UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea (triennale) in Informatica

Elaborato finale

Data mining applicato ai dati meteorologici: sviluppo di un prototipo software per l'analisi meteorologica locale

Relatore: prof. Paolo Giorgini

Laureando: Paolo Cestari

Anno Accademico 2007-2008

a Cristina, Irene e Lorenzo

Indice dei contenuti

Introduzione

La meteorologia si avvale ormai da molti anni di gran parte delle più moderne tecnologie del settore dell'informazione e della comunicazione. Se pensiamo al miglioramento della previsione meteorologica avvenuto negli ultimi decenni e dovuto principalmente all'introduzione dei rilevamenti a terra, dei rilevamenti in libera atmosfera, dei rilevamenti satellitari, della modellistica numerica e del radar meteorologico, possiamo comprendere la quantità di lavoro svolto dal mondo scientifico e la quantità di nuove tecnologie applicate a questo settore.

All'interno della meteorologia generale, la specifica attività di previsione meteorologica è quindi caratterizzata dal pesante utilizzo delle informazioni fornite dai mezzi di acquisizione ed elaborazione dei dati sopra citati. Questi strumenti vengono utilizzati al fine di tradurre il risultato fornito in forma complessa da questi sistemi in una prognosi diffusa in linguaggio naturale attraverso i bollettini di previsione.

La ricerca in questo settore è in continua evoluzione nell'intento di migliorare la previsione sia in termini di miglioramento nello spazio (previsioni più puntuali), che in termini di miglioramento nel tempo (previsioni temporalmente più precise). I due aspetti spazio-tempo non sono affatto separati, anzi, la riduzione dell'orizzonte temporale comporta generalmente per l'utente finale l'attesa di una previsione più precisa sia in termini di spazio che di tempo mentre è generalmente tollerata una certa imprecisione nel medio periodo (2-5 giorni).

Le previsioni puntuali sono sempre più richieste e numerosi sono i siti web che forniscono questo tipo di servizio basandosi quasi esclusivamente sui risultati dei modelli numerici. Il grado di attendibilità per questo tipo di previsioni è perlomeno discutibile se non è filtrato da un operatore specializzato (il previsore meteorologico), in particolare se queste vengono utilizzate in un territorio con un'orografia complessa come ad esempio quello alpino.

La strada intrapresa dalla comunità scientifica per questo tipo di miglioramento è quella di integrare molte informazioni diverse fornite da questi strumenti; si utilizzano ad esempio

diverse fonti di informazione dei monitoraggi ambientali (satelliti, stazioni a terra, radiosondaggi) nella modellistica matematica, per prevedere le evoluzioni atmosferiche.

Esiste però ancora un campo di attività che si reputa possa ottenere un miglioramento qualitativo con l'utilizzo di strumenti di integrazione software e di apprendimento automatico. Si tratta del settore della previsione a brevissimo termine (fino a poche ore o al massimo 1 giorno) allorquando ci si trova nelle condizioni di non poter più avvalersi esclusivamente delle previsioni numeriche dei modelli matematici e dove entra quindi in gioco l'esperienza e la conoscenza locale del meteorologo che deve prevedere l'evoluzione atmosferica attraverso il dato grezzo fornito dagli strumenti di monitoraggio ambientale.

Questo tipo di attività viene ad assumere sempre maggiore importanza sia per gli aspetti di protezione civile locale, ma anche per un uso quotidiano in numerose situazioni (molte attività antropiche quali lavori nei campi, lavori all'aperto in genere, gestione viabilità, spettacoli e manifestazioni, ecc ...).

Dal punto di vista organizzativo questo si è tradotto ad esempio anche in Italia nel rivedere le modalità con cui si controllano le evoluzioni dei fenomeni meteorologici intensi trasferendo le competenze di vigilanza meteo, finora attribuite alle strutture centrali dello stato, alle regioni ed alle province, cioè ad organi decentrati sul territorio.

Gli strumenti utilizzati per la previsione a brevissimo termine sono prevalentemente le immagini trasmesse in tempo quasi reale dai satelliti, le immagini del radar meteo ed i dati rilevati dalle stazioni automatiche di misura a terra. Sui dati rilevati da queste ultime si concentrerà il presente lavoro.

I monitoraggi a terra sono effettuati attraverso una rete di stazioni meteorologiche elettroniche. Una parte molto limitata di queste informazioni entrano nel circuito internazionale delle stazioni sinottiche e sono finalizzate all'utilizzo nei sistemi di predizione matematica a medio termine (10 giorni). Un numero molto più consistente di queste stazioni automatizzate, sono invece gestite da enti e strutture regionali e sono finalizzate esclusivamente al miglioramento della conoscenza meteo climatica locale e, viste anche le nuove competenze delle regioni, all'attività di previsione a brevissimo termine.

Come si è verificato quindi anche in altri settori dell'attività umana, lo sviluppo dell'elettronica prima e dell'informatica e delle telecomunicazioni poi, ha comportato anche per il settore dei monitoraggi ambientali, un notevole aumento della quantità di informazioni acquisite ed una conseguente necessità di gestire in modo agevole un patrimonio informativo molto più consistente che nel passato.

Il presente lavoro nasce quindi dalla consapevolezza di disporre di una nuova e rilevante risorsa informativa ed è stato quindi il punto di stimolo che ha portato ad indagare negli schemi caratteristici delle serie di dati meteorologici delle stazioni a terra. Questi schemi, se scoperti, possono contribuire a migliorare la conoscenza locale, ma soprattutto la loro scoperta ha una finalità predittiva in quanto, una volta classificati, possono essere applicati a istanze di dati contingenti per predirne valori futuri.

L'obiettivo principale di questo lavoro è quindi quello di utilizzare tecniche di data mining per estrarre regole ed estrapolare schemi ricorrenti dalle serie di misure meteorologiche con la finalità di proporre poi uno strumento di supporto alla previsione a brevissimo termine basato sullo sfruttamento del patrimonio informativo fornito dalle stazioni a terra. Si intende quindi sviluppare un prototipo software orientato ad una consultazione via web in grado di agevolare il meteorologo regionale impegnato nel contesto della analisi dei fenomeni atmosferici.

La tesi è strutturata in 6 capitoli. Il capitolo 1 è dedicato all'analisi del contesto organizzativo per focalizzare lo stato dell'arte nel settore dell'utilizzo dei dati di monitoraggio delle stazioni meteorologiche a terra per la previsione a brevissimo termine.

Verranno poi presentate nel capitolo 2 le analisi svolte sui dati meteorologici attraverso l'utilizzo interattivo dello strumento software di data mining denominato WEKA (Waikato Environment for Knowledge Analysis) sviluppato dall'università di Waikato in Nuova Zelanda [9], con il quale sono state applicati metodi classici di data mining (l'analisi di regole dedotte con l'algoritmo “Apriori”, l'analisi delle giornate “meteorologicamente vicine”, l'analisi di regressione e la classificazione basata sul teorema di Bayes). Saranno valutate quindi le potenzialità di questi strumenti ai fini predittivi nel settore della previsione a brevissimo termine.

Si continuerà con la progettazione del sistema prototipale nella sua architettura principale (Capitolo 3) e si entrerà poi nei dettagli implementativi (Capitolo 4).

Nel capitolo 5 verrà presentato il prototipo di sistema software integrato ed orientato alla consultazione web che, sfruttando anche le interfacce di programmazione fornite dallo stesso strumento WEKA, consentirà di agevolare il decisore nell'analisi dei dati e di applicare delle predizioni automatiche attraverso le tecniche sopra citate.

Nel Capitolo 6 si presenteranno i test svolti e le validazioni effettuate sul prototipo per terminare poi con alcune conclusioni finali.

Capitolo 1 - La gestione e l'analisi delle informazioni meteorologiche locali

In questo capitolo vengono riportate le motivazioni che sono alla base del presente lavoro esponendo un'analisi parallela tra la gestione dei dati meteorologici e le attuali linee di tendenza nel campo della scienza e della tecnologia dell'informazione. Si descrive quindi il patrimonio delle informazioni meteorologiche locali esistenti rilevando l'opportunità di sfruttare questo patrimonio per estrarre conoscenze finalizzate alla previsione meteorologica a breve termine.

1.1 - La previsione meteorologica locale

L'uomo ha dapprima iniziato ad osservare il tempo atmosferico e, successivamente, ha affinato una certa sensibilità alle variazioni; in questo modo è riuscito a cogliere con l'esperienza le relazioni esistenti tra i fenomeni atmosferici e, da queste, è riuscito ad effettuare delle predizioni empiriche. Numerosi sono infatti i proverbi e le conoscenze tramandati dalla cultura popolare che hanno alla base un'acquisizione implicita di conoscenza attraverso l'esperienza.

A titolo di esempio riporto un paragrafo del capitolo 14 del libro di J.Kerkmanm e G.Kappenberger [6] che ben si presta ad essere adottato come modello di conoscenza meteorologica acquisita con l'esperienza.

> ***Il freddo delle pecore****: una singolare situazione che si ripete puntualmente ogni anno, è chiamato il freddo delle pecore. D'abitudine, dopo un primo riscaldamento del continente, nella prima decade del mese di giugno si ha un irruzione di aria nuovamente fredda, con correnti che tendono ad orientarsi a nord-ovest e da nord e che riportano il limite della neve a quote relativamente basse "cosicchè le pecore, già tosate, patiscono il freddo" (da cui sembra provenire l'espressione).*

Nel libro di G.Kappenberger e J.Kerkmann [6] si dedicano parecchi paragrafi alle situazioni meteorologiche tipiche, alle regole empiriche di previsione meteorologica ed alle

regole dedotte da altri fenomeni naturali che sono associabili a variazioni meteorologiche. Ci si riferisce unicamente a regole che hanno una motivazione scientifica, non di semplici proverbi che sono giustificabili solo dal fatto che hanno una rima.

Alcune regole, molto più specifiche, sono state dedotte in tempi più recenti attraverso l'analisi svolta sui dati dei rilievi strumentali; tipicamente questi dati sono rilevati da stazioni meteorologiche poste al suolo: a titolo di esempio si riporta, sempre da G.Kappenberger e J.Kerkmann [6], la seguente affermazione proprio per far comprendere l'importanza dell'analisi del dato meteorologico per la scoperta di schemi.

> ***Regole empiriche di interpretazione (cap. 12 par. 4.16)****: il rialzo della pressione atmosferica lento e graduale preannuncia un miglioramento più duraturo (...). Un rialzo più veloce corrisponde piuttosto ad un miglioramento passeggero(...).*

Numerose sono quindi le regole apprese con l'esperienza dalle popolazioni di montagna ed anche quelle dedotte con misurazioni strumentali dal mondo scientifico, ma in questo contesto possiamo affermare che anche **il tempo atmosferico ha delle regole proprie** e che queste regole si affinano con il passare del tempo e con la disponibilità di informazioni sempre più dettagliate.

E' necessario però riportare, sempre dal libro di G.Kappenberger e J.Kerkmann [6], l'affermazione che "**in meteorologia le eccezioni sono molte di più delle regole**". Questa affermazione racchiude in se una doppia interpretazione: da un lato ci fa capire che in meteorologia le evoluzioni sono molto complesse da analizzare, perché non disponiamo di informazioni così dettagliate sullo stato dell'atmosfera tali da poterle comprendere, dall'altro lato ci induce a credere, o perlomeno a sperare, che i dati in nostro possesso rilevati dalle stazioni automatiche a terra, nascondano al loro interno qualche relazione non immediatamente comprensibile tale da giustificare un'attività di data mining.

Per completezza si ricorda che accanto alle osservazioni e le rilevazioni umane e strumentali a terra, nel tempo si sono evolute molte altre tecniche di misurazioni strumentali che hanno migliorato la conoscenza dei fenomeni meteorologici.

Si ricorda il grande passo in avanti della previsione meteorologica dopo l'introduzione dei radio sondaggi[1] , o ancora con l'introduzione dei monitoraggi satellitari[2] , oppure l'utilizzo del radar in meteorologia[3].

Negli ultimi decenni, grazie anche all'avvento del calcolo numerico, queste tecnologie hanno permesso di sviluppare modelli fisico-matematici dell'evoluzione a medio termine (10 giorni) su larga scala (scala planetaria) o ad area limitata (poche decine di km). Il più importante per l'area europea è il modello fisico-matematico dell'ECMWF (European Centre for Medium Range Weather Forecats) di Reading (Inghilterra).

