UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI INFORMATICA

Dottorato di ricerca in Informatica e Matematica
XXX ciclo

Progetto di ricerca

Dottorando: Dott. Roberto Corizzo
Tutor: Prof. Michelangelo Ceci

Coordinatore
Prof. Donato Malerba

Firma del dottorando _____

Firma del tutor _____

1) Titolo della ricerca:

Predictive Models for Streams of Sensor Data

2) Area nella quale si inquadra la ricerca:

Apprendimento automatico e Data Mining

3) Obiettivi della ricerca:

Il presente progetto di ricerca ha come obiettivo quello di sviluppare metodi di data mining per la predizione dei valori di attributi dipendenti in una rete di sensori georeferenziati.

L'applicazione di algoritmi di data mining spazio temporale consentirà di costruire dei modelli a partire dai dati storici disponibili e di predire i valori assunti delle variabili di interesse per un determinato orizzonte temporale (da pochi minuti a pochi giorni). Sarà inoltre possibile sollevare eventi negativi qualora i valori non dovessero essere coerenti con i modelli costruiti in precedenza.

4) Motivazioni della ricerca

Nonostante la presenza di approcci già noti in letteratura per lo specifico obiettivo di ricerca, bisogna considerare alcune sfide che devono essere affrontate:

- La presenza di **autocorrelazione spaziale** introdotta dalla prossimità spaziale dei sensori viola l'assunzione che le istanze siano indipendenti e identicamente distribuite, che sottostà ai principali algoritmi di Data Mining. Risulta opportuno integrare e sperimentare tecniche statistiche per l'autocorrelazione spaziale note in letteratura (quali ad esempio la tecnica PCNM o la tecnica LISA);

- Occorre prendere in considerazione anche l'**autocorrelazione temporale**, indotta dall'evidente ciclicità dei giorni dell'anno e delle stagioni. A questo scopo si rivelano utili ad esempio le **directional statistics**, che consentirebbero di considerare la collocazione temporale della giornata da predire, pesando anche opportunamente il contributo delle altre giornate (dati storici) sulle quali è stato costruito il modello;

- Quando i processi non sono strettamente stazionari, il concetto target potrebbe cambiare gradualmente nel tempo (es: attacchi di rete, frodi in transazioni con carte di credito, etc). Questa proprietà è nota come **concept drift**, ovvero il cambio delle proprie caratteristiche nel tempo. Per gestire questi aspetti occorre applicare algoritmi di predizione adattivi e di serie temporali, che di conseguenza richiedono una fase di training continua;

- Occorre considerare la probabilità di **rumore nei dati**: infatti i dati che sono acquisiti dai sensori potrebbero non essere trasmessi a causa di problemi tecnici. Gli approcci di **semi-supervised learning** rispondono all'esigenza di lavorare con dati mancanti o non etichettati;

- Creare un modello di predizione generale, in grado di gestire grandi moli di dati diversificati e che preveda metodi che lavorino a differenti livelli di granularità spaziale e temporale. Sarà pertanto opportuno integrare tecniche di predizione short-term e long-term, e che in ogni caso prendano in considerazione l'autocorrelazione spaziale.

I sistemi esistenti si pongono e affrontano solo alcune di queste sfide. L'obiettivo del progetto di ricerca è quello di affrontare in maniera sistematica e combinata le problematiche elencate.

5) Stato dell'arte

Un data stream è una sequenza ordinata di istanze che può essere letta solo una volta o un numero limitato di volte utilizzando limitate risorse di computazione e memorizzazione. Queste fonti di dati sono senza limiti, scorrono ad alta velocità e sono generate da distribuzioni non stazionarie in ambienti dinamici [2].

Nel modello data stream, gli elementi di input *a1, a2, ..., aj, ...* arrivano sequenzialmente, item per item e descrivono una funzione sottostante *A*. Modelli di stream differiscono su come ai descrive *A*. Si distinguono tra:

- **Insert only Model**: quando un elemento a_i è visto, non può essere cambiato;
- **Insert-Delete Model**: gli elementi a_i possono essere eliminati o aggiornati;
- **Accumulative Model**: ciascun a_i è un incremento ad $A[j] = A[j-1] + a_i$

Gli operatori di query definiti "blocking" sono operatori non in grado di produrre la prima tupla dell'output prima di aver visto l'intero input. Ad esempio sono SORT, SUM, COUNT, MAX, MIN. Nella situazione degli stream, query continue che utilizzano questi operatori sono problematiche.

