionale e Bioinformatica	M
07/08	Bioinformatica	Informatica	M
07/08	Informatica	Fisioterapia e Dietistica	T
06/07	Linguaggi di programmazione per la bioinformatica	Genomica Funzionale e Bioinformatica	M
06/07	Bioinformatica	Informatica	M
06/07	Informatica	Podologia ed Igiene Dentale	T
05/06	Linguaggi di programmazione per la bioinformatica	Genomica Funzionale e Bioinformatica	M
05/06	Bioinformatica	Informatica	M
05/06	Informatica	Podologia ed Igiene Dentale	T
04/05	Linguaggi di programmazione per la bioinformatica	Genomica Funzionale e Bioinformatica	M
04/05	Algoritmi per la Bioinformatica	Genomica Funzionale e Bioinformatica	M
04/05	Laboratorio di Bioinformatica	Biologia	T

M : laurea magistrale; T : laurea triennale; D : corso di dottorato

- Precedentemente ho tenuto il corso di Bioinformatica Funzionale 1 per il corso di laurea in Fisica dell'Università di Genova negli a.a. 2002/03 e 2004/04 e le esercitazioni di Reti Neurali 1 ed il Laboratorio di programmazione di Sistema (a.a. 2001/02) per il corso di laurea in Scienze dell'Informazione dell'Università di Genova.
- Ho avuto incarichi di insegnamento per corsi di dottorato in diverse Università, tra cui l'Universita' di Palermo (Dottorato in Informatica e Matematica), l'Università di Salerno (Dottorato in Informatica), l'Università of Brno (Czech Republic) (Summer Doctoral School in Bioinformatics), il Dipartimento di Computer Science della Royal Holloway, University of London.
- Sono stato docente e coordinatore di corsi sul linguaggio R per l'analisi di dati bio-medici indirizzati a dottorandi, assegnisti e ricercatori presso diversi centri di ricerca, tra cui:
  - "An R course for Oncology Bioinformatics", CINECA - Casalecchio di Reno, giugno 2008

- “Corso sul linguaggio R per l’ analisi di dati di DNA microarray con metodi di machine learning”, IST Genova, Novembre 2009
- “Analisi di dati biomolecolari con R”, Parco Tecnologico Padano di Lodi, 21-23 gennaio 2009

Ho tenuto diversi seminari e corsi di dottorato in diverse Università italiane ed europee, tra cui

- "Discovering significant structures in bio-molecular data" corso di dottorato 3rd International Summer School on Computational Biology, Mikulov, Czech Republic, 13 -15 August 2007
- “Ensemble methods in bioinformatics” corso per il Dottorato in Matematica ed Informatica all'Università di Palermo (22-26 giugno 2008)
- "Supervised Gene Set Analysis for the molecular characterization of patients: a machine learning approach", seminario presso il Centro de Investigacion Principe Felipe di Valencia, 6-10-2009
- "True Path Rule hierarchical ensembles for genome-wide gene function prediction: a machine learning approach", seminario per il corso di dottorato in Computer Science de l'Universitat Politecnica de Valencia, 13-10-2009
- "Hierachical ensemble methods for gene function prediction" seminario per il dottorato in Information Engineering, Dipartimento di Ingegneria Elettrica ed Elettronica (DIEE) dell'Università di Cagliari (maggio 2010)
- “True Path Rule and H-Bayes hierarchical cost-sensitive ensembles for gene function prediction” tenuto presso l’Istituto Internazionale per gli Alti Studi Scientifici “E.R. Caianiello”, Vietri sul mare (Salerno) il 27 maggio 2010
- “True Path Rule hierarchical ensemble methods for gene function inference” seminario per il corso di dottorato in Informatica tenuto presso il presso il Dipartimento di Informatica, Università di Salerno il 6 giugno 2011
- “Stability-based methods for the assessment of clusters discovered in bio-molecular data” corso per il Dottorato in Matematica ed Informatica all'Università di Palermo, giugno 2012
- “Ensemble methods for multi-target classification and regression”, seminario al Dipartimento di Informatica dell' Aristotle University of Thessaloniki, Grecia, 14 ottobre 2014
- “Scalable methods for bio-molecular network analysis and for structured prediction in biological ontologies” seminario all'European Center for Living technologies - Univ. Ca' Foscari, Venezia, 27 febbraio 2015
- "Analysis of bio-molecular networks through semi-supervised graph-based learning methods" seminario per il dottorato in Computer Science, Computer Science Dept. Royal Holloway, University of London, 9 March 2015
- "Multi-label hierarchical prediction methods and their application to the automatic function prediction of proteins" seminario per il dottorato in Computer Science, Computer Science Dept. Royal Holloway, University of London, 10 March 2015
- “Scalable methods for the analysis of complex biomolecular networks”, seminario tenuto presso il Dipartimento di Informatica, Università di Salerno il 5 Maggio 2015
- "Semi-supervised graph-based learning methods for the analysis of bio-molecular networks" seminario per il corso di dottorato in Computer Science della Aalto University, Helsinki, Finland, 13 June 2015
- “Semi-supervised graph-kernel methods for disease gene prioritization”, seminario presso il Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, Spagna, 15-6-2016
- “A Machine Learning and High Performance Computing tool to prioritize pathogenic variants in the human genome”, seminario per il dottorato in Computer Science, Computer Science Dept. Royal Holloway, University of London, 26 Febbraio 2019