Tuttavia per la previsione meteorologica del giorno in corso e del successivo non è generalmente utilizzato il solo risultato dei modelli numerici, poiché questi non sempre riescono a descrivere bene la dinamica atmosferica, specie in zone con orografia complessa come le Alpi.

Sempre da G.Kappenberger e J.Kerkmann [6] riportiamo infatti le seguenti affermazioni:

- *"Per la previsione a corta scadenza, ossia per il giorno stesso e per quello successivo, oltre al sistema di previsione numerica, si utilizzano altri metodi, quali il metodo sinottico che si basa sull'estrapolazione empirica della velocità di spostamento dei sistemi meteorologici, oppure diversi metodi statistici";*

- *"Se da un lato la qualità delle previsioni numeriche diminuisce con il numero dei giorni, dall'altro lato è anche possibile che il terzo, quarto e quinto giorno siano meglio previsti del primo".*

Da queste affermazioni possiamo comprendere che esiste un periodo temporale (i primi

1 - I radiosondaggi sono misurazioni di alcune variabili meteorologiche effettuate attraverso una strumentazione attaccata ad un pallone sonda che viene rilasciato in atmosfera.

2 - Specifici satelliti lanciati in orbita per scopi di analisi e studio dei fenomeni meteorologici che rilevano e trasmettono al suolo ogni 15 minuti immagini nei canali del visibile, dell'infrarosso e del vapor d'acqua.

3 - Il radar è uno strumento che, attraverso l'emissione di onde elettromagnetiche e l'analisi degli stessi impulsi riflessi, consente di rilevare la presenza di oggetti distanti, di localizzarli nello spazio e di ottenere informazioni sulla loro natura fisico-geometrica. Nel caso particolare di un radar meteorologico, tali oggetti sono tipicamente le idrometeore, siano esse gocce di pioggia oppure neve, grandine o pioggia ghiacciata.

due giorni di previsione) in cui è possibile impiegare metodi di lavoro diversi; in particolare se sostituiamo la parola “statistici” della prima affermazione, con la parola “data mining”[4] che, in un'accezione abbastanza comune altro non è che l'applicazione di metodi statistici ad ampie basi di dati, possiamo circoscrivere un campo di studio che potrebbe risultare di interesse pratico. Si tratta cioè di cercare di sfruttare appieno, all'interno di sistemi di apprendimento automatico, il patrimonio informativo fornito dai monitoraggi ambientali e, per quanto riguarda in particolare il presente lavoro, quello fornito dalle stazioni meteorologiche a terra.

Più avanti verrà spiegato con maggior dettaglio cosa intendiamo con il termine “stazione meteorologica”, per ora accontentiamoci di pensare a questa come ad una fonte di informazioni abbastanza dettagliate in senso temporale che descrivono lo stato della bassa atmosfera in una circoscritta località.

Per quanto riguarda quindi l'utilizzo delle informazioni delle stazioni a terra per la previsione a breve scadenza possiamo vedere qual è lo scenario attuale. Con l'evoluzione delle tecnologie elettroniche e del trattamento automatico delle informazioni, è aumentato il numero di punti di misura (stazioni meteorologiche automatiche) e la velocità con la quale queste misure possono essere elaborate e trasmesse ai centri di analisi meteorologica.

Esistono oggi ad esempio in Italia numerose reti di stazioni elettroniche di misura a terra, gestite da enti regionali o provinciali, i cui dati non entrano nei flussi informativi dedicati alla predizione numerica a scala planetaria o in quelli ad area limitata, ma che sono finalizzate allo studio della meteorologia e della climatologia locale. Queste informazioni, acquisite elettronicamente, forniscono l'opportunità di tentare di estrarre della conoscenza e di sfruttarla all'interno dei sistemi informativi che supportano il decisore nell'analisi delle condizioni meteorologiche locali.

4 - Ci sono diverse definizioni per il termine data mining, tutte accomunate dal fatto di contenere espressioni tipo “grandi mole di dati”, “relazioni tra i dati”, “tecniche di analisi avanzate”; da wikipedia: il data mining è una tipica applicazione informatica (solitamente facente parte di un sistema esperto), usata per rintracciare (ed accorpare) dati significativi sepolti sotto una montagna di informazioni irrilevanti. Il termine inglese mining fa proprio riferimento al lavoro di estrazione che viene fatto nelle miniere. Altra definizione: è un processo atto a scoprire correlazioni, relazioni e tendenze nuove e significative, setacciando grandi quantità di dati immagazzinati, usando tecniche di riconoscimento delle relazioni e tecniche statistiche e matematiche".

Dobbiamo a questo punto però allargare per un attimo il campo visivo e proporre una breve analisi parallela tra la gestione delle informazioni di carattere meteorologico e tutto il settore della scienza e delle tecnologie informatiche poiché l'esperienza di quest'ultimo settore, applicata in altre e diverse discipline, ben si presta ad essere adottata come modello di sviluppo per il trattamento dell'informazione meteorologica. Possiamo riassumere l'esperienza fatta nel data warehousing e nel data mining dicendo che dalle basi informative, ed in particolare da quelle strutturate o semistrutturate, è possibile estrarre automaticamente della conoscenza per analogia.

Questa “estrazione di conoscenza”, detta appunto anche “data mining”, si effettua in generale sottoponendo ad una serie di noti algoritmi un certo numero di esempi, quello che viene chiamato insieme delle istanze di addestramento, dai quali il sistema apprende (o meglio sarebbe dire con il quale facciamo apprendere il calcolatore).

E' opportuno introdurre brevemente anche le tappe evolutive che hanno motivato l'introduzione del concetto di “data mining” nel settore delle scienze e delle tecnologie informatiche. Ricordando queste tappe evolutive, possiamo renderci conto che anche il settore dell'acquisizione dei dati meteorologici delle reti a terra ha avuto un'evoluzione analoga e può quindi rifarsi a questi modelli generali di evoluzione tecnologica nel trattamento dell'informazione.

Il concetto di “data mining” o “estrazione di conoscenza dai dati”, può essere visto come il risultato del percorso naturale nell'evoluzione della tecnologia dell'informazione così come ben evidenziato da Han e Kamber [5].

Dalle prime collezioni di dati gestite nei primi anni sessanta con le sole “primitive di accesso ai files”, siamo passati, negli anni settanta-ottanta, all'introduzione del concetto di “basi di dati relazionali” (DBMS, RDBMS). Quindi negli anni ottanta e novanta sono stati introdotti i concetti di “sistemi avanzati di gestione strutturata dell'informazione” che estendono il concetto di database relazionale in tre direzioni diverse: database avanzati (ad esempio per la gestione di dati multimediali), database per l'analisi (data warehousing) e sistemi adatti per l'integrazione e la condivisione delle informazioni sul web come lo standard XML (eXtensible Markup Language).

Come si può facilmente capire, la motivazione comune che ha comportato l'introduzione di nuovi concetti nella gestione dell'informazione, è il continuo aumento della quantità di informazione disponibile in forma digitale.

Da questo continuo aumento delle informazioni disponibili, nasce l'esigenza di pensare a nuove modalità di gestione del dato e, negli ultimi anni, si sente parlare sempre più spesso di “data mining”.

E' ben descritta infatti in Han e Kamber [5] la situazione in cui si presenta prepotentemente in un'organizzazione la necessità di introdurre il concetto di “data mining”. Laddove si nota infatti una situazione di abbondanza di informazioni accoppiata alla necessità di potenti strumenti software di elaborazione si presenta con evidenza il problema dell'incapacità umana di “comprensione” di questi dati.

Questa situazione viene descritta come “**la tomba dell'informazione**”. Conseguenza di questa situazione è che gli archivi vengono visitati raramente e le decisioni vengono prese dai decisori sulla base di intuizioni (non sempre corrette).

Sfortunatamente le procedure di costruzione e gestione di un sistema di gestione di vaste banche dati, sono operazioni dispendiose sia in termini di tempo che di denaro e vengono quindi spesso viste come non necessarie, mentre potrebbero dare dei contributi importanti nei settori di strategie di business, gestione della conoscenza e nella ricerca scientifica.

Il sistematico sviluppo degli strumenti di “data mining” sono il punto cruciale per trasformare la situazione di “tomba dell'informazione” in quella di miniera da cui estrarre dall'informazione le famose “**pepite d'oro**“, concetto che, come vedremo, ha comportato una esagerata attesa di risultati sorprendenti dall'attività di data mining .

Il concetto di “pepita d'oro”, che viene spesso citato come risultato finale delle attività di data mining, non deve essere inteso come un sorprendente risultato concretamente ed immediatamente utilizzabile che esce come da una bacchetta magica dall'applicazione di algoritmi informatici applicati ad una grossa base di dati; piuttosto può essere visto come

fiduciosa aspettativa riposta nell'informazione stessa, che porta ad una sistematica attività di riduzione del divario esistente tra la situazione di “ricchezza di dati e povertà di strumenti di trattamento” e la situazione in cui questa ricchezza di informazione viene continuamente salvaguardata cercando di organizzarla in modo tale che strumenti di analisi sempre più potenti ed in continua evoluzione possano gestirla al meglio oggi ma, possibilmente, anche in futuro. In altre parole se siamo altamente convinti che i dati di cui siamo in possesso sono molto importanti, li custodiremo e li tratteremo come fossero delle pepite d'oro.

Il concetto viene ripreso in modo simile anche da Ian H.Witten e Eibe Frank [7], non esistono “pepite d'oro” nel data mining, non esistono segreti da scoprire impostando algoritmi di calcolo su oceani di dati, ma piuttosto esistono **tecniche ben conosciute che possono estrarre delle conoscenze utili** da un patrimonio informativo importante.

Ecco quindi che l'attenzione è più spostata verso il processo di gestione dei dati piuttosto che nel risultato finale, proprio per non riporre aspettative troppo grandi alle dispendiose attività di gestioni di grossi archivi di dati. Possiamo anche affermare in modo poco rigoroso che, in fondo, con l'attività di data mining cerchiamo di creare **negentropia**[5] nel campo dell'informazione digitale laddove, negli ultimi anni sono aumentate le quantità di dati e, molto spesso, anche il disordine in questi.

Era importante premettere queste osservazione sul data mining poiché anche in campo meteorologico, ed all'interno di questo anche pensando ai soli rilievi automatici delle stazioni a terra, esistono enormi quantità di dati che vanno valorizzati e accuditi come un tesoro.

Ritornando allo specifico tema di questo lavoro, se disponiamo quindi di un'insieme di informazioni strutturate, come quelle fornite dai rilievi automatici a terra, possiamo cercare di far apprendere al calcolatore regole e relazioni importanti nei dati e dedurne degli schemi evolutivi caratteristici di una specifica località, con poca differenza rispetto a quanto ha sempre

5 - Con negentropia si esprime in un’unica parola un concetto che, in modo raro, si usa per indicare in fisica, in biologia, nonché nella scienza dell’informazione, una situazione di entropia negativa. Poiché l’entropia si può vedere come uno stato di disordine e di degrado, l’opposto, cioè la negentropia (o entropia negativa), può essere vista come fattore ordinante o di rinnovamento vitale nelle fasi cruciali degli esseri viventi (ripreso da http://www.silviaratti.com/nna.htm).

fatto l'uomo anche quando non disponeva di strumenti di analisi così potenti come ai giorni nostri.

In questo modo, oltre che cercare di migliorare la conoscenza locale, possiamo anche avere l'ambizione di costruire un ponte di congiunzione tra le tradizioni antiche in campo meteorologico e la scienza moderna. Non è una pura nostalgia che porta a voler gettare un ponte tra l'antico ed il moderno ma, come spesso accade, con l'avvento di nuove tecnologie che sostituiscono l'attività umana, si fanno dei progressi, ma si perde qualcosa per strada; in questo caso specifico si possono perdere delle conoscenze locali.

Da queste motivazioni di fondo che riguardano la scienza informatica in generale ma che, come abbiamo visto sono perfettamente calzanti anche per la gestione di una banca dati meteorologica, scaturiscono le motivazioni del presente studio.

Riassumendo, l'obiettivo del presente lavoro è quello di verificare se, applicando tecniche di data mining ad un insieme di dati rilevati dalle stazioni meteorologiche automatiche a terra, è possibile estrarre delle associazioni, delle regole e/o dei pattern significativi tali da consentire di migliorare l'analisi e la conseguente previsione a breve o brevissimo termine.