Si effettuano quindi:

- Tecniche di **query processing approssimate** per valutare query che richiedono una quantità non limitata di memoria;
- Query processing **sliding window,** come tecnica di approssimazione e opzione nel linguaggio di query;
- **Campionamento** per gestire situazioni in cui il tasso di flusso dello stream di input è più veloce del processore di query;
- Il significato e l'**implementazione di operatori blocking** (come aggregazione e ordinamento) in presenza di flussi senza fine.

Processamento dati	Dati tradizionali	Data streams
Numero di passi	*multipli*	*singolo*
Tempo di processamento	*illimitato*	*limitato*
Utilizzo di memoria	*illimitato*	*limitato*
Tipo di risultato	*accurato*	*approssimato*
Distribuito	*no*	*si*

Uno stimatore è una funzione del campione di dati osservabile che è usato per stimare un parametro della popolazione sconosciuta.

Si possono ridefinire operatori come media, deviazione standard e coefficiente di correlazione in forma ricorsiva, da poter calcolare sui campioni dei flussi di dati senza dover memorizzare o avere a disposizione tutti i valori.

Tuttavia, queste statistiche sono di uso limitato nei problemi tipici dei data streams: nella maggior parte delle applicazioni i dati recenti sono i più rilevanti: per soddisfare questo scopo, un approccio popolare consiste nel definire una finestra temporale (nota come Sliding Window) che copre i dati più recenti.

Le **sliding window** sono un approccio usato comunemente per risolvere query nel contesto di flussi di dati non limitati (open-ended). Anzichè computare una risposta basata sull'intero flusso di dati, la query (o l'operatore) è computato, eventualmente più volte, su un sottoinsieme limitato di tuple. In questo modello, un timestamp è associato a ciascuna tupla. Il timestamp definisce quando una specifica tupla è valida (ad esempio dentro la finestra) o meno. Le query sono eseguite sulle tuple nella finestra. Tuttavia, nel caso si mettano insieme più flussi la semantica dei timestamp è meno chiara (ad esempio il timestamp di una tupla di output). Sono stati utilizzati più modelli a finestra in letteratura, ma i più rilevanti sono i seguenti:

- **Landmark windows**

  I punti rilevanti identificati (i landmark) nei data stream e gli operatori aggregati usano tutti i record visti fino a quel momento dopo il landmark. Le finestre successive condividono alcuni punti iniziali e sono di **dimensione crescente**. In alcune applicazioni, i landmark hanno una semantica naturale. Ad esempio, su base giornaliera aggregare l'inizio del giorno è un landmark.

- **Sliding windows**

  Il più delle volte, non siamo interessati a calcolare statistiche sull'intero passato, ma solo nel recente passato. L'approccio più semplice sono le sliding windows di **dimensione fissata**. Questo tipo di finestre sono simile alle strutture dati first in, first out. Ogni volta che un elemento ***j*** è osservato e inserito nella finestra, un altro elemento ***j-w***, dove ***w*** rappresenta la dimensione della finestra, è tenuto fuori.

- **Tilted windows**

  I modelli di finestra precedenti prevedono che qualunque osservazione passata ricada o non ricada nella finestra. Nelle finestre *tilted*, la scala temporale è compressa. I dati più recenti sono memorizzati nella finestra al livello di granularità più fine. Le informazioni più vecchie sono memorizzate a un livello più grossolano, in forma aggregata. Il livello di granularità dipende dall'applicazione. Nel caso delle finestre ***natural tilted time***, i dati sono memorizzati con granularità rispetto alla naturale tassonomia del tempo: l'ultima ora a una granularità di 15 minuti (4 punti), l'ultimo giorno a una granularità ad ore (24 punti), l'ultimo mese a una granularità a giorni (32 punti) e l'ultimo anno a una granularità in mesi (12 punti). Nel caso delle ***logarithmic tilted windows***, data una granularità massima con periodi di ***t***, la granularità decresce logaritmicamente quando i dati sono più vecchi. Allo scorrere del tempo, la finestra memorizza l'ultimo periodo ***t***, il periodo prima di questo e consecutivamente aggrega a granularità inferiore (2 periodi, 4 periodi, 8 periodi, etc).