Sono stato membro di diverse commissioni per l'esame finale di dottorato e revisore di tesi di dottorato per diverse Università nazionali ed europee:

- Membro della commissione d'esame finale per il dottorato in Computer Science dell'Universita' di Salerno (aprile 2010)
- Membro della commissione d'esame finale per il dottorato in Information Engineering organizzato dal Dipartimento di Ingegneria Elettrica ed Elettronica (DIEE) dell'Università di Cagliari (maggio 2010).
- Membro della commissione d'esame finale per il dottorato in Computer Science dell'Universita' di Salerno (febbraio 2011)
- Membro della commissione d'esame finale per il dottorato in informatica del DIBRIS Università di Genova (maggio 2018)
- Membro della commmisione d'esame finale per il dottorato in Computer Science della Freie Universitat Berlin (settembre 2018)
- Membro della commmisione d'esame finale per il dottorato in Computer Science della Royal Holloway – University of London (febbraio 2019)
- Revisore della tesi di dottorato di Isabel Segura Bedmar (Computer Science Department, Universidad Carlos III de Madrid) "Application of Information Extraction techniques to pharmacological domain: Extracting drug-drug interactions" (2010)
- Revisore della tesi di dottorato di Francesco Iorio (Dottorato in Computer Science - Universita' degli Studi di Salerno) "Automatic Discovery of Drug Mode of Action and Drug Repositioning from Gene Expression Data" (2010)
- Revisore della tesi di dottorato di Luca Pinello (Dipartimento di Matematica e Informatica - Universita' degli Studi di Palermo) "Multi Layer Analysis" (2012)
- Revisore della tesi di dottorato di Carmen Navarro (Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Espana) "Approach to Personalized Medicine through the development of new Artificial Intelligence methodologies" (2017).
- Revisore della tesi di dottorato di Guido Zampieri (Ph.D. Course in Biosciences - Curriculum: Genetics, Genomics and Bioinformatics, Università degli Studi di Padova) "Prioritisation of candidate disease genes via multi-omics data integration". (2017)
- Revisore della tesi di Master of Science in Artificial Intelligence di Jeremy Borg (Faculty of ICT, University of Malta) "Improved Performance of Error Correcting Output Codes for Multiclass Classification" (2017)
- Revisore della tesi di dottorato di Samuele Fiorini (Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi Università degli Studi di Genova) "Challenges in biomedical data science: data-driven solutions to clinical questions". (2018)
- Revisore della tesi di dottorato di Max Schubach (Department of Mathematics and Computer Science, Freie Universitat Berlin) "Learning the Non-Coding Genome" (2018)
- Revisore della tesi di dottorato di Juan Caceres Silva (Department of Computer Science, Royal Holloway – University of London), "Network Medicine Characterisation of Genetic Disorders by Propagation of Disease Phenotypic Similarities". (2019)

Attività di didattica integrativa e di sostegno agli studenti.

- Tutor, relatore o correlatore di tesi per i seguenti dottorandi di ricerca per il Dottorato di Ricerca in Informatica, Università degli Studi di Milano:

  1) Francesca Ruffino: "Supervised Learning Methods for the Analysis of Gene Expression Data"

2) Raffaella Folgieri: “Ensembles based on Random Projection for gene expression data analysis”
3) Roberto Avogadri: “Unsupervised clustering methods for high-dimensional data anslysis”
4) Francesco Saccà (dottorato in Matematica e Statistica per le Scienza Computazionali): “Problemi di Clustering con vincoli: algoritmi e complessita'”
5) Marco Frasca: “Graph-based approaches for imbalanced data in functional genomics”
6) Rajab Ali Keshavarz Emami: ”Machine learning methods for the prediction of epileptic seizures”
7) Marco Notaro: “Hierarchical Ensemble Methods for Ontology-based Predictions in Computational Biology”
8) Alessandro Petrini: “Machine learning methods for the the prediction of pathogenic variants in non coding regions of the human genome”

- Supervisore dei seguenti assegnisti di ricerca: 1) Matteo Re; 2) Marco Frasca
- Supervisore della laureanda Jessica Gliozzo vincitrice di una borsa all’estero presso il Dept. of Computer Science della Royal Holloway - University of London, ai fini della predisposizione della tesi di laurea magistrale a.a. 2015/2016, finanziata dall' Università degli Studi di Milano.
- Relatore o correlatore di piu' di 40 tesi di laurea (triennale o magistrale o del vecchio ordinamento), in ambito Machine Learning e Bioinformatica, per le lauree in Informatica, Tecnologie dell’Informazione e della Comunicazione, Biotecnologie Molecolari e Bioinformatica, Matematica, dal 2005 ad oggi.
- Ho svolto e svolgo attività di tutor per decine di studenti nell'ambito delle laurea triennali e magistrali in Informatica.