1.2 - Gestione dei dati meteorologici locali

Figura1:Stazione meteorologica

Iniziamo quindi a vedere con maggior dettaglio di che cosa stiamo parlando.

Tralasciando gli stati primordiali delle osservazioni meteorologiche allorquando le “pseudo conoscenze” non venivano archiviate ma erano patrimonio degli stregoni, da più di cento anni oramai l'uomo è abituato ad archiviare informazioni di carattere meteorologico con strumenti che, nel tempo, si sono evoluti.

Dapprima su tabelle cartacee, poi via via con strumenti

meccanici (es. termografi, pluviografi, etc...) per arrivare all'era elettronica dei giorni nostri.

A partire dagli anni '80 hanno incominciato a diffondersi metodologie di rilevamento dati meteorologici basate sulla strumentazione elettronica che permette di raccogliere le informazioni con una frequenza temporale molto più elevata che nel passato (come minimo oraria ma spesso anche sub oraria).

Figura 2: Stazione meteorologica installata

Ogni punto di raccolta di informazioni di carattere meteorologico è in gergo chiamato stazione. Una stazione meteo è costituita da un insieme di sensori che trasformano la misura fisica in un segnale più facilmente gestibile (es. una temperatura può essere trasformata in un valore elettrico di tensione o corrente); quindi il segnale può essere digitalizzato e trasformato in un informazione numerica che viene gestita da un'unità centrale (datalogger) la quale mantiene il dato in una memoria; un'esempio di strumentazione meteorologica è rappresentato nella figura 1, mentre in figura 2 è rappresentata una reale installazione in campo.

Normalmente le stazioni sono destinate almeno a raccogliere informazioni di tipo termo-pluviometrico (temperatura e precipitazione), ma molto spesso sono anche equipaggiate per raccogliere dati anemometrici (vento), igrometrici (umidità), barometrici (pressione atmosferica), radiometrici (radiazione solare). In alcuni casi raccolgono informazioni molto più specifiche quali l'idrometria (altezza livello del fiume), nivometria (altezza neve), la freatimetria (altezza acqua in falda), etc... .

La stazione meteo è spesso dotata di un'unità aggiuntiva che svolge il compito di comunicare alla centrale l'informazione. Generalmente la stazione viene interrogata dall'ufficio deputato alla vigilanza attraverso uno specifico software di centrale, ma i datalogger di ultima generazione (ormai spesso dei mini computer con sistema operativo a bordo), permettono di gestire autonomamente anche la politica di trasmissione dei dati al centro. La comunicazione

avviene attraverso una rete radio dedicata oppure attraverso la rete pubblica GSM[6]/GPRS[7] o, in qualche raro caso, ancora su rete telefonica PSTN[8]. A causa dei costi elevati di trasmissione solo in rari casi si usano invece reti satellitari.

Con frequenza sub oraria ed un ritardo di pochi minuti, si può disporre quindi di gran parte dei dati rilevati a terra dalle stazioni locali. Un insieme di stazioni meteo forma quindi quella che viene chiamata rete meteorologica.

Le reti possono essere classificate anche secondo la distanza infra-strumentale, cioè la distanza esistente tra una stazione e l'altra: se questa è nell'ordine di pochi km, sono classificate dalla letteratura di settore come reti meteorologiche a mesoscala (territori di poche decine di kmq distanze interstrumentali dell'ordine di 10 km) e sono finalizzate a studi e caratterizzazioni di tipo meteo-climatici locali; se invece la distanza intrastrumentale è dell'ordine di decine o centinaia di km sono classificate come reti sinottiche e sono finalizzate al supporto della modellistica fisico-matematica di tipo meteorologico.

In provincia di Trento esistono 2 stazioni meteorologiche che fanno parte della rete sinottica, ed i cui dati, di proprietà del Servizio Meteorologico Nazionale dell'Aeronautica Militare, entrano nel circuito internazionale finalizzato alla modellistica fisico-matematica di previsione a medio termine.

Esistono inoltre tre principali reti meteorologiche caratterizzate ciascuna da una propria finalità che fanno capo a differenti gestioni amministrative ed insieme formano una rete di circa 300 punti di misura collegati:

- la rete dell'Ufficio Previsioni e Organizzazione del Dipartimento Protezione Civile della Provincia Autonoma di Trento, che è orientata principalmente allo studio ed alla previsione dei fenomeni meteorologici, nivologici e climatologici;

6 - Global System for Mobile Communications (GSM) è attualmente lo standard di telefonia mobile più diffuso nel mondo.

7 - General Packet Radio Service (GPRS) tecnologia per trasmissioni dati a pachetto su canali GSM condivisi attraverso tecniche di multiplazione numerica.

8 - Public Switched Telephone Network o, più semplicemente, la rete telefonica tradizionale.

- la rete del Servizio di Piena della Provincia Autonoma di Trento che è orientata principalmente all'acquisizione e gestione dell'informazione di tipo idrologico (altezza idrometrica dei corsi d'acqua);

- la rete dell'Istituto Agrario di S.Michele All'Adige - Fondazione Edmund Mach, orientata principalmente alla ricerca in campo agro-meteorologico, alla climatologia ed al supporto informativo agli agricoltori (ad esempio prevenzione gelate primaverili).

Come è intuibile quindi, la crescita in quantità di queste informazioni, ha comportato un progressivo sviluppo anche dei sistemi di gestione centrale. Ognuna delle strutture proprietarie delle reti, ha a disposizione uno o più sistemi gestionali (diversificati anche in relazione alle differenti tecnologie utilizzate nel corso del tempo), alcuni basati su formati dati binari proprietari, altri hanno strutturato le proprie archiviazioni su database relazionali. Esiste quindi una certa disomogeneità nel trattamento dei dati causata sia dalla differente gestione amministrativa, sia dalla differente tecnologia impiegata, sia dalla diversa finalità con la quale vengono raccolte queste informazioni. Agevolare la condivisione di queste informazioni risulterebbe sicuramente proficuo per tutti i soggetti interessati all'analisi di questi dati.

Dopo questa breve illustrazione della tecnologia utilizzata per i rilevamenti meteorologici, possiamo spostare l'analisi dello stato dell'arte volgendo l'attenzione alla modalità di gestione delle informazioni. Anche in questo caso, come è stato fatto sopra, un parallelo con lo stato dell'arte in campo meteorologico e le attuali tendenze del settore delle scienze e tecnologie dell'informazione, consente di comprendere l'importanza del settore di “integrazione dell'informazione” in molti campi di attività.

L'esempio della gestione dell'informazione meteorologica della provincia di Trento qui riportato, fa emergere questo aspetto: l'opportunità di integrare informazioni provenienti da sistemi diversi per poter svolgere analisi ad ampio spettro sui dati. Anche da questo punto di vista quindi, in meteorologia si manifesta una tendenza attuale comune nel campo della tecnolgia della comunicazione e dell'informazione, come ben evidenziato da Hector Garcia-Molina, Jeffrey D. Ullman e Jennifer Widom [2] laddove si dice infatti che “una nuova e larga famiglia di applicazioni si sta sviluppando, sulla base delle ampie fonti di dati disponibili, sotto il nome di

information integration".

Le opportunità fornite da un sistema integrato sono molte: quella più interessante, per i nostri scopi, è di poter gestire tutti i dati provenienti da diverse fonti come se fossero un'unica fonte e, conseguentemente, anche le funzioni di estrazione di conoscenza e di analisi, sono meglio e più omogeneamente applicabili in modo indipendente dalla sorgente.

Queste considerazioni sull'integrazione dei dati in campo meteorologico, sembrano far deviare dall'obiettivo principale del presente lavoro. Infatti non è propriamente il tema in esame la costituzione di un data warehouse di dati meteorologici; potrebbe essere sufficiente sviluppare un'applicazione che non consideri il lavoro di integrazione dei dati. Ma la realizzazione di un sistema di integrazione di dati, comporta necessariamente la progettazione di un sistema ottimizzato per l'analisi dei dati anziché di un sistema ottimizzato per la gestione delle transazioni operative. Quindi le due cose sono intimamente legate: da un lato fare analisi vuol dire costruire un database ottimizzato per l'analisi, dall'altro si costruisce un data warehouse proprio per mettere le basi ad un sistema ottimizzato per analisi.

I sistemi per l'analisi si differenziano da quelli operazionali perlomeno per i seguenti aspetti fondamentali:

- le attività operazionali coinvolgono il livello tecnico dell'organizzazione mentre quelle di analisi generalmente coinvolgono i responsabili delle decisioni al vertice dell'organizzazione ;
- nei sistemi operazionali tipicamente si tende ad ottimizzare i vari processi software finalizzati ad intervenire su pochi record nel più breve tempo possibile (inserimento di pochi nuovi record o ricerca di piccole quantità di dati) mentre sul database di analisi si tende ad ottimizzare il processo di computazione delle aggregazioni di dati che coinvolge spesso anche l'intero database;
- nel database operazionale si mantengono tutti i dati di dettaglio, mentre sul data warehouse si mantengono i dati più significativi e, spesso, solo in forma aggregata;
- nel database operazionale tipicamente ci sono spesso molti processi concorrenti o molti utenti che richiedono e/o modificano solamente pochi singoli record mentre nei database

orientati all'analisi ci sono generalmente pochi processi (o utenti) che richiedono un'attività computazionale anche pesante sull'intero database;

Se parliamo delle sole informazioni fornite dalle stazioni a terra di un singolo ufficio regionale, va infatti valutata con estrema attenzione l'introduzione di un'attività così onerosa qual è quella di costituzione e gestione di un data warehouse.

Se però allarghiamo la visione anche al fatto che ci sono altre informazioni meteorologiche che possono essere integrate o che lo potranno essere tra non molto (penso ai rilievi radar meteorologici o alle numerose fotocamere che si stanno installando in diversi punti del territori o alle immagini trasmesse dal satellite meteosat) allora potremmo pensare ad un **data warehouse** che consenta l'analisi simultanea di più informazioni provenienti da sistemi di osservazione anche concettualmente molto diversi tra loro.

Anche nelle analisi propedeutiche all'installazione del radar meteorologico ora installato sul monte Macaion in provincia di Trento, viene riportato il concetto della necessità di integrazioni di fonti informative diverse laddove si dice che “il miglioramento delle previsioni di precipitazione a breve termine sembra conseguibile utilizzando, secondo un albero decisionale, i modelli numerici di previsione e i flussi cospicui di dati forniti, in tempo reale, da radar meteorologico e da satellite” [12] e, si aggiunge per completezza, dalle stazioni a terra.

Non è un'attività impensabile in campo meteorologico anzi, per la verità, qualcuno ci ha già pensato. Un tentativo di costituzione di un data warehouse è stato fatto nell'anno 2002 dalla Provincia Autonoma di Trento con il progetto Aquarium, peraltro limitato alla gestione delle sole informazioni termo-pluviometriche e con finalità unicamente di controllo del ciclo idrologico. Allo stato attuale il progetto si è però interrotto per una serie di motivazioni non solo tecniche, ma anche organizzative.

D'altro canto **i rischi** che ci sono nell'intraprendere progetti di data warehouse sono ben evidenziati in letteratura; M.Golfarelli e S.Rizzi nel suo **“***Data warehouse : teoria e pratica della progettazione”* [4] espongono in maniera estremamente chiara i rischi a cui ci si espone nella realizzazione di un database per l'analisi che sono legati principalmente alla modalità di

gestione del progetto, alle tecnologie impiegate, all'organizzazione dei dati, all'organizzazione in genere.

Uno progetto avviato ed ancora in fase di realizzazione, ma estremamente interessante nel settore del data warehousing meteorologico, è quello dell'Ufficio Federale di Meteorologia e Climatologia della Svizzera dove, da alcuni anni, è in corso di realizzazione un sistema integrato per l'analisi dei dati meteorologici [10] .