Modelli di predizione

I modelli di predizione impiegabili spaziano tra reti neurali, clustering predittivo di flussi di dati (stream) e ensemble bayesiani. Tutti questi modelli lavorano su un flusso (uno stream) di dati collezionati ad intervalli di tempo regolari e di conseguenza aggiornano i modelli di predizione.

Modelli di predizione basati su reti neurali

Modelli matematici che rappresentano l'interconnessione tra elementi definiti neuroni artificiali, ossia costrutti matematici che in qualche misura imitano le proprietà dei neuroni viventi. Questi modelli matematici possono essere utilizzati per risolvere problemi ingegneristici di intelligenza artificiale come quelli che si pongono in diversi ambiti tecnologici (in elettronica, informatica, simulazione, e altre discipline).

I suddetti neuroni ricevono in ingresso degli stimoli e li elaborano. L'elaborazione può essere anche molto sofisticata ma in un caso semplice si può pensare che i singoli ingressi vengano moltiplicati per un opportuno valore detto peso, il risultato delle moltiplicazioni viene sommato e se la somma supera una certa soglia il neurone si attiva attivando la sua uscita. Il peso indica l'efficacia sinaptica della linea di ingresso e serve a quantificarne l'importanza, un ingresso molto importante avrà un peso elevato, mentre un ingresso poco utile all'elaborazione avrà un peso inferiore.

Le reti neurali per come sono costruite lavorano in parallelo e sono quindi in grado di trattare molti dati. Si tratta in sostanza di un sofisticato sistema di tipo statistico dotato di

una buona immunità al rumore. I singoli neuroni vengono collegati alla schiera di neuroni successivi, in modo da formare una rete di neuroni. Normalmente una rete è formata da tre strati. Nel primo abbiamo gli ingressi (I), questo strato si preoccupa di trattare gli ingressi in modo da adeguarli alle richieste dei neuroni. Se i segnali in ingresso sono già trattati può anche non esserci. Il secondo strato è quello nascosto (H, hidden), si preoccupa dell'elaborazione vera e propria e può essere composto anche da più colonne di neuroni. Il terzo strato è quello di uscita (O) e si preoccupa di raccogliere i risultati ed adattarli alle richieste del blocco successivo della rete neurale. Queste reti possono essere anche molto complesse e coinvolgere migliaia di neuroni e decine di migliaia di connessioni. Per costruire la struttura di una rete neurale multistrato si possono inserire N strati Hidden; vi sono però alcune dimostrazioni che mostrano che con 1 o 2 strati di Hidden si ottiene una stessa efficace generalizzazione da una rete rispetto a quella con più strati Hidden. L'algoritmo di backpropagation è utilizzato nell'apprendimento supervisionato. Esso permette di modificare i pesi delle connessioni in modo tale che si minimizzi una certa funzione errore E.

La funzione errore che si deve minimizzare si può scrivere come:

$$E(w) = \frac{1}{2} \sum_h \sum_k (out_k^h - y_k^h)^2$$

L'errore commesso dalla rete è propagato all'indietro (backward) e i pesi sono aggiornati in maniera appropriata.

Il metodo del gradiente discendente rappresenta la più utilizzata classe di algoritmi per l'apprendimento supervisionato di reti neurali. Il più popolare tra questi è il Back Propagation, un metodo del primo ordine che minimizza la funzione di errore aggiornando i pesi w utilizzando il metodo steepest descent.

Un esempio di impiego di tale modello su stream di dati e, specificatamente in ambito di energy prediction, è quello di Joao Gama e del suo gruppo di ricerca che applica tecniche di machine learning: gli autori affrontano questo problema in [1] modificando un algoritmo del gradiente discendente computazionalmente poco costoso (RPROP, Resilient Backpropagation), al fine di migliorare la velocità di apprendimento. Uno dei problemi inerenti i metodi del gradiente discendente riguarda la convergenza ai minimi locali. Queste tecniche utilizzano solo informazioni sul gradiente come la derivata parziale dell'errore rispetto ai pesi, per adattare parametri specifici del peso.