Attività istituzionali

- Docente referente per la Laurea Magistrale in Informatica, Università degli Studi di Milano
  (da settembre 2019 ad oggi)
- Referente UNIMI per l'Area 06 – Scienze Informatiche per l'organizzazione e il censimento, in collaborazione con UniMITT, delle attività di ricerca dell'area informatica con ricadute in campo bio-medico e bio-tecnologico.
  (dal 2005 al 2007).
- Membro associato del CNR - Istituto di Elettronica e di Ingegneria dell’Informazione e delle Telecomunicazioni per il progetto “Machine Learning Techniques for Modeling and Growing Up” (2007-2008)
- Membro del collegio di dottorato in Informatica, Università degli Studi di Milano dal 2008
- Membro della commissione di valutazione per la Facolta' di Scienze MFN dell'Universita' degli Studi di Milano per il corso ERASMUS INTENSIVE PROGRAMME (IP) a.a. 2009/2010 "Interdisciplinary approaches to microarray data analysis" organizzato dalla Università di Helsinki, Warwick, Napoli (Federico II) e Milano.
- Membro della commissione trasferimenti e responsabile dei trasferimenti al corso di laurea in Informatica F1X dell'Università degli Studi di Milano da altri corsi di laurea dall'a.a. 2010/11 al 2014/15.
- Membro della comissione del dipartimento di Informatica per i rapporti di scambio didattico e scientifico con l'Università di Pechino e per il trasferimento tecnologico con la società di genomica cinese Rose Genomics (Beijing) nell'ambito di un costituendo accordo tra Fondazione UNIMI, UNIMI e Rose Genomics (dal 2018)
- Membro Data Mining and Big Data Analytics Technique Committee (DMTC) – IEEE Computational Intelligence Society (IEEE-CIS)
  (dal 2015)
- Membro del comitato scientifico del working group on Machine Learning and Data Mining, all'interno di AI*IA, Italian Association for Artificial Intelligence
  (dal 2015)

- Membro della commissione IEEE Computational Intelligence Society 2017 PhD Thesis Award (dal 2018)

Pubblicazioni

Articoli in riviste internazionali

R52. M. Frasca, G. Grossi, J. Gliozzo, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini and G. Valentini A GPU-based algorithm for fast node label learning in large and unbalanced biomolecular networks, *BMC Bioinformatics 19:Suppl 10 Oct. 15, 2018* *doi.org/10.1186/s12859-018-2301-4*

R51. S. Vascon, M. Frasca, R. Tripodi, G. Valentini, M. Pelillo Protein Function Prediction as a Graph-Transduction Game, *Pattern Recognition Letters (in press), 2018* doi.org/10.1016/j.patrec.2018.04.002

R50. M. Notaro, M. Schubach, P.N. Robinson, G. Valentini Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods, *BMC Bioinformatics, vol. 18 (1), 2017* doi.org/10.1186/s12859-017-1854-y

R49. M. Schubach, M. Re, P.N. Robinson and G. Valentini Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants, *Scientific Reports, Nature Publishing*, 7:2959, 2017. doi.org/10.1038/s41598-017-03011-5

R48. D. Smedley, M, Schubach, J. Jacobsen, S. Kohler, T. Zemojtel, M. Spielmann, M. Jager, H. Hochheiser, N. Washington, J. McMurry, M. Haendel, C. Mungall, S. Lewis, T. Groza, G. Valentini and P.N. Robinson A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease, *The American Journal of Human Genetics*, 99:3, pp.595--606, September 2016. doi.org/10.1016/j.ajhg.2016.07.005

R47. Y. Jiang, P. Oron, ... G. Valentini, ... I. Friedberg and P. Radivojac An expanded evaluation of protein function prediction methods shows an improvement in accuracy, *Genome Biology*, 17:184 September 2016. doi.org/10.1186/s13059-016-1037-6

R46. G. Valentini, G. Armano, M. Frasca, J. Lin, M. Mesiti and M. Re RANKS: a flexible tool for node label ranking and classification in biological networks, *Bioinformatics*, 32(18), September 2016. doi:10.1093/bioinformatics/btw235

R45. M. Frasca, S.Bassis, G. Valentini Learning node labels with multi-category Hopfield networks, *Neural Computing and Applications*, 27(6), pp 1677-1692, 2016 doi:10.1007/s00521-015-1965-1

R44. M. Frasca, G. Valentini COSNet: an R package for label prediction in unbalanced biological networks, *Neurocompting*, 2016. doi:10.1016/j.neucom.2015.11.096