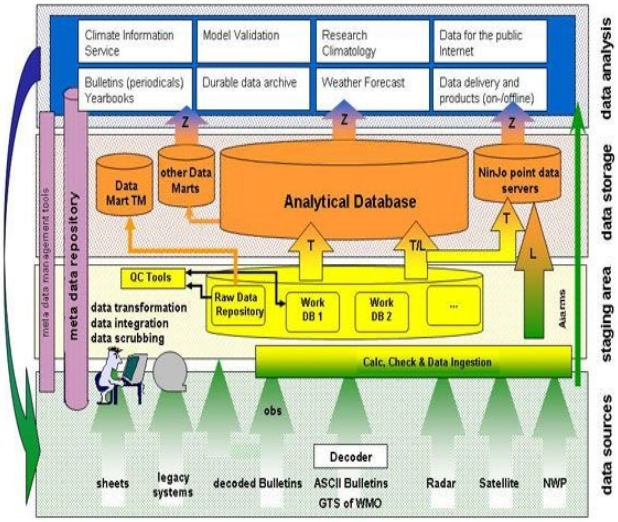


Figura 3: progetto data warehouse del centro meteo Svizzera

Il progetto si propone di costruire un'efficiente ed estensibile infrastruttura hardware e software per potenziare le capacità di analisi dei dati meteorologici.

L'elemento chiave del sistema è la centralizzazione dei dati e la costituzione di un database ottimizzato per l'analisi dei dati.

Nel sistema i dati sono consolidati seguendo un processo di trasformazione e correzione in modo da preservarli nella forma più consistente per un utilizzo anche futuro. Secondo quanto riportato dall'ufficio federale Svizzero, **la rapidità e la flessibilità di accesso alle informazioni consolidate è di fondamentale importanza per le previsioni del tempo e per l'analisi climatologica.**

Questo oltre che confermare la validità degli obiettivi che ci si propone è fondamentale per definire le linee di tendenza nello sviluppo di un sistema informativo meteorologico, per preservare nel tempo i dati e poter effettuare analisi sui dati anche per le previsioni a brevissimo termine.

L'immagine di figura 3, ripresa dal sito web di meteo Svizzera [10], evidenzia come il sistema di analisi sia il livello più elevato di elaborazione delle sorgenti di dati e fa comprendere l'onerosa attività necessaria per integrare i dati.

Nella maggioranza dei casi comunque, nelle regioni e province italiane la gestione dell'informazione delle reti meteorologiche locali è attualmente affidata a sistemi operazionali perlopiù proprietari e spesso disomogenei.

Nella previsione a breve termine si incontrano quindi dei limiti nell'operatività immediata per la difficoltà di applicare tecniche di analisi e sarebbe quindi auspicabile l'introduzione di sistemi di analisi quasi real-time. Un real-time data warehousing è un concetto non sempre ben accettato nel campo del data warehousing perché non si ritiene opportuno gestire un flusso di dati quasi real-time con le normali tecniche di caricamento dei dati nel sistema di analisi attraverso gli strumenti ETL (extraction, transformation e loading) solitamente operanti in modalità batch; chi è favorevole invece suggerisce di utilizzare strumenti CTF (capture, transform e flow) per il caricamento dei dati perché questi ultimi si possono attivare a seguito di nuovi eventi di inserimento dati nei database operazionali (trigger).

Tornando al tema del presente lavoro, una delle opportunità in cui vale la pena investire per il miglioramento della previsione a breve o brevissimo termine, è fornito dalla presenza di numerose stazioni meteorologiche locali come di seguito esposto.

1.3 - Un'opportunità: le numerose stazioni meteo provinciali

Come citato da J.Kerkmanm e G.Kappenberger [6], il paradosso della meteorologia è quello di voler fare delle previsioni esatte sulla base dello stato dell'atmosfera non conosciuto esattamente nei dettagli.

Nelle analisi svolte in questo lavoro, abbiamo un'ulteriore limitazione: vogliamo fare delle previsioni conoscendo solo, nella migliore delle ipotesi, le misure dello strato atmosferico al suolo in una zona molto limitata. Anche se le evoluzioni meteorologiche al suolo sono sicuramente correlate con quelle in quota questo è sicuramente un limite, ma si ritiene che sia compensato dal fatto che ci accontentiamo di una previsione molto limitata sia in senso temporale che in senso spaziale.

Nulla vieta tuttavia in futuro di integrare un eventuale sistema informativo con le fonti informative dei modelli numerici che descrivono la situazione prevista in quota e la tendenza nell'evoluzione temporale in modo da migliorare ulteriormente il sistema oppure con altre sorgenti di informazione come quelle fornite dal radar meteorologico o dai satelliti.

Dopo queste anticipazioni, che ne circoscrivono anche le aspettative, riassumo per punti principali, le motivazioni che hanno dato inizio al presente lavoro.

L'esistenza di numerose stazioni meteo regionali e quindi di una mole di informazioni in continuo aumento da queste rilevate, unitamente alla sensazione di poterle gestire in modo più proficuo è un sintomo che, comunemente a ciò che accade in altri settori, è ben conosciuto nel campo della scienza e della tecnologia della gestione dell'informazione e questo suggerisce di valutare l'opportunità di intraprendere la strada del data warehousing in campo meteorologico come processo di gestione e analisi dell'informazione meteorologica.

Una prima, anche se limitata, valutazione di quanto possa essere efficace un sistema ottimizzato per l'analisi, può essere apprezzata attraverso l'utilizzo di un prototipo software che consenta agli esperti del settore di testare le potenzialità di una gestione integrata delle informazioni in campo meteorologico, dopodiché questa potrebbe essere la linea di tendenza evolutiva tracciata nell'ottica di valorizzare il patrimonio informativo accumulato e di integrare fonti informative diverse.

Come si vedrà più avanti, dalle analisi svolte sui dati meteorologici attraverso il tool di data mining WEKA [9] emergono tre cose fondamentali: la prima è quella che si intravedono delle opportunità di estrarre della conoscenza dalle informazioni, la seconda che questa conoscenza può essere di aiuto al previsore meteo nella delicata fase di analisi predittiva dell'evoluzione meteorologica a breve termine, la terza che lo stesso strumento WEKA, può essere sfruttato non solo come strumento di analisi interattiva, ma può essere “pilotato” anche da interfacce di programmazione Java e quindi può essere agevolmente integrato in una propria applicazione software.

Come si può comprendere inoltre da quanto in precedenza evidenziato, nella previsione a breve scadenza non si utilizzano solo i risultati dei modelli numerici, ma anche i dati forniti da

altri strumenti. Tra questi meritano di essere meglio considerate tutte le stazioni meteorologiche a terra.

Infine, come aspetto molto particolare, l'analisi delle regole associative dei parametri meteorologici, potrebbe fornire anche un interessante metodo per proporre in un modo facilmente comprensibile le regole meteorologiche locali “apprese automaticamente” le più importanti delle quali sono state patrimonio, nel passato, delle sole persone che vivono costantemente in ambiente naturale (contadini, gente di montagna, etc...). Questo nuovo e molto più numeroso patrimonio di regole “apprese automaticamente” dal calcolatore ed esposte in linguaggio comprensibile, può quindi essere consultato dai previsori meteo come supporto all'analisi meteorologica locale, ma anche dai non addetti ai lavori come strumento di supporto all'acquisizione di conoscenza meteorologica locale.

In questo contesto si riuscirà forse a far comprendere solo in minima parte le numerose potenzialità di analisi fornite dallo strumento di data mining utilizzato ed utilizzabili anche nello specifico settore dei dati meteorologici,.

Si tratta comunque di una nuova opportunità che, se avrà un seguito, potrà svilupparsi secondo una linea di tendenza abbastanza comune nel campo del data mining: si inizia cioè con l'applicazione di metodi di analisi più semplici per definire successivamente schemi di apprendimento più complessi.

Capitolo 2 - L'analisi dei dati meteorologici attraverso uno strumento di data mining

In questo capitolo vengono descritti alcuni metodi di estrazione di conoscenza disponibili anche nello strumento di analisi WEKA ed utilizzati per l'analisi dei dati meteorologici. Vengono quindi descritti i risultati ottenuti dall'applicazione su un insieme di istanze di una stazione meteorologica locale. Infine viene descritto come le conoscenze estratte possono essere rappresentate.

Nel presente lavoro, sono stati analizzati i dati disponibili in rete e fruibili dal sito web dell'Istituto Agrario di S.Michele All'Adige - Fondazione Edmund Mach[13]. In particolare sono stati costruiti e mantenuti aggiornati nel corso del lavoro due insiemi di dati con frequenza oraria, il primo proveniente dalla stazione meteorologica di Trento Sud, il secondo dalla stazione meteorologica di Arco per un periodo temporale di circa 8 anni a partire dal primo gennaio 2001 fino ad oggi.

Sono stati organizzati ed utilizzati i dati orari dei seguenti parametri:

- TA: temperatura media oraria rilevata a due metri dal suolo (espressa in °C);
- RH: umidità relativa media oraria rilevata a due metri dal suolo (%);
- RR: pioggia totale oraria (mm)
- RS: radiazione solare globale (MJ/mq)
- VV: media oraria velocità vento rilevata a 10 m dal suolo (m/s)
- DV: media oraria direzione vento rilevata a 10 m dal suolo (gradi/360)
- BAR: media oraria pressione atmosferica (mbar)

Da questi parametri sono state inoltre derivate alcune ulteriori variabili, in particolare per la temperatura dell'aria, la pressione atmosferica e l'umidità relativa sono state derivate le rispettive tendenze definite come il coefficiente angolare della retta di regressione dei dati nel periodo utilizzato ai fini predittivi.

Al fine di poter analizzare i dati, valutare l'accuratezza delle classificazioni o delle associazioni e verificarne quindi le capacità predittive, è stato utilizzato lo strumento di data mining WEKA, per il quale si fornisce una breve illustrazione del suo utilizzo rimandando al sito web del progetto WEKA [9] ed al libro di I.H.Witten e E.Frank [7] per maggiori dettagli.

2.1 - Lo strumento di analisi WEKA

Lo strumento WEKA (Waikato Environment for Knowledge Analysis) è una collezione di algoritmi di apprendimento automatico che possono essere applicati sia direttamente ad un insieme di dati, attraverso l'interfaccia con la quale l'utente interagisce con il sistema, oppure richiamandoli dal proprio codice di programmazione in linguaggio java.

WEKA[9], sviluppato dall'Università di Waikato in Nuova Zelanda, è giunto alla versione 3.4, nella versione collegata al libro di I.H.Witten e E.Frank [7] ed alla versione 3.5.8 nella versione di sviluppo; richiede che sul sistema sia installato l'ambiente di run time Java RE almeno nella versione 1.4 ed esegue quindi su diverse piattaforme operative quali Linux, Windows, Mac Os. E' disponible con licenza “GNU General Public License” ed è liberamente scaricabile da http://www.cs.waikato.ac.nz/ml/WEKA/ da dove si può facilmente accedere inoltre a numerose pagine di supporto e documentazione.

WEKA fa uso di una propria terminologia il cui significato, per i termini di uso più frequente, viene di seguito specificato:

- **classe** (o concetto): rappresenta ciò che stiamo cercando, cioè quello che deve essere appreso dal modello automatico; è un'informazione di input nel senso che ognuna delle istanze utilizzate per l'addestramento deve contenere il valore della classe di appartenenza di se stessa rispetto a tutti i valori della classe che rappresenta il concetto in esame;

- **istanze**: insieme di informazioni (attributi) che descrivono i casi in esame;

9 - La sigla WEKA corrisponde anche al nome di un curioso uccello in via di estinzione che è presente solo nelle isole della Nuova Zelanda.

- **attributi**: serie di valori ognuno dei quali descrive un particolare aspetto delle varie istanze;

- **rappresentazione della conoscenza**: è il modo con il quale la conoscenza appresa in modo automatico viene rappresentata; tipicamente vengono rappresentate attraverso liste di decisioni, regole di associazione, regole basate su alberi di decisione, regole basate su istanze (vicinanza del caso in esame con altre istanze nel set di training), cluster di istanze vicine, cluster gerarchici;

- **algoritmi di apprendimento**: gli algoritmi che implementano lo schema di apprendimento, sono classificabili in varie tipologie a seconda dei metodi utilizzati; tipicamente statistici, divide-et-impera, alberi di copertura, associativi, lineari, basati sulle istanze o basati sulla clusterizzazione;

- **training set**: è l'insieme di tutte le istanze che sono disponibili per l'algoritmo al fine dell'apprendimento;

- **test set:** è l'insieme di istanze utilizzate dall'algoritmo per valutarne la sua credibilità; il test set deve essere diverso dal training set;

- **credibilità**: la credibilità del risultato viene valutata attraverso specifici metodi di controllo; gli strumenti più diffusi sono cross-validation, leave-one-out, bootstrap; si tratta poi di valutare anche il costo di un'errata classificazione; per questo si usano tabelle di contingenza, matrici di confusione, roc analisi;

- **error rate**: è la proporzione tra gli errori fatti sull'intero set delle istanze;

Il software WEKA può gestire i dati di input per l'apprendimento in diversi modi; collegandosi ad un database con standard SQL attraverso i driver java JDBC, importando i dati da file CSV, leggendoli da sorgenti XML o da URL, e altro ancora.