Gli autori hanno testato il metodo su dati relativi a due parchi eolici in combinazione con dati ambientali provenienti dal modello MM5, per la predizione di energia ad un orizzonte di 72 ore. La presenza di distribuzioni non gaussiane relativamente all'errore di predizione ha motivato la ricerca di tecniche di apprendimento che minimizzano il contenuto informativo della distribuzione dell'errore anzichè minimizzare la sua varianza (RMSE).

Una misura tipica di contenuto informativo è l'entropia e la particolarità del setting di apprendimento utilizzato dagli autori consiste nell'applicare metriche basate su entropia come criterio di ottimizzazione (responsabile dell'aggiornamento dei pesi della rete neurale), come ad esempio l'MCC (massima correntropia), l'MEE (minima entropia), il MEEF (minimo errore di entropia con punti di fiducia).

Si osserva che l'uso di entropia come criterio di performance porta a predizioni migliori (in termini di maggiore frequenza di errori vicini allo zero e insensibilità agli outlier) rispetto alla scelta dell'RMSE.

In seguito a una fase di training offline mediante utilizzo di dati storici, la rete neurale effettua predizioni online (incrementalmente) per i timestamp successivi al timestamp attuale..

Modelli di predizione basati su clustering

Il metodo proposto in [3] utilizza dati storici in forma di stream per la predizione a breve termine (da 10 minuti a 3 ore) e si basa sulla ricerca degli analoghi locali (o k vicini più vicini) sfruttando la collaborazione di un vicinato di 11 siti.

In una prima fase, le *k* situazioni più simili alla situazione corrente sono estratte a partire da un archivio di dati passati, in cui *k* è un parametro del metodo. I timestamp e i pesi di questi analoghi sono scambiati con gli altri sistemi e in una seconda fase vengono identificati gli analoghi globali, consentendo di escludere gli analoghi locali che non fanno parte del trend complessivo (intero vicinato).

L'approccio per descrivere le situazioni è detto *template matching*: è una tecnica per discretizzare una serie temporale in segmenti di lunghezza desiderata. Un template è un vettore binario di m elementi in cui un valore 1 indica che il valore misurato al timestamp corrispondente a quella posizione è di interesse. Il template è applicato esaustivamente a tutti i timestamp del passato e una volta coperto l'intero spazio di ricerca, viene applicata una misura di distanza per estrarre le k situazioni più simili a quella corrente.

$$w_i(p) = \frac{1}{k-1} * \frac{\left(\sum_{j \in c_i} dist(q_i, T_i(t_j))^2\right)^{-dist(q_i, T_i(t_p))^2}}{\sum_{j \in c_i} dist(q_i, T_i(t_j))^2}$$

In seguito, sia:

- **α** un parametro che indica il peso o l'importanza data all'informazione dei vicini;
- **ε** un parametro che indica che la differenza in tempo massima per due analoghi per essere considerati parte della stessa situazione globale

La funzione punteggio per gli analoghi globali è calcolata come:

$$score(p) = (1-\alpha) * w_i(p)$$
$$+\alpha * \frac{1}{|neighbors(i)|} \sum_{j \in neighbors(i)} \sum_{l=1}^{k} w_j(l) * ts(t_p, t_l)$$

$$ts(t_1, t_2) = \begin{cases} 1 & |t_1 - t_2| \leq \varepsilon \\ 0 & |t_1 - t_2| > \varepsilon \end{cases}$$

Dopo aver scelto i **b** analoghi globali con più alto punteggio, viene scelto il mediano per effettuare la predizione.

Il metodo risulta efficace per un orizzonte di predizione fino a 3 ore (si ottengono buoni risultati con un orizzonte di 10, 20, 30 minuti, 1, 2 e 3 ore). Ad eccezione dell'orizzonte di 10 minuti, il metodo degli analoghi ha performance migliori del metodo di persistenza. Il modello di regressione tuttavia presenta performance migliori del metodo degli analoghi, ma richiederebbe la centralizzazione di tutti i dati e il ricalcolo del modello all'arrivo di nuovi dati.