R43. M. Frasca, A. Bertoni, G. Valentini UNIPred: Unbalance-aware Network Integration and Prediction of protein functions, *Journal of Computational Biology*, 22(12): 1057-1074, 2015. doi:10.1089/cmb.2014.0110

R42. M. Mesiti, M. Re, G. Valentini Think globally and solve locally: secondary memory-based network learning for automated multi-species function prediction, *GigaScience*, 3:5, 2014

R41. G. Valentini, A. Paccanaro, H. Caniza, A. Romero, M. Re, An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, *Artificial Intelligence in Medicine*, Volume 61, Issue 2, pages 63-78, June 2014

R40. H. Caniza, A. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, M. Mesiti, G. Valentini, A. Paccanaro, GOssTo: a user-friendly stand-alone and web tool for calculating semantic similarities on the Gene Ontology, *Bioinformatics*, Vol. 30 no. 15, pages 2235-2236, 2014

R39. G. Valentini, Hierarchical Ensemble Methods for Protein Function Prediction, *ISRN Bioinformatics*, vol. 2014, Article ID 901419, 34 pages, 2014

R38. M. Re, and G. Valentini, Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories,

IEEE ACM Transactions on Computational Biology and Bioinformatics 10(6), pp. 1359-1371, Nov-Dec 2013 IEEE link Supplemental Material

R37. I. Cattinelli, G. Valentini, E. Paulesu, A. Borghese A Novel Approach to the Problem of Non-uniqueness of the Solution in Hierarchical Clustering ,
IEEE Transactions on Neural Networks and Learning Systems 24(7) pp.1166-1173, July 2013

R36. M. Frasca, A. Bertoni, M. Re, and G. Valentini, A neural network algorithm for semi-supervised node label learning from unbalanced data,
Neural Networks 43, pp.84-98, July 2013

R35. M. Re, M. Mesiti and G. Valentini, A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks,
IEEE ACM Transactions on Computational Biology and Bioinformatics 9(6) pp. 1812-1818, 2012.

R34. A. Beghini, F. Corlazzoli, L. Del Giacco, M. Re, F. Lazzaroni, M. Brioschi, G. Valentini, F. Ferrazzi, A. Ghilardi, M. Righi, M. Turrini, M. Mignardi, C. Cesana, V. Bronte, M. Nilsson, E. Morra and R. Cairoli, Regeneration-associated Wnt signaling is activated in long-term reconstituting AC133bright acute myeloid leukemia cells,
Neoplasia 14:12, pp. 1236-1248, 2012

R33. M. Re and G. Valentini Cancer module genes ranking using kernelized score functions
BMC Bioinformatics 13 (Suppl 14): S3, 2012.

R32. N. Cesa-Bianchi, M. Re, G. Valentini, Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference,
Machine Learning, vol.88(1), pp. 209-241, 2012.

R31. M. Re, M. Mesiti, G. Valentini, Drug repositioning through pharmacological spaces integration based on networks projection,
EMBnet.journal, vol 18, Supplement A, pp.30-31, 2012.

R30. M. Frasca, A. Bertoni, G. Valentini, Regularized Network-Based Algorithm for Predicting Gene Functions with High-Imbalanced Data,
EMBnet.journal, vol 18, Supplement A, pp.41-42, 2012.

R29. G. Valentini, True Path Rule hierarchical ensembles for genome-wide gene function prediction,
IEEE ACM Transactions on Computational Biology and Bioinformatics, vol.8 n.3 pp. 832-847, 2011.

R28. 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, vol.8 n.5 pp. 1385-1392, 2011.

R27. M. Re, G. Valentini, Noise tolerance of Multiple Classifier Systems in data integration-based gene function prediction,
Journal of Integrative Bioinformatics, 7(3):139, 2010

R26. 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, vol.8: Machine Learning in Systems Biology, pp. 98-111, 2010.

R25. N. Cesa-Bianchi, G. Valentini, Hierarchical cost-sensitive algorithms for genome-wide gene function prediction,
Journal of Machine Learning Research, W&C Proceedings, vol.8: Machine Learning in Systems Biology, pp.14-29, 2010.

R24. M. Re, G. Valentini, Integration of heterogeneous data sources for gene function prediction using Decision Templates and ensembles of learning machines,
Neurocomputing, 73:7-9 pp. 1533-37, 2010 doi:10.1016/j.neucom.2009.12.012

R23. 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

R22. 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

R21. G.Valentini, R.Tagliaferri, F.Masulli, Computational Intelligence and Machine Learning in Bioinformatics
Artificial Intelligence in Medicine 45(2), pp. 91-96, 2009

R20. 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

R19. G.Pavesi, G.Valentini, Classification of co-expressed genes from DNA regulatory regions,
Information Fusion 10(3), pp. 233-241, 2009

R18. A. Bertoni, G.Valentini, Discovering multi-level structures in bio-molecular data through the Bernstein inequality
BMC Bioinformatics **9**(Suppl 2)**:**S4, 2008

R17. G.Valentini, N. Cesa-Bianchi, HCGene: a software tool to support the hierarchical classification of genes,
Bioinformatics, 24(5), pp. 729-731, 2008.