Il modo nativo, che spesso è conveniente usare nella fase iniziale di analisi dei dati è il

formato “Attribute-Relation File Format (ARFF)”; si tratta di un file di testo composto da un'intestazione che descrive il tipo di attributi utilizzati e da un elenco di istanze usate per l'addestramento, una per ogni riga, in ognuna delle quali si ripetono i valori degli attributi secondo l'ordine descritto nella testata.

Il semplice esempio qui sotto riportato illustra in un modo che si ritiene sufficientemente chiaro il formato dati ARFF; si noti in particolare la differenza tra i due tipi di attributi ammessi: numerici (nell'esempio gli attributi “temperatura” e “velocita_vento” sono dichiarati numeric) o discreti (nell'esempio “direzione_vento” e “pioggia” sono enumerati).

L'ultimo attributo (la pioggia) descrive la classe di appartenenza dell'istanza; la classe è quell'attributo che istruisce il sistema facendogli comprendere che, con la situazione in cui si presentano gli attributi in precedenza, ne consegue questo risultato.

Esempio di file ARFF

```
@relation <NOME-RELAZIONE>
@attribute temperatura numeric
@attribute velocitaVento numeric
@attribute direzione_vento {Nord,NordEst,Est,....,Sud,...}
@attribute pioggia Assente,Debole,Moderata,....,}
@data
-3.6, 0.4, Nord, Assente
2.1, 1.2, Sud, Assente
5.4, 2, 0, Est, Debole
```

Una volta preparati i dati ci sono diversi modi per interagire con il sistema. Per il nostro scopo di analisi è conveniente utilizzare l'interfaccia di esplorazione (explorer). Le restanti funzionalità sono perlopiù adatte ad un utilizzo diverso degli stessi algoritmi presenti in explorer, come l' interfaccia a riga di comando (CLI), l'interfaccia di comparazione delle performance di differenti schemi di apprendimento (experimenter), e l'interfaccia di combinazione dei flussi di apprendimento che consente di connettere in sequenza le varie fasi (input dei dati, preprocessamento, classificazione,valutazione) e salvare il flusso operativo (knowledge flow gui); l'immagine riportata di seguito in figura 4 rappresenta l'interfaccia “explorer” di WEKA.

WEKA dispone di strumenti di pre-processamento dei dati (filtri), di classificazione, di

regressione, di clusterizzazione, di analisi delle regole associative, di selezione degli attributi rilevanti e di visualizzazione ai fini di sommarizzazione dei dati. Inoltre WEKA mette a disposizione anche strumenti per la costruzione di propri schemi di apprendimento.

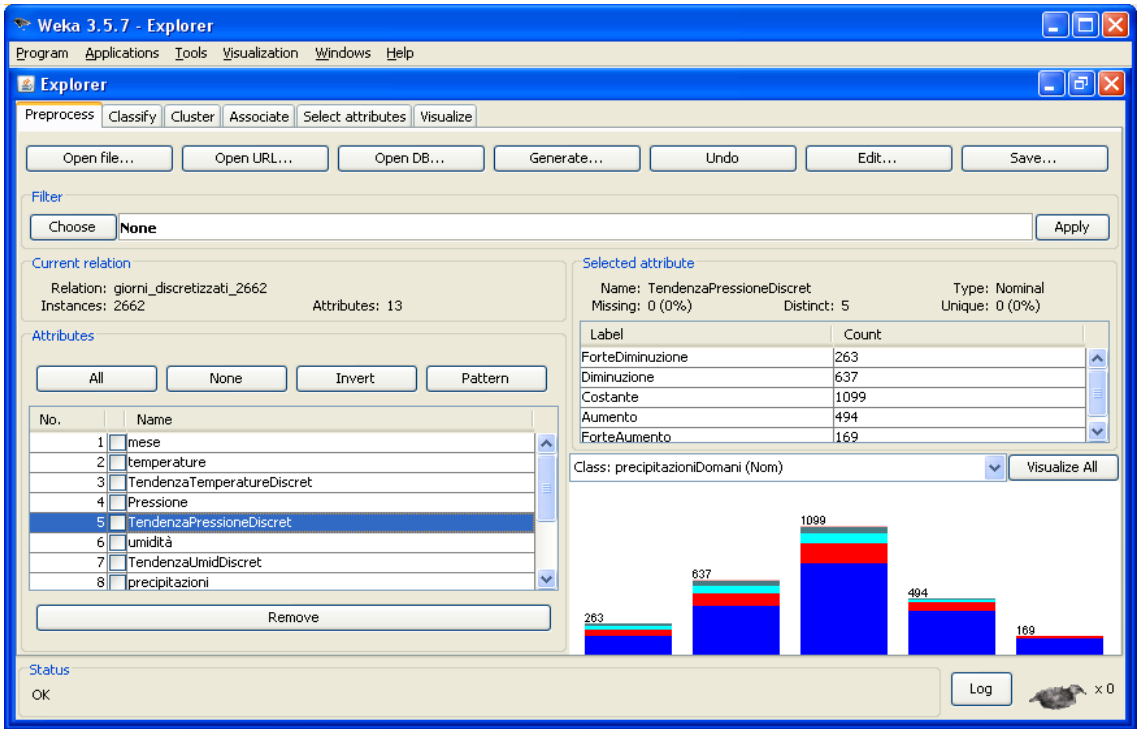


Figura 4: schermata interfaccia di WEKA explorer

Per ognuno degli algoritmi di analisi WEKA implementa anche controlli di credibilità e strumenti di comparazione finalizzati alla verifica delle capacità di apprendimento.

2.2 - Il processo generale di analisi dei dati

Molto spesso, in qualsiasi campo di attività, la scoperta di nuove conoscenze è il risultato di un processo che viene svolto progressivamente all'interno di un'organizzazione ed è quindi una conseguenza di una serie di passi che normalmente vengono intrapresi nell'intento di migliorare le proprie condizioni di lavoro. Se vogliamo analizzarla anche da questo punto di vista, **l'attività di data mining è quindi la formalizzazione di un modello comportamentale di un'organizzazione**.

Esiste un'ampia letteratura al proposito che schematizza in sette passi fondamentali il processo di estrazione di conoscenza:

- data cleaning (rimozione dei dati inconsistenti e dei rumori);
- data integration (combinazione delle diverse sorgenti in un unico sistema);
- data selection (selezione dei soli dati interessanti per l'analisi);
- data trasformation (trasformazione dei dati nel modo più opportuno rispetto alle analisi che si intende svolgere);
- data mining (applicazione dei metodi di analisi);
- pattern evaluation (valutazione dei pattern scoperti);
- knowledge rappresentation (presentazione delle conoscenze acquisite);

I primi quattro punti in realtà identificano un unico processo chiamato (anche all'interno dello strumento WEKA) pre-processamento dei dati, mentre gli ultimi tre sono l'applicazione delle tecniche di mining e l'esposizione dei risultati finali.

Analizzerò brevemente le fasi sopra elencate, lasciando al capitolo successivo un approfondimento più specifico in relazione anche al lavoro svolto sulla serie di dati meteorologici.

La prima attività da svolgere è quindi quella di prendere coscienza dei dati in possesso; l'attività è generalmente indicata come **descrizione sommaria dei dati** ed aiuta nello studio delle caratteristiche generali dei dati in possesso. Essa ha generalmente il compito di descrivere come sono distribuiti i dati, evidenziare errori o valutare i dati isolati (outliers) presenti nel database.

Una tecnica per rappresentare la distribuzione dei dati si ottiene con diagrammi “box plot” che rappresentano in modo efficace la mediana, il primo ed il terzo quartile , il valore massimo, quello minimo ed i punti isolati esterni all'intervallo $[q1-1{,}5*(q3-q1)],[q3+1{,}5*(q3-q1)]$ (dove q1 e q3 sono rispettivamente il primo ed il

terzo quartile).

Esistono inoltre anche altre modalità di rappresentazione quali gli istogrammi, gli istogrammi di frequenza, i diagrammi scatter e altro ancora.

Dopo avere descritto sommariamente i dati è necessario procedere con la loro "**pulizia"**. I dati presenti nei database operazionali, cioè i database utilizzati per l'operatività continua, sono spesso affetti da errori, inconsistenze o mancanze e, dal momento che l'attività di estrazione di conoscenza è direttamente dipendente dalla qualità dei dati utilizzati, si rende necessario effettuare una pulizia preventiva dei dati.

Per questo si può procedere in diversi modi:

I. ignorare le registrazioni mancanti o eliminare le registrazioni che non hanno un numero sufficientemente elevato di attibuti corretti; è una tecnica piuttosto povera ma adatta in alcuni casi quando la registrazione è affetta da una notevole percentuale di mancanza di dati;

II. riempire le mancanze di dati nelle registrazioni manualmente: adatta generalmente solo nei casi in cui le mancanze sono relativamente basse;

III. riempire i buchi con un valore costante che sia indicativo del dato mancante per un determinato algoritmo di analisi: esempio un dato numerico con valore 99999 può indicare all'algoritmo la mancanza del dato;

IV. usare la media tra il/i valori precedenti ed il/i valori successivi al periodo di mancanza del dato come valore da utilizzare per riempire le posizioni mancanti;

V. usare la media della classe di appartenenza dell'istanza: cioè se l'istanza è classificabile attraverso gli altri attributi come appartenente ad una certa classe, si utilizza la media di quella classe per il valore mancante;

VI. usare il valore più probabile ottenuto dall'applicazione di algoritmi statistici usando gli altri attributi dell'istanza in esame.

Quando i dati provengono da fonti diverse è necessario procedere con l'attività di **integrazione delle fonti**, cioè si deve ricondurre il tutto ad una forma omogenea. E' un'attività molto delicata e spesso onerosa ed è soggetta a problemi che si devono considerare attentamente. Dallo schema di due database sorgenti differenti che gestiscono informazioni analoghe, è infatti possibile valutare, erroneamente, che due attributi sono uguali (per esempio nella codifica di un attributo dobbiamo chiederci se effettivamente sono state utilizzate da entrambe le sorgenti gli stessi codici oppure no).

E' spesso anche necessario effettuare l'attività di **selezione dei soli attributi interessanti** ai fini dell'analisi. Molto spesso è la conoscenza del dominio applicativo che sovrintende quest'attività, ma non è sbagliato ricorrere a specifici algoritmi di selezione automatica: l'analisi dei sottoinsiemi di attributi rilevanti (capitolo 7.1 di I.H.Witten e E. Frank [7]) o l'analisi delle componenti principali (capitolo 7.3 di I.H.Witten e E. Frank [7]) , sono due tecniche abbastanza comuni che possono ridurre gli attributi utilizzati ad un suo sottoinsieme più limitato, mantenendo inalterate le potenzialità di analisi e/o consentendo di migliorare le performance dell'algoritmo di data mining .

Ai fini di migliorare le performance degli algoritmi, è possibile anche **ridurre il numero di attributi** con tecniche di accorpamento di due o più attributi in un solo nuovo attributo che complessivamente li descriva, oppure discretizzando e/o la gerarchizzando dei concetti. E' anche possibile **ridurre il numero delle istanze** aggregando i dati.

Talvolta è necessario procedere con la **trasformazione dei dati** in ingresso in modo da poter applicare determinati algoritmi, oppure di migliorare le performance dell'analisi. Un tipico esempio è quello di normalizzare tutti i valori di alcuni attributi numerici in valori compresi in un range (es [-1,+1]) in modo che algoritmi di analisi delle distanze o di classificazione, possano essere applicati e/o agevolati.

Una volta effettuate queste attività, si può procedere con l'applicazione degli algoritmi di apprendimento automatico. I principali algoritmi di apprendimento, tendono ad estrarre

conoscenza dai dati cercando schemi ricorrenti (frequent pattern), associazioni e correlazioni tra attributi, classificazione delle istanze, raggruppamento (clusterizzazione) delle istanze, ricerca di similarità.