Modelli di predizione basati su Bayesian Ensemble

Consiste nell'impiego di più predittori i cui risultati si combinano per ottenere la predizione finale, utilizzando un approccio di mediatura. In [4] viene proposto un approccio Bayesian Ensemble con 3 predittori distinti: Naive Bayes, k-NN e mining di motivi.

Tecniche per l'autocorrelazione spaziale

Principal Coordinate analysis of Neighbour Matrices (PCNM) [5]

La tecnica propone di estrarre predittori spaziali facilmente incorporabili in modelli di regressione o analisi tradizionali. Il metodo consiste nella diagonalizzazione di una matrice spaziale di pesi (ottenuta a partire dalla matrice delle distanze geografiche tra i nodi) e in seguito dell'estrazione di autovettori che massimizzano l'indice di autocorrelazione di Moran.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

Local Indicators of Spatial Association (LISA) [6]

Con le misure globali un singolo valore si applica all'intero dataset. Pertanto lo stesso pattern o processo avviene nell'intera area geografica.

Con le misure locali è calcolato un valore per ciascuna unità di osservazione. Pattern differenti o processi differenti potrebbero avvenire in diverse parti di una regione. Il valore è calcolato per ogni località. La tecnica LISA è considerata come la versione locale della statistica I di Moran. Ad esempio: per ciascun poligono, l'indice è calcolato considerando come vicinato l'insieme di poligoni che condividono un lato con il poligono corrente.

$$I_i = z_i \sum_j w_{ij} z_j \qquad z_i = \frac{x_i - \bar{x}}{SD_x}$$

L'indice di Moran univariato è la correlazione tra una variabile X e la stessa variabile X in aree vicine.

L'indice di Moran bivariato è la correlazione tra una variabile X e una variabile Y distinta nelle aree vicine. E' possibile considerare la LISA bivariata come una versione locale del

coefficiente di correlazione. Essa mostra la natura e la forza dell'associazione tra due variabili nella regione di interesse.

6) Approccio al problema

1) Acquisizione e preprocessing di dati simulati e misurati (ad esempio provenienti da impianti di energia rinnovabile) e caratterizzati dalla presenza di autocorrelazione, (in grado di descrivere direttamente un sito della rete oppure tramite le features del vicinato di un impianto, che infine vengono aggregate al sito stesso).

2) L'applicazione di metodi di predizione quali ad esempio reti neurali, clustering predittivo ed ensemble learning sui dati a disposizione (ad esempio per ottenere delle predizioni dell'energia erogata dai siti ad un orizzonte temporale di 24 e 48 ore).

3) La valutazione sperimentale degli approcci proposti, in modo da individuare la combinazione più promettente tra dati di input, output del sistema e metodo utilizzato.

7) Ricadute applicative

Le reti di sensori ricoprono un ruolo fondamentale nel task di monitoraggio di qualsiasi ambiente. I dati collezionati dai sensori possono inoltre essere impiegati per predire lo stato delle variabili di interesse ad un determinato orizzonte temporale (da pochi minuti fino a pochi giorni) mediante tecniche di predizione spazio-temporale.

Esempi di applicazioni includono il monitoraggio di rete, la modellazione utente in applicazioni web, reti di sensori in reti elettriche o impianti di energia rinnovabile, gestione di dati di telecomunicazioni, predizione in borsa, monitoraggio di frequenze radio, etc.

L'importanza e l'utilità della predizione dipende dallo specifico contesto. Ad esempio, nell'ambito dell'energy prediction, si osserva che le fonti rinnovabili (quali array fotovoltaici o parchi eolici) sono caratterizzati da variabilità e intermittenza relativamente alla quantità di energia prodotta, rendendo difficile la loro integrazione nella griglia energetica. Inoltre, nel mercato energetico il contributo delle singole fonti contribuisce alla definizione del prezzo nel mercato orario o giornaliero: variazioni rispetto alle stime di energia prodotta influenzeranno il prezzo finale di acquisto. Una predizione quanto più precisa possibile comporterebbe minori perdite per i player del mercato energetico (distributori, compagnie, etc).

Conseguentemente, l'impiego di soluzioni di data mining che prendano in considerazione aspetti quali il concept drift e l'autocorrelazione spaziale e temporale, consentirebbero di raffinare le predizioni grezze messe a disposizione dalle autorità di energia elettrica nazionali, con un conseguente miglioramento nell'offerta per il mercato energetico e la definizione di una strategia di vendita e acquisto accurata.