R16. F. Ruffino, M. Muselli, G.Valentini, Gene expression modelling through positive Boolean functions,
International Journal of Approximate Reasoning, 47(1), pp. 97-108, 2008.

R15. A.Bertoni, G.Valentini, Model order selection for biomolecular data clustering,
BMC Bioinformatics, vol.8, Suppl.3, 2007.

R14. G.Valentini, Mosclust: a software library for discovering significant structures in bio-molecular data.
Bioinformatics 23(3):387-389, 2007.

R13. G. Valentini, F.Ruffino, Characterization of Lung tumor subtypes through gene expression cluster validity assessment,
RAIRO - Theoretical Informatics and Applications, 40:163-176, 2006.

R12. 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.*

R11. G.Valentini, Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data,
Bioinformatics 22(3):369-370, 2006.

R10. 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

R9. P. Campadelli, E. Casiraghi, G.Valentini, Support Vector Machines for candidate nodules classification, *Neurocomputing* vol.68 pp. 281-289, 2005 Science Direct access

R8. A. Bertoni, R. Folgieri, G. Valentini, Bio-molecular cancer prediction with random subspace ensembles of Support Vector Machines,
Neurocomputing vol. 63C pp. 535-539, 2005 Science Direct access

R7. 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, 2004, MIT Press, JMLR link

R6. F. Masulli, G. Valentini, An experimental analysis of the dependence among codeword bit errors in ECOC learning machines.
Neurocomputing 57 pp. 189-214, 2004, science direct link

R5. G. Valentini, M. Muselli and F. Ruffino, Cancer recognition with bagged ensembles of Support Vector Machines,
Neurocomputing 56 pp. 461-466, 2004.

R4. F. Masulli, G. Valentini, Effectiveness of output coding decomposition schemes in ensemble and monolithic learning machines.
Pattern Analysis and Applications 6 pp. 285-300, 2003.

R3. 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

R2. G. Valentini, F. Masulli, NEURObjects: an object-oriented library for neural network development, *Neurocomputing* 48(1-4) pp. 623-646 , 2002.

R1. M. Pardo, G. Sberveglieri, A.Taroni, F. Masulli, G. Valentini Decompositive classification models for electronic noses.
Anal. Chim. Acta (446) pp. 223-232, 2001.

Editor di libri

E5. O. Okun, G. Valentini, M. Re (eds.), Ensembles in Machine Learning Applications,
Studies in Computational Intelligence, vol. 373 Springer, ISBN: 978-3-642-22909-1, 2011.

E4. O. Okun, M. Re, G. Valentini (eds.), Proceedings of the the Third Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications (SUEMA), European Conference on Machine Learning, Barcelona, Spain, 2010.

E3. O. Okun, G. Valentini (eds.), Applications of Supervised and Unsupervised Ensemble Methods,
Studies in Computational Intelligence, vol. 245 Springer, ISBN: 978-3-642-03998-0, 2010.

E2. 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.

E1. O. Okun, G. Valentini (eds.), Supervised and Unsupervised Ensemble Methods and their Applications, *Studies in ComputationalIntelligence*, vol. 126 Springer, ISBN: 978-3-540-78980-2, 2008.

Proceedings di conferenze internazionali e capitoli di libri

C84. A. Cuzzocrea, L. Cappelletti, G. Valentini A neural model for the prediction of pathogenic genomic variants in Mendelian diseases, 1st International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI), Barcelona, 2019

C83. M. Notaro, M. Schubach, M.Frasca, M. Mesiti, P.N. Robinson, G. Valentini Ensembling Descendant Term Classifiers to Improve Gene - Abnormal Phenotype Predictions, *Lecture Notes in Bioinformatics*, vol. 10834, pp. 70-80, 2019

C82. Marco Frasca, Jean Fred Fontaine, Giorgio Valentini, Marco Mesiti, Marco Notaro, Dario Malchiodi et al. Disease-Genes Must Guide Data Source Integration in the Gene Prioritization Process *Lecture Notes in Bioinformatics*, vol. 10834, pp. 60-69, 2019

C81. C. T. Ba, E. Casiraghi, M. Frasca, J. Gliozzo, G. Grossi, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, M. Re and G. Valentini, A Graphical Tool for the Exploration and Visual Analysis of Biomolecular Networks, *CIBB 2018 - Computational Intelligence methods for Bioinformatics and Biostatistics*, Lisboa (Portogallo) (accepted)

C80. C. Cano, M. Re, M. Verbeni, M. Notaro, G. Valentini, A. Blanco, Characterization of coding and non-coding RNA interactions through topological descriptors, *CIBB 2018 - Computational Intelligence methods for Bioinformatics and Biostatistics*, Lisboa (Portogallo) (accepted)

C79. M. Notaro, M. Schubach, P.N. Robinson, G. Valentini Predicting new relationships between genes and Human Phenotype Ontology terms, presented at the *26th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Chicago, 2018.