Dopo aver appreso le caratteristiche dei dati in esame, è necessario circoscrivere l'oggetto di nostro interesse cioè la classe obiettivo del nostro "cercare" e quindi individuare un attributo che descriva la categoria di appartenenza delle istanze in esame.

Quindi si può provare ad elaborare il nostro insieme di istanze con uno o più tra gli algoritmi disponibili. Questi algoritmi hanno campi di applicazione in cui sono meglio performanti rispetto ad altri e quindi la successiva fase di valutazione e comparazione consentirà di individuare quelli che meglio si prestano al nostro problema.

Illustrerò ora un piccolo sottoinsieme di tecniche di data mining che troveranno poi riscontro nell'applicazione al caso dei dati meteorologici e nelle quali si enfatizzeranno gli aspetti più importanti riferibili alle analisi svolte.

2.2.1 - Analisi degli schemi frequenti, delle associazioni e delle regole

L'analisi degli schemi frequenti, delle associazioni e delle **regole associative** (Frequent Pattern & Associations Analysis), presuppone che i dati in esame siano di tipo discreto. Nulla toglie però che si possano discretizzare i dati in ingresso oppure utilizzare dei preprocessori che facciano il lavoro automaticamente.

Una volta che i dati sono disponibili in forma discretizzata, possiamo analizzare ogni singola istanza tra tutte quelle in esame estraendo una serie di indici di associazione. Cioè possiamo vedere se nel dataset ci sono combinazioni, sequenze o strutture di attributi che appaiono più frequentemente di altre. Nella sua forma più semplice, un tipico esempio è la Market Basket Analysis (MBA) in cui si analizzano gli oggetti che il cliente di un negozio mette nel carrello verificando poi ad esempio che, tra tutti gli utenti, esiste una determinata percentuale di essi che acquista contemporaneamente il prodotto A e il prodotto F (poniamo 5%), e che tra tutti quelli che comperano A un'altra determinata percentuale compera anche F (poniamo l'80%).

Ho indicato con un esempio il tipo di analisi che si può fare sui dati discretizzati ma implicitamente ho riportato gli indici che ne danno la misura di significatività dell'associazione, l'indice di supporto e l'indice di confidenza. Espresso nel linguaggio probabilistico, il supporto dell'esempio precedente s = P(A U F) , cioè il 5% dei clienti compera A e F insieme mentre la confidenza è c = P(F|A), cioè tra tutti quelli che acquistano A, l'80% compera anche F (probabilità condizionata).

Il passo successivo è proporre questo risultato come una regola: una regola è un'espressione del tipo “quando <A> allora <F>”; “quando <A>” è la premessa, “allora <F>” è la conseguenza; in altre parole quando ci sono determinate premesse <A> ne consegue <F>.

La regola dell'esempio precedente viene generalmente scritta nella forma: A (0.05) => F (0.8), cioè l'acquisto di A implica l'acquisto di F con un supporto del 5% ed una confidenza dell 80%.

Ovviamente in una serie di istanze non ci sono solo due attributi, per cui si possono trovare numerosissime regole del tipo A=>F, A^B => C, B^F^H=>A; si possono trovare quindi regole più generali (che coinvolgono pochi attributi) e regole più specifiche (che coinvolgono molti attributi in più rispetto a quella generale).

L'**algoritmo “Apriori”**[10] è una delle tecniche per effettuare questo tipo di analisi, esso cioè analizza le istanze iterativamente in due passi: prima genera insiemi di attributi frequenti e, da questi, estrapola regole del tipo sopra indicato. L'importanza di questo algoritmo risiede però per molte applicazioni nella possibilità di fissare la conseguenza delle regole da minare. Possiamo ad esempio chiedere all'algoritmo di restituire solo le regole in cui la conseguenza è la classe da minare, cioè richiedere l'estrazione delle sole regole che hanno come conseguenza il concetto che vogliamo conoscere.

10 - Apriori è un algoritmo per la costruzione automatica di regole associative. Esso costruisce associazioni di n item a partire da associazioni con n-1 item se queste ultime superano il supporto desiderato; cioè viene costruito progressivamente un itemset a partire da un itemset più piccolo. L'algoritmo apriori è ben descritto in wikipedia all'URL http://it.wikipedia.org/wiki/Algoritmo_apriori e sul libro di Han e Kamber [5] al capitolo 5.2.

Il limite di questo algoritmo è invece il fatto che può produrre un numero molto elevato di regole da interpretare. Anche se queste possono essere filtrate, nel senso che è possibile richiedere all'algoritmo di restituire solo le regole che superano una certa soglia di supporto e/o confidenza, rimane comunque il problema di capire se sono state escluse regole magari importanti.

Il tool di analisi WEKA espone le regole trovate in questo modo:

mese=feb media_ta=Basse tendenza_ta=Aumento tendenza_rh=Costante totale_rr=Assenti 47 ==> totale_rr_domani=Assenti 47 conf:(1)

che sta a significare che nel mese di febbraio, quando la media di temperatura dell'aria è bassa, la tendenza della temperatura è in aumento, l'umidità media dell'aria è costante e le precipitazioni sono assenti, anche le precipitazioni domani saranno assenti; inoltre la regola ci dice che è supportata dal fatto che 47 istanze del database corrispondono ai requisiti sopra indicati e che di queste 47 istanze, 47 (cioè tutte, il 100%) hanno la conseguenza indicata (cioè pioggia domani assente).

Ci sono anche altri metodi per derivare regole da insiemi di dati, come ad esempio dedurle da alberi di decisione derivati a sua volta dall'applicazione sui dati di tecniche di “divide et impera” o di “algoritmi di copertura” (capitoli 4.3 e 4.4 di I.H.Witten e E. Frank [7]).

2.2.2 - Analisi della regressione

La correlazione è una tipologia di analisi applicabile ad attributi di tipo numerico che fa parte di una famiglia di tecniche di classificazione lineare i cui principi sono alla base di molti altri algoritmi.

L'idea è quella di esprimere la classe di appartenenza di un'istanza da classificare come una combinazione lineare degli attributi con un predeterminato peso secondo la seguente:

$x = w_0 + w_1 * a_1 + w_1 * a_1 + ... + w_n * a_n$ dove x è la classe, $a_1,a_2,..,a_n$ sono i valori degli attributi e $w_0,w_1,...,w_n$ sono i pesi degli attributi.

I pesi sono calcolati attraverso i dati di addestramento scegliendoli tra quelli che minimizzano la differenza al quadrato tra il valore della classe attuale ed il valore della classe predetta, l'algoritmo cioè tende minimizzare la: $\Sigma^{n}_{(i=1)}(x^{i}-\Sigma^{k}_{(j=0)}(w_j * a^{i}_{j}))^2$ dove i è l'istanza i-esima e j l'attributo j-esimo.

La regressione lineare è un buon metodo per le predizioni numeriche ed ampiamente usato in applicazioni statistiche, anche se soffre dello svantaggio della linearità; se i dati in esame hanno una dipendenza non lineare, viene presa la migliore combinazione lineare che meglio si adatta all'insieme di training.

Molto spesso la regressione lineare viene utilizzata per costruire blocchi di schemi di apprendimento più complessi.

2.2.3 - Classificazioni statistiche

Classificare le istanze significa raggrupparle in gruppi omogenei o anche associare ad ognuna di esse un'etichetta in modo tale che la classe di appartenenza di successive istanze ancora non classificate possa essere predetta dalle caratteristiche dei raggruppamenti fatti in precedenza.

Una famiglia di tecniche abbastanza performanti per categorizzare un insieme di attributi discreti, è quella basata sul teorema della probabilità condizionata di Bayes. Una variante molto pratica di questo algoritmo è quella chiamata "**Naive Bayes**" che usa tutti gli attributi in modo tale che ognuno di essi posa dare un contributo alla decisione, cosicché, se questi sono statisticamente indipendenti, la probabilità condizionata di un evento H_i (H_i è uno dei possibili eventi di una data classe di eventi H) dato un insieme E di attributi è la seguente:

$$P[H_i|E]=(P[E_1|H_i]*P[E_2|H_i]*...*P[E_n|H_i]*P[H_i])/P[E]$$

Questa formula, applicata ad ogni possibile valore i della classe H che descrive il concetto da cercare, permette di poter classificare l'istanza, scegliendo quella la cui probabilità condizionata $P[H_i|E]$ è più alta.

Nonostante l'assunzione che gli attributi siano statisticamente indipendenti, l'algoritmo

funziona spesso abbastanza bene anche se questi non lo sono, inoltre si possono usare filtri di pre-processamento per la selezione dei soli attributi rilevanti, così da eliminare le ridondanze.

Uno dei problemi che affliggono l'algoritmo si presenta quando una delle probabilità $P[E_j|H_i]$ è uguale a zero, cioè quando il valore di un attributo si presenta con probabilità 0 in congiunzione con il valore della classe. In questo caso la predizione sarebbe compromessa (perchè uguale a 0), ma con l'applicazione di una variante dell'estimatore di Laplace[11] si riesce ad evitare il problema.

Il classificatore Naive Bayes si presta bene ad essere utilizzato in prima applicazione per essere poi sostituito, magari dopo anni di esperienza, qualora siano stati sviluppati classificatori più sofisticati. Esso apprende quindi in modo probabilistico le conoscenze dalle istanze di addestramento.

2.2.4 - Analisi di serie temporali

Una serie temporale è una sequenza di valori acquisiti ad intervalli di tempo regolari ed è una tipologia di attributo molto popolare nello studio dei fenomeni naturali. L'enfasi di una serie temporale viene riposta nel fatto che gli eventi sono ordinati nel tempo.

Una variabile che rappresenta una grandezza che evolve nel tempo, può essere vista come una funzione del tempo $Y=f(t)$ e per essa diventa particolarmente efficace una sua rappresentazione cartesiana.

Diversi sono gli aspetti rilevanti nell'analisi delle serie temporali: l'analisi delle ciclicità, l'analisi delle stagionalità, l'analisi delle irregolarità, l'analisi della tendenza (trend analysis) e

11 - L'estimatore di Laplace viene applicato in questo modo: se abbiamo n numeri frazionari, che possono rappresentare ad esempio delle probabilità (es. 2/9 , 3/9 , 4/9, 0/9), aggiungiamo 1 ad ogni numeratore e compensiamo ogni denominatore con l'aggiunta di un numero N di 1 pari a quante sono le frazioni (nell'esempio si otterrà 3/13, 4/13, 5/13, 1/13 eliminando quindi la frazione che condurrebbe ad un probabilità 0). Una variante di questo estimatore, consente di fissare N a piacere ed aggiungere ad ogni numeratore il valore N/n dove n è il numero di frazioni (nell'esempio se poniamo N=1 le frazioni diventano (2 + 1/4)/10 , (3 + 1/4)/10, (4 + 1/4)/10, (0 + 1/4)/10).

l'analisi di similarità.

La formula di cui sopra può quindi essere modellata con una funzione a quattro variabili $Y = f(T, C, S, I)$ dove T è la modellazione della tendenza, C è la modellazione dei movimenti ciclici, S è la modellazione dei movimenti stagionali, I è la modellazione delle irregolarità. Esistono numerosi studi molto approfonditi nel settore dell'analisi delle serie temporali.

Nell'analisi della tendenza e delle ciclicità si può suddividere la tendenza a lungo termine, la ricerca di movimenti ciclici, la ricerca di movimenti stagionali e la ricerca di irregolarità sporadiche.

Nell'analisi delle **similarità** si tende invece a cercare nella serie storica una sequenza di eventi che differiscono il meno possibile dalla sequenza di eventi in esame. Un metodo abbastanza semplice per la ricerca delle similarità è il calcolo della distanza euclidea. Il calcolo della distanza euclidea è applicabile tipicamente ad attributi di tipo numerico; per gli attributi di tipo enumerato si parla invece di misure di prossimità, per le quali esistono tecniche di calcolo basate sulla binarizzazione dei dati discreti, come riportate ad esempio da P.Giudici [3].