8) Riferimenti bibliografici

[1] Entropy and Correntropy Against Minimum Square Error in Offline and Online Three-Day Ahead Wind Power Forecasting - R.Bessa, V.miranda, J.Gama - IEEE 2009

[2] A Framework for Clustering Evolving Data Streams - Charu C. Aggarwal, Jiawei Han, Jianyong Wang, Philip S. Yu - Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003

[3] Analog Method for Collaborative Very-Short-Term Forecasting of Power Generation from Photovoltaic Systems - Veronica Gomez Berdugo , Cristophe Chaussin, Laurent Dubus, Georges Hebrail, Vivianle Leboucher

[4] Fine-grained Photovoltaic Output Prediction using a Bayesian Ensemble - Prithwish Chakraborty , Manish Marwas, Martin Arlitt, Naren Ramakrishnan - AAAI 2012

[5] Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM) - Legendre et al.

[6] Local Indicators of Spatial Association-LISA Geographical Analysis - Luc Anselin - 1995

9) Fasi del progetto

Anno 1°: studio della letteratura, dello stato dell'arte e del materiale di ricerca di base:

- **Attività 1A:** studio approfondito di aspetti teorico-formali dell'apprendimento automatico e dello sviluppo di sistemi per la scoperta di conoscenza dai dati;
- **Attività 1B:** approfondimento delle tematiche relative alle tecniche di predizione di dati numerici (quali la regressione e le reti neurali);
- **Attività 1C:** ricerca e studio di metodi per la predizione applicati a dati numerici;
- **Attività 1D:** partecipazione a scuole internazionali inerenti all'attività e agli obiettivi previsti.

Anno 2°: sintesi, realizzazione e implementazione di metodi:

- **Attività 2A:** confronto con l'attività svolta da gruppi di ricerca con obiettivi affini;
- **Attività 2B:** sintesi, progettazione e implementazione di metodi per la predizione applicati a dati numerici, che soddisfino gli obiettivi previsti;
- **Attività 2C:** valutazione dei metodi realizzati, confronto con approcci esistenti e pubblicazione dei risultati conseguiti in riviste e conferenze internazionali.

Anno 3°: applicazione al dominio applicativo scelto e sviluppo della tesi di dottorato:

- **Attività 3A:** stage presso università straniera e confronto con l'attività svolta presso altri gruppi di ricerca con obiettivi affini;
- **Attività 3B:** raffinamento dei metodi e realizzazione di caratteristiche specifiche per il dominio applicativo scelto;
- **Attività 3C:** analisi dei risultati sperimentali ottenuti sul particolare dominio applicativo scelto;
- **Attività 3D:** stesura della tesi di dottorato.

10) Valutazione dei risultati.

In letteratura sono state proposte varie metriche per quantificare l'accuratezza delle predizioni di produzione.

L'uso di metriche differenti dipende dalle esigenze dell'utente: gli operatori necessitano di metriche per valutare il costo di errori di previsione, mentre i ricercatori necessitano di indicatori di performance dei differenti modelli di previsione.

Tra quelli maggiormente impiegati vi sono i seguenti:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{x_{\mathrm{Mod},i} - x_{\mathrm{Mea},i}}{x_{\mathrm{Mea},i}}\right)^2}{n}} \times 100\%$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where A_t is the actual value and F_t is the forecast value.

I metodi di previsione semplici possono servire come baseline rispetto a quelli di interesse per valutare le previsioni. Un modello molto semplice di riferimento è la previsione di persistenza, che assume che "le cose rimangono le stesse", ovvero implica il proiettare nel futuro i valori passati della variabile da predire (ad esempio la produzione di un sito, l'indice del ciel sereno, etc…).

11) Eventuali referenti esterni al Dipartimento

Durante gli studi, nonché durante la partecipazione a scuole estive inerenti agli obiettivi prefissati, saranno selezionati alcuni referenti stranieri, operanti presso Università della Comunità Europea, al fine di supportare il lavoro di stesura della tesi di dottorato. Possibili referenti esterni sono:

Laboratory of Artificial Intelligence and Decision Support, and Faculty of Economics, University of Porto Porto, Portugal