C78. A. Petrini, M. Schubach, M. Re, M. Frasca, M. Mesiti, G. Grossi, T. Castrignano', P.N. Robinson, G. Valentini Parameters tuning boosts hyperSMURF predictions of rare deleterious non-coding genetic variants,
PeerJ Preprints 5:e3185v1, 2017 presented at Methods, tools & platforms for Personalized Medicine in the Big Data Era - NETTAB 2017, Palermo, Italy

C77. M. Schubach, M. Re, P.N. Robinson, G. Valentini Variant relevance prediction in extremely imbalanced training sets,

F1000Research 2017, 6(ISCB Comm J):1392 (poster) (doi: 10.7490/f1000research.1114637.1), presented at the 25th International Conference on Intelligent Systems for Molecular Biology (ISMB), Prague 2017

C76. M. Notaro, M. Schubach, P.N. Robinson, G. Valentini Ensembling Descendant Term Classifiers to Improve Gene - Abnormal Phenotype Predictions,
CIBB 2017, The 14th International Conference on Bioinformatics and Biostatistics, Cagliari, Italy, 2017.

C75. M. Frasca, J.F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi and M.A. Andrade-Navarro Disease Genes must Guide Data Source Integration in the Gene Prioritization Process ,
CIBB 2017, The 14th International Conference on Bioinformatics and Biostatistics, Cagliari, Italy, 2017.

C74. J. Lin, M. Mesiti, M. Re and G. Valentini Within network learning on big graphs using secondary memory-based random walk kernels,
Complex Networks & Their Applications V: Proceedings of the 5th International Workshop on Complex Networks and their Applications (COMPLEX NETWORKS 2016), Studies in Computational Intelligence, Springer, pp. 235-245, 2017, doi.org/10.1007/978-3-319-50901-3_19

C73. P. Perlasca, G. Valentini, M. Frasca, M. Mesiti Multi-species Protein Function Prediction: Towards Web-based Visual Analytics,
Proceedings of the 18th International Conference on Information Integration and Web-based Applications & Services, Singapore, ACM, New York, USA pp. 1-5, 2016.
doi.org/10.1145/3011141.3011222

C72. H. Su, G. Valentini, S. Szedmak and J. Rousu Transport Protein Classification through Structured Prediction and Multiple Kernel Learning ,
NIPS Workshop on Machine Learning in Computational Biology (MLCB) & Machine Learning in Systems Biology (MLSB) 2015 - Montreal, Canada, December 2015

C71. P.N. Robinson, M.Frasca, S. Kohler, M. Notaro, M. Re, G. Valentini, A hierarchical ensemble method for DAG-structured taxonomies ,
Multiple Classifier Systems - MCS 2015 - Gunzburg, Germany *Lecture Notes in Computer Science*, vol. 9132, pp. 15-36, Springer, 2015

C70. G. Valentini, S. Kohler, M. Re, M. Notaro, P.N. Robinson, Prediction of human gene - phenotype associations by exploiting the hierarchical structure of the Human Phenotype Ontology,
3rd International Work-Conference on Bioinformatics and Biomedical Engineering - IWBBIO 2015, Granada, Spain *Lecture Notes in Bioinformatics*, vol. 9043, pp. 66-77, Springer, 2015

C69. M. Re, M.Mesiti, G. Valentini, An automated pipeline for multi-species protein function prediction from the UniProt Knowledgebase,
Automated Function Prediction SIG 2014 - ISMB 2014, Boston, USA

C68. M. Re, M.Mesiti, G. Valentini, On the Automated Function Prediction of Big Multi-Species Networks,
Network Biology SIG 2014 - ISMB 2014, Boston, USA

C67. M.Frasca, A. Bertoni, G. Valentini An unbalance-aware network integration method for gene function prediction,
MLSB 2013 - Machine Learning for Systems Biology, Berlin, 2013

C66. G. Valentini, A. Paccanaro, H. C. Vierci, A. E. Romero, M. Re, Network integration boosts disease gene prioritization,
Network Biology SIG 2013 - ISMB 2013, Berlin

C65. M.Mesiti, M. Re, G. Valentini Scalable Network-based Learning Methods for Automated Function Prediction based on the Neo4j Graph-database,
Automated Function Prediction SIG 2013 - ISMB 2013, Berlin

C64. H. C. Vierci, A. E. Romero, S. Heron, H. Yang, M. Frasca, M. Mesiti, G. Valentini and A. Paccanaro GOssTo & GOssToWeb: user-friendly tools for calculating semantic similarities on the Gene Ontology,
Bio-Ontologies SIG 2013 - ISMB 2013, Berlin

C63. M. Re, M.Mesiti, G. Valentini Comparison of early and late omics data integration for cancer modules gene ranking ,
NETTAB 2012 Workshop on Integrated Bio-Search, Como 14-16 November, 2012.