Per gli attributi di tipo numerico come in genere sono le serie temporali, la **distanza euclidea** è definita come: $d_{(i,j)} = \sqrt{(\Sigma^{n}_{(r=1)}(x_{(ir)} - x_{(jr)})^2)}$ dove la distanza tra due sequenze i e j di n istanti temporali ciascuna è calcolata come la radice quadrata della sommatoria delle differenza dei valori che assumono nell'istante r-esimo elevata al quadrato.

Un problema sorge quando l'analisi di similarità si esegue su più di un attributo numerico contemporaneamente, in quanto attributi differenti hanno valori che spaziano in intervalli differenti (es. la temperatura ha un intervallo di valori che, nelle potenzialità dello strumento di misura, possono variare tra -30 e + 50 °C, l'umidità tra 0 e 100 %, la pressione atmosferica a Trento tra 950 e 1020 mBar).

Quindi, nel calcolo della distanza euclidea, le variabili verrebbero ad assumere un peso assolutamente diverso ed è pertanto necessario riportare le variabili ad una forma normalizzata

tale che ognuna di queste possa contribuire in egual misura alla valutazione dell'indice di similarità. Ogni variabile è quindi ricondotta ad una forma adimensionale con la normalizzazione $HN_i = (H_i - E(H_i))/s(H_i)$, dove: HN_i è la variabile normalizzata, H_i è la variabile, $E(H_i)$ è la media e $s(H_i)$ è la deviazione standard.

E' possibile poi eventualmente assegnare un peso ragionato ad una variabile piuttosto che ad un'altra; in questo caso si può optare per un fattore di moltiplicazione da applicare appena dopo la normalizzazione delle variabili e prima del calcolo della distanza.

2.2.5 - La credibilità dell'apprendimento

La valutazione dei risultati è la chiave per fare dei progressi nel campo del data mining e per queste valutazioni abbiamo quindi la necessità di utilizzare dei metodi sistematici ed oggettivi.

Ogni insieme di istanze manifesta in genere una particolare struttura tale da suggerire l'applicazione di un certo tipo di algoritmo piuttosto che un altro. E' necessario quindi introdurre dei metodi per verificare la credibilità della classificazione e per poter confrontare un classificatore rispetto ad un altro.

Ci troviamo quindi con un insieme di istanze utilizzate per l'addestramento del sistema ma non è opportuno usare le stesse istanze di addestramento per misurare la credibilità. Per una valutazione della credibilità è necessario che il classificatore sia applicato ad un insieme di istanze mai utilizzate in precedenza. Questo insieme di istanze è chiamato insieme di test e solo dall'applicazione dell'algoritmo a queste ultime istanze si possono calcolare degli indici di credibilità.

Un metodo abbastanza comune ed in ascesa nel campo del data mining è la validazione incrociata o "**cross-validation**": si suddivide l'insieme delle istanze disponibili in un certo numero di partizioni, poniamo 10 (impostazione standard) e poi, ripetendo la procedura 10 volte, ognuna di queste viene usata una volta come insieme di test mentre la parte restante viene usata come addestramento. In questo modo ogni istanza è stata usata almeno una volta come test.

Una metodologia spinta di “cross-validation”, chiamata “leave-one-out” consiste nel partizionare l'insieme delle istanze di addestramento in n parti, dove n è il numero di istanze e quindi permette di usare per il test un'istanza alla volta e per l'addestramento n-1 istanze (un utile riferimento bibliografico per la cross-validation sono i capitoli 5.3 e 5.4 di I.H.Witten e E. Frank [7]).

Gli indici utilizzati per le misure di credibilità sono diversi laddove ci siano attributi numerici rispetto a quelli utilizzati per gli attributi discreti. Ne citerò alcuni che sono stati utilizzati nell'analisi dei dati meteorologici. Il coefficiente di correlazione e l'errore medio assoluto sono due indici usati per valutare le classificazioni di istanze la cui classe sia numerica. Il numero di istanze correttamente classificate, la tabella di contingenza, la matrice di confusione, l'indice kappa-statistic sono indici e metodi per valutare i classificatori di istanze discrete.

Il **coefficiente di correlazione** ci consente di verificare se le variabili hanno andamenti che sono fra loro in relazione lineare (diretta o inversa). Anche se non può essere utilizzato come un indice di causa-effetto l'indice è molto utilizzato specie se usato insieme ad altri indici di errore.

L'**errore medio assoluto** è un indice che ci consente di capire qual è mediamente l'errore qualora utilizzassimo un determinato classificatore per una predizione numerica; si tratta cioè della media degli errori assoluti commessi dal classificatore.

Il numero di **istanze correttamente classificate** è un indicatore della bontà di un certo classificatore per istanze di tipo discreto, ma da solo non ci fa capire quanto le previsioni corrette siano casuali. Per questo si usa rappresentare la classificazione di istanze la cui classe è discreta, con le matrici di contingenza o con la matrice di confusione.

La **tabella di contingenza** qui sotto riportata (figura 5), consente ad esempio di valutare un classificatore che ammette due classi “Sì-No”.

Da questa si riescono a dedurre degli indici numerici quali la percentuale di veri

positivi, cioè la percentuale di istanze correttamente individuate come “Sì” tra tutte quelle che effettivamente erano “Sì”: TP rate = TP/(TP+FN) , la percentuale di falsi positivi cioè la percentuale di istanze classificate come “Sì” tra quelle che invece erano “No” FP rate = FP/(FP+TN), la percentuale di successi: S = (TP+TN)/(TP+TN+FP+FN) e la percentuale di errori E=1-S.

		Classe predetta	
		Sì	**No**
Classe attuale	**Sì**	Veri positivi (TP)	Falsi negativi (FN)
	No	Falsi positivi (FP)	Veri negativi (TN)

Figura 5: esempio di tabella di contingenza

Per istanze che hanno una possibile classificazione multipla, la tabella di contingenza non è più utilizzabile. Per questo si utilizza **la matrice di confusione** dove esiste una riga ed una colonna per ogni classe; la colonna di una classe rappresenta la predizione fatta, la riga invece è la classe a cui appartiene effettivamente l'istanza.

La tabella qui sotto riportata rappresenta il risultato di un'ipotetica classificazione a tre classi (Debole, Moderato, Forte) dove ad esempio il numero 12 evidenziato in colonna 3 e riga 3 rappresenta il numero di istanze classificate come “Forte” e che effettivamente erano di tipo “Forte” (cioè istanze correttamente classificate), il numero 6 della cella appena superiore rappresenta invece 6 istanze che sono state classificate come “Forte” ma che invece erano di tipo “Moderato”.

	Debole	**Moderato**	**Forte**	
Debole	88	10	2	100
Moderato	14	40	6	60
Forte	18	10	12	40
	120	60	20	**200**

Esempio di matrice di confusione

Come è intuibile, tutte le istanze sulla diagonale rappresentano istanze correttamente

classificate e questo valore (nell'esempio 88 + 40 + 12 = 140) diviso per il numero totale delle istanze (200) è la percentuale di classificazioni corrette (70%) .

Ma questo indice da solo non basta; classificare un'istanza come “Debole” quand'essa era “Forte” , è diverso che classificarla come “Debole” quand'essa era “Moderata”. Per questo si introduce l'indice kappa-statistic.

L'indice **kappa-statistic**[12], un numero compreso tra 0 e 1, rappresenta una misura che ci consente di capire che accordo c'è tra le classificazioni del modello in esame e quelle di un modello che sceglie casualmente la classe. Un indice pari a 1 indica un classificatore in totale disaccordo con il modello casuale, cioè un ottimo classificatore. L'esempio che segue ci aiuterà a capire come viene dedotto l'indice kappa-statistic.

Riferendosi sempre alla matrice di figura 6 e leggendo le somme per ogni colonna, notiamo che il nostro classificatore ha classificato 120 istanze come “Debole”, 60 come “Moderato” e 20 come “Forte”. Leggendo invece le somme sulle righe, notiamo che nella realtà abbiamo 100 istanze di tipo “debole”, 60 di tipo “moderato” e 40 di tipo “forte”.

Possiamo quindi chiederci: cosa farebbe un classificatore che sceglie casualmente la classe ma che abbia lo stesso numero totale di istanze classificate per ogni classe? La risposta è che tenderebbe a distribuire le 100 istanze della prima riga (“Debole”) secondo le percentuali 120/200, 60/200 e 20/200. Allo stesso modo distribuirebbe le 60 istanze della seconda riga e le 40 della terza e quindi da un classificatore casuale ci attendiamo la seguente matrice di confusione:

	Debole	**Moderato**	**Forte**	
Debole	60	30	10	100
Moderato	36	18	6	60
Forte	24	12	4	40
	120	60	20	**200**

Matrice attesa da un predittore casuale

Il numero di istanze utilizzate (somma delle righe) sono le stesse della matrice precedente ed abbiamo assicurato inoltre per il predittore casuale lo stesso numero di

12 - Per maggiori dettagli sul calcolo dell'indice kappa-statistic, vedere capitolo 5.7 di I.H.Witten e E. Frank [7]

classificazioni per le classi “Debole”,”Moderato” e “Forte”. Quello che cambia invece è il numero di classificazioni corrette di questo predittore casuale 60+18+4=82 (41%).

Se confrontiamo i due predittori, vediamo che per il nostro classificatore ci sono 58 successi in più rispetto al classificatore casuale (140 – 82=58) su un totale di 118 possibili successi in più (200 – 82). La proporzione di extra successi rispetto ai possibili extra successi è l'indice kappa-statistic che cerchiamo; nel nostro esempio 58/118=0,49.

2.3 - L'analisi dei dati meteorologici

L'analisi dei dati meteorologici è stata effettuata attraverso la descrizione sommaria e la pulizia dei dati, descritta di seguito nel capitolo 2.3.1 e l'applicazione di metodi di data mining, descritta nei seguenti capitoli dal 2.3.2 al 2.3.5.

2.3.1 - Il pre-processamento dei dati meteorologici

Nella fase di sommarizzazione e pulizia dei dati, il metodo di verifica si è basato principalmente sulla visualizzazione delle distribuzioni dei dati (svolta attraverso gli strumenti WEKA di visualizzazione), sulla ricerca dei dati mancanti e sulla ricerca dei dati isolati attraverso una visualizzazione grafica.

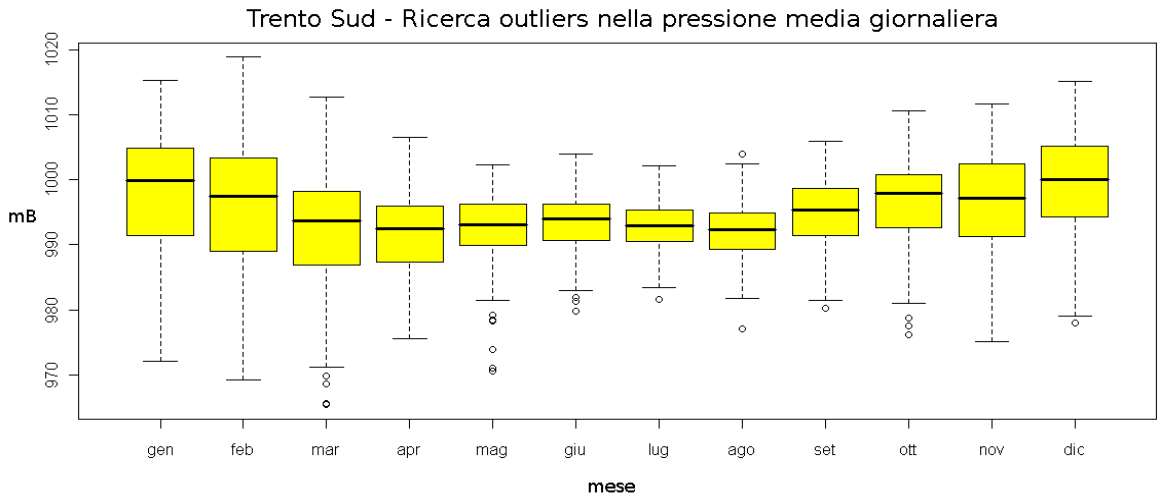


Figura 6: esempio di ricerca di casi isolati nei dati

Sono state quindi fatte per ogni stazione e per ognuna delle variabili meteorologiche (Temperatura, Pressione, etc...) delle rappresentazioni grafiche stile boxplot[13] come quelle riportate in figura 6 che mettono in evidenza le misure di centralità, di dispersione ed i casi isolati.