C62. M. Re and G. Valentini Random walking on functional interaction networks to rank genes involved in cancer
2nd Artificial Intelligence Applications in Biomedicine Workshop, in: AIAI 2012 - Artificial Intelligence Applications and Innovations, pp. 66-75, *IFIP AICT Series*, Springer, 2012

C61. M. Re, G. Valentini Large Scale Ranking and Repositioning of Drugs with Respect to DrugBank Therapeutic Categories, slides
In: L. Bleris et al. (Eds.): International Symposium on Bioinformatics Research and Applications (ISBRA 2012), Dallas, USA, *Lecture Notes in Bioinformatics* vol.7292, pp. 225-236, Springer, 2012.

C60. M. Re, G. Valentini, Ensemble methods: a review,
In: Advances in Machine Learning and Data Mining for Astronomy, Chapman & Hall Data Mining and Knowledge Discovery Series, Chap. 26, pp. 563-594, 2012.

C59. M. Re, G. Valentini Genes prioritization with respect to Cancer Gene Modules using functional interaction network data , *NETTAB 2011* Workshop on Clinical Bioinformatics, Pavia 12-14 October, 2011.

C58. A. Bertoni, M. Frasca, G. Valentini COSNet: a Cost Sensitive Neural Network for Semi-supervised Learning in Graphs.,
In: "Machine Learning and Knowledge Discovery in Databases". European Conference, ECML PKDD 2011, Athens, Greece, Proceedings, Part I, *Lecture Notes in Artificial Intelligence*, vol. 6911, pp.219-234, Springer, 2011.

C57. A. Rozza, G. Lombardi, M. Re, E. Casiraghi, G. Valentini and P. Campadelli A Novel Ensemble Technique for Protein Subcellular Location Prediction ,
In: "Ensembles in Machine Learning Applications", *Studies in Computational Intelligence* vol. 373, pp. 151-167, Springer, 2011

C56. M. Frasca, A. Bertoni, G. Valentini A cost-sensitive neural algorithm to predict gene functions using large biological networks.,
Network Biology SIG: On the Analysis and Visualization of Networks in Biology, ISMB 2011, Wien

C55. A. Bertoni, M. Re, F. Sacca, G. Valentini Identification of promoter regions in genomic sequences by 1-dimensional constraint clustering,
Frontiers in Artificial Intelligence and Applications, vol. 234, Neural Nets WIRN11 - Proceedings, pp. 162-169, 2011.

C54. A. Rozza, G. Lombardi, M. Re, E. Casiraghi, and G. Valentini, DDAG K-TIPCAC: an ensemble method for protein subcellular localization,
Proc. of the Third Edition of SUEMA, pp. 75-84 , ECML, Barcelona, Spain, 2010.

C53. N. Cesa-Bianchi, M. Re, G. Valentini, Functional Inference in FunCat through the Combination of Hierarchical Ensembles with Data Fusion Methods,
ICML Workshop on learning from Multi-Label Data MLD'10 , Haifa, Israel, pp.13-20, 2010

C52. A. Bertoni, M. Frasca, G. Grossi, G. Valentini, Learning functional linkage networks with a cost-sensitive approach ,
Neural Networks - WIRN 2010, IOS Press, pp. 52-61, 2010

C51. M. Re, G. Valentini, An experimental comparison of Hierarchical Bayes and True Path Rule ensembles for protein function prediction,
In: (N. El Gayar, J. Kittler and F. Roli, Eds) Nineth International Workshop on Multiple Classifier Systems MCS 2010, *Lecture Notes in Computer Science*, vol. 5997, pp. 294-303, Springer, 2010.

C50. 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.

C49 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.

C48 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.

C47. M. Re, G. Valentini, Predicting gene expression from heterogeneous data,
CIBB 2009, The Sixth International Conference on Bioinformatics and Biostatistics, Genova, Italy, 2009.

C46. 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.

C45. 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.

C44. 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.

C43. 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, pp. 249-278, 2009.

C42. 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.

C41. 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*, vol.245, pp. 79-91, Springer, 2010.

C40. 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.

C39. 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,
in: I. Lovrek, R. J. Howlett, L. C. Jain (Eds.): Knowledge-Based Intelligent Information and Engineering Systems, 12th International Conference, KES 2008, Zagreb, Croatia, September 3-5, 2008, Proceedings, Part III. *Lecture Notes in Computer Science*, vol.5179 pp. 764-770, Springer 2008.

C38. 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.

C37. O. Okun, G.Valentini, Dataset Complexity Can Help to Generate Accurate Ensembles of K-Nearest Neighbors,
IEEE International Joint Conference on Neural Networks - IJCNN 2008 (IEEE World Congress on Computational Intelligence), pp. 450-457, 2008.

C36. 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.

C35. 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.

C34. 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.

C33. 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

C32. 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.

C31. 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.

C30. F. Ruffino, M. Muselli, G. Valentini, Modeling 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.

C29. A.Bertoni, G. Valentini, Model order selection for clustered bio-molecular data,
In: *Probabilistic Modeling 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, slides

C28. 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.

C27. 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.

C26. 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. 3849, pp. 277-283, 2006.

C25. P. Campadelli, E. Casiraghi, G.Valentini, Lung nodules detection and classification,
ICIP 05, The IEEE International Conference on Image Processing, Genova, Italy, 2005.

C24. 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.

C23. 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.

C22. 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.

C21. 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, Powerpoint slides

C20. 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.

C19. 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.

C18. 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.

C17. 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.

C16. 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.

C15. 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.

C14. G. Valentini, Supervised gene expression data analysis using Support Vector Machines and Multi-Layer Perceptrons,
In: Knowledge-Based Intelligent Information Engineering Systems and Allied technologies - Sixth International Conference on Knowledge-Based Intelligent Information & Engineering Systems *KES'2002* , special session Machine Learning in Bioinformatics, pp. 482-487, 2002.

C13. 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.

C12. 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), Copenaghen, Denmark, 2001.

C11. 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.

C10. 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.

C9. 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

C8. 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.

C7. 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

C6. 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, 2001

C5. 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.

C4. 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.

C3. 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.

C2. 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.

C1. G. Valentini, F. Masulli, NEURObjects, 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 di Conferenza Nazionali

N21. J. Gliozzo, M. Notaro, A. Petrini, P. Perlasca, M. Mesiti, E. Casiraghi, M.Frasca, G. Grossi, M. Re, A. Paccanaro, G. Valentini Modeling biomolecular profiles in a graph-structured sample space for clinical outcome prediction with melanoma and ovarian cancer patients ,
BITS 2017, Bioinformatics Italian Society Meeting, Cagliari, Italy, 2017.

N20. A. Petrini, M. Notaro, J. Gliozzo, G. Valentini, G. Grossi, M. Frasca Speeding up node label learning in unbalanced biomolecular networks through a parallel and sparse GPU- based Hopfield model
BITS 2017, Bioinformatics Italian Society Meeting, Cagliari, Italy, 2017.

N19. P. Perlasca, M. Mesiti, M. Notaro, A. Petrini, J. Gliozzo, G. Valentini, M. Frasca A Web Graphical Tool for the Integration of Unbalanced Biomolecular Networks ,
BITS 2017, Bioinformatics Italian Society Meeting, Cagliari, Italy, 2017.

N18. M. Re, M. Mesiti, M. Frasca, J. Lin, G. Valentini Analysis of bio-molecular networks through semi-supervised graph-based learning methods ,
*Third Italian Workshop on Machine Learning and Data Mining - XIII AI*IA Symposium on Artificial Intelligence* (invited talk), Pisa December 2014.

N17. M. Dugo, M. Callari, P. Miodini, V. Cappelletti, M.L. Carcangiu, R. Orlandi, G. Valentini, MG Daidone, Performance of single sample predictors in defining breast cancer molecular subtypes ,
53rd Annual Meeting of the Italian Cancer Society , Torino, October 2011.

N16. A. Bertoni, M. Frasca, G.Valentini, An efficient supervised method to integrate multiple biological networks ,
BITS 2011, Bioinformatics Italian Society Meeting, Pisa, Italy, 2011.

N15. A. Rozza , G. Lombardi, M. Re, E. Casiraghi, G. Valentini, P. Campadelli, A Novel Ensemble Approach for the Subcellular Localization of Proteins ,
BITS 2011, Bioinformatics Italian Society Meeting, Pisa, Italy, 2011.

N14. D. Malchiodi, M. Re and G. Valentini, Uso di Mathematica per la classificazione di dati di qualitÃ variabile ,
Mathematica Italia User Group Meeting - Atti del Convegno 2010, Adalta (ISBN 978-88-96810-00-2), 2010.

N13. M. Re, G.Valentini, Data fusion based gene function prediction using ensemble methods,
BITS 2009, Bioinformatics Italian Society Meeting, Genova, Italy, 2009.

N12. N. Cesa-Bianchi, G. Valentini, Genome-Âwide hierarchical classification of gene function,
BITS 2009, Bioinformatics Italian Society Meeting, Genova, Italy, 2009.

N11. 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.

N10. 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.

N9. 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.

N8. G.Pavesi, G.Valentini, Classification of co-expressed genes from DNA regulatory regions
BITS 2007, Bioinformatics Italian Society Meeting, Napoli, Italy, 2007.

N7. G. Pavesi , G. Valentini, G. Mauri, G. Pesole, Motif Based Classification of Coregulated Genes,
BITS 2006, Bioinformatics Italian Society Meeting, Bologna Italy, 2006.

N6. A. 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

N5. 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

N4. M. Muselli, F. Ruffino, and G. Valentini, An Artificial Model for Validating Gene Selection Methods,
BITS 2004, Bioinformatics Italian Society Meeeting, Padova, Italy, 2004

N3. 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.

N2. 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.

N1. 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.