Non sempre gli outliers identificano situazioni di dati errati. In molti casi sono situazioni meteo particolari: ad esempio vento di foehn particolarmente intenso con conseguenti temperature elevate in mesi invernali, che possono essere segnalati come degli outliers.

Nei casi di dati mancanti o errati si è proceduto, dipendentemente dalla variabile meteorologica analizzata, in questo modo:

- se esistono per il corrispondente periodo delle misure nelle zone limitrofe (stazione meteo di Aldeno o in alternativa Trento Nord), allora desumo l'andamento della temperatura nell'intervallo mancante e lo adatto alla stazione di Trento Sud traslando la serie di un valore pari alla differenza media tra le due stazioni, sotto la condizione che l'intervallo di dati mancanti non superi le 6 ore;

- per la pressione atmosferica e la radiazione solare ho adottato i metodi di correzione che in letteratura vengono citati come “smoothing by bin boundaries” (che riempie la posizione mancante con i valori vicini) o “smoothing by bin mean” (che riempie la posizione mancante con la media dei valori vicini);

- per la velocità del vento ho usato l'interpolazione lineare;

- in alcuni rari casi ho eliminato delle registrazioni dal database perchè affette da troppe mancanze di dati.

A questo punto ho considerato che l'insieme di dati fosse sufficientemente “pulito” per iniziare a fare delle analisi.

13 - WEKA non dispone di visualizzazioni stile boxplot, per queste elaborazioni è stato usato il pacchetto di statistica free “R” (http://www.r-project.org/).

2.3.2 - L'analisi basata sulle regole

Per l'analisi delle regole ho individuato un intervallo temporale all'interno del quale sia possibile determinare l'insieme di attributi che devono essere utilizzati nelle istanze di addestramento. Per questo ho definito il giorno come intervallo e, all'interno di questo, ho utilizzato come attributi i valori aggregati (medie, totali, tendenze) dei dati orari di ognuna delle variabili meteorologiche.

Questo aspetto temporalmente rigido delle regole, ne può limitare in parte il loro utilizzo applicativo ma, in questa prima applicazione prototipale, è utile semplificare il quadro di analisi ponendo delle basi di ragionamento al fine di poter costruire eventualmente in futuro schemi di apprendimento più complessi che possono riconsiderare questa particolarità.

La scelta del giorno come periodo di analisi, è stata fatta perché già di per se esso ha una sua regolarità meteorologica determinata dalla rotazione terrestre, quindi sembra plausibile ottenere delle regole, tuttavia altri schemi di apprendimento potrebbero considerare periodi più ridotti e/o diversi e/o attributi non necessariamente riconducibili allo stesso periodo temporale.

Premesso questo, ho utilizzato le seguenti variabili al fine di determinare gli attributi dell'insieme di istanze che utilizzerò per l'addestramento:

- mese (attributo discreto che serve per tenere in considerazione l'aspetto stagionale dell'andamento meteorologico);

- temperatura (temperatura media dei 24 valori orari di temperatura dell'aria espressa in gradi centigradi – sigla “media_ta”);

- tendenza della temperatura dell'aria (coefficiente angolare della retta di regressione che descrive il trend dei 24 valori giornalieri di temperatura del giorno in esame – sigla “tendenza_ta”);

- pressione atmosferica (pressione atmosferica media dei 24 valori giornalieri espressa in millibar – sigla “media_bar”);

- tendenza della pressione atmosferica (coefficiente angolare della retta di regressione che descrive il trend dei 24 valori giornalieri di pressione del giorno in esame – sigla “tendenza_bar”);

- umidità relativa dell'aria (umidità media dei 24 valori giornalieri espressa in % – sigla “media_rh”);

- tendenza dell'umidità relativa (coefficiente angolare della retta di regressione che descrive il trend dei 24 valori giornalieri di umidità del giorno in esame – sigla “tendenza_rh”);

- velocità vento (velocità media dei 24 valori giornalieri espressa in m/s – sigla “media_vv”);

- precipitazioni (totale della precipitazione giornaliera espresso in mm – sigla “totale_rr”);

- radiazione solare globale (valore integrato sul giorno della radiazione solare espresso in MJ/mq – sigla “totale_rad”);

- vento prevalente tra le ore 12 e le ore 20 (attributo che ci fornisce indicazioni sull'attività delle brezze locali (brezza di valle), dedotto dalla direzione che ha assunto il vento nel momento di massima intensità all'interno dell'intervallo temporale considerato – sigla “cls_vento_prevalente_12_20”);

- vento prevalente tra le ore 21 e le ore 23 (attributo che ci fornisce indicazioni sull'attività delle brezze locali (brezza di monte) – sigla “cls_vento_prevalente_21_23”);

- classe (è l'attributo utilizzato per classificare l'istanza; a seconda delle regole che ho voluto minare, ho utilizzato rispettivamente la temperatura media giornaliera del

giorno successivo, la precipitazione media giornaliera del giorno successivo, il vento medio giornaliero del giorno successivo, il vento medio nell'intervallo tra le ore 6 e le ore 12 del giorno successivo ed il vento medio nell'intervallo tra le ore 12 e le ore 18 del giorno successivo – siglate come "<parametro>_domani").

I valori ottenuti sono stati discretizzati secondo il criterio riportato di seguito in queste tabelle.

TEMPERAT. MEDIA GG (°C)	
Molto basse	< -1
Basse	[-1 ; 7)
Medie	[7 ; 14)
Alte	[14 ; 21)
Molto alte	[21 ;25)
Altissime	>=25

PRECIP. TOT. GG (mm)	
Assenti	< 0.2
Deboli	[0.2 ; 5)
Moderate	[5 ; 15)
Forti	[15 ; 40)
Molto forti	>= 40

VENTO MEDIO GG (m/s)	
Debole	[0 ; 1)
Moderato	[1 ; 2)
Forte	[2 ; 4)
Molto Forte	>= 4

PRESS. ATM. MED. GG (mBar)	
Bassa	< 985
Media	[985 ; 990)
Alta	[990 ; 1000)
Molto alta	>= 1000

VENTO MEDIO h06-12 (m/s)	
Debole	[0 ; 1)
Moderato	[1 ; 2)
Forte	[2 ; 4)
Molto Forte	>= 4

VENTO MEDIO h12-18 (m/s)	
Debole	[0 ; 1.5)
Moderato	[1.5 ; 3)
Forte	[3 ; 5)
Molto Forte	>= 5

UMIDITA' REL. MEDIA GG (%)	
Secco	< 40
Poco umido	[40 ; 60)
Umido	[60 ; 80)
Molto umido	>= 80

RADIAZ. SOL. TOT. GG (MJ/mq)	
Bassa	< 4
Media	[4 ; 12)
Alta	[12 ; 24)
Molto alta	>= 24

VENTO PREVAL. h12-20 (m/s)	
Debole	[0 ; 1)
Moderato	[1 ; 3)
Forte	[3 ; 6)
Molto Forte	>= 6

VENTO PREVAL. H21-23 (m/s)	
Debole	[0 ; 1)
Moderato	[1 ; 2)
Forte	[2 ; 4)
Molto Forte	>= 4

N.B. Ognuna di queste 4 classi di vento prevalente può essere associata a 8 direzioni (N,NE,E,SE,S,SO,O,NO) per cui un vento moderato da E apparterrà alla classe "ModeratoE"

I dati sono stati quindi predisposti in formato ARFF-compatibile per il tool WEKA ed analizzati tramite l'algoritmo "Apriori". L'algoritmo Apriori deve essere inizializzato prima di essere eseguito con almeno i principali seguenti parametri:

- il tipo di regole da estrarre (tutte, oppure le sole regole associative riferite all'attributo classe delle istanze);
- il numero massimo di regole da estrarre;
- un limite di massimo supporto ammesso per la regola (default = 100%);
- un limite di minimo supporto ammesso per la regola (default = 1%);
- un limite di minima confidenza che deve avere una regola (default = 90%);

La parametrizzazione di questo algoritmo diventa particolarmente delicata per i dati meteorologici; porre dei limiti molto elevati per la confidenza significa cercare regole forti in un settore dove le eccezioni sono molto numerose, quindi per trovare regole forti dobbiamo abbassare il limite minimo di supporto e, di conseguenza, si trovano spesso regole poco interessanti.

Il vantaggio di rappresentare le conoscenze attraverso l'apparente semplicità dell'esposizione delle regole, risiede nel fatto che queste ultime, dedotte in modo assolutamente automatico, ci consentono di fornire uno strumento di analisi oggettivo e facilmente interpretabili anche dai non addetti al lavoro.

Proprio per testare le potenzialità dell'algoritmo, in questa prima fase di analisi sono state ricercate regole che abbiano un elevato livello di confidenza; per l'applicazione prototipale si cercherà invece di fare una scelta diversa; saranno cioè messe in evidenza le regole che, pur con minor grado di confidenza, sono compatibili con la giornata in esame.

Di seguito si riportano le prime 10 regole ottenute da un'elaborazione WEKA effettuata sui dati della stazione di Trento Sud in cui sono state ricercate quelle che abbiano una percentuale di confidenza > del 99%; per questo ho dovuto abbassare i limiti di supporto ammesso allo 0,8 %, il che significa che giornate in cui valgono regole di questo tipo si presentano mediamente 8 giorni ogni mille. L'elaborazione di questo tipo ha comportato la

generazione di circa 200 regole per la sola previsione del vento medio.

```
=== Run information ===
Scheme:       WEKA.associations.Apriori -N 1000 -T 0 -C 0.99 -D 0.0010 -U 1.0 -M 0.0080 -S -1.0 -A -c -1
Relation:     export_vento32
Instances:    2858
Attributes:   13
              mese
              media_ta
              tendenza_ta
              media_bar
              tendenza_bar
              media_rh
              tendenza_rh
              totale_rr
              media_vv
              cls_vento_prevalente_12_20
              cls_vento_prevalente_21_23
              totale_rad
              media_vv_domani
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.01 (23 instances)
Minimum metric <confidence>: 0.99
Number of cycles performed: 992
Best rules found:
 1. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante 44 ==> media_vv_domani=Debole 44    conf:(1)
 2. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante totale_rr=Assenti 43 ==> media_vv_domani=Debole 43    conf:(1)
 3. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante media_vv=Debole 42 ==> media_vv_domani=Debole 42    conf:(1)
 4. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante totale_rr=Assenti media_vv=Debole 41 ==> media_vv_domani=Debole 41    conf:(1)
 5. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante totale_rad=Media 38 ==> media_vv_domani=Debole 38    conf:(1)
 6. mese=gen media_bar=MoltoAlta tendenza_bar=Costante media_vv=Debole totale_rad=Media 38 ==> media_vv_domani=Debole 38    conf:(1)
 7. mese=dic media_bar=MoltoAlta cls_vento_prevalente_21_23=DeboleNE 37 ==> media_vv_domani=Debole 37    conf:(1)
 8. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante totale_rr=Assenti totale_rad=Media 37 ==> media_vv_domani=Debole 37    conf:(1)
 9. mese=gen media_bar=MoltoAlta tendenza_bar=Costante tendenza_rh=Costante media_vv=Debole totale_rad=Media 37 ==> media_vv_domani=Debole 37    conf:(1)
10. mese=gen media_bar=MoltoAlta tendenza_bar=Costante totale_rr=Assenti media_vv=Debole totale_rad=Media 37 ==> media_vv_domani=Debole 37    conf:(1)
```

Essendo l'algoritmo “Apriori” un algoritmo generico adatto a trovare regole associative tra gli attributi e non quindi un classificatore, è stato necessario predisporre un metodo oggettivo per valutare la sua credibilità.

Ho scelto quindi di verificare un periodo di dati (un anno circa) operando in questo modo:

(a) scelta di una giornata per la quale prevedere una determinata variabile (esempio il vento) e applicazione dei successivi punti b,c,d;

(b) generazione delle regole a partire dall'insieme delle istanze dalla quale ho preventivamente escluso i dati del giorno precedente a quello in esame in modo da non facilitare l'algoritmo (dalla giornata precedente infatti l'algoritmo estrarrebbe sicuramente almeno una regola perfettamente calzante con la

giornata da prevedere);

(c) con una procedura Java sviluppata ad hoc, ho filtrato tutte le regole generate automaticamente dall'algoritmo “Apriori” eliminando quelle non compatibili con la giornata da prevedere (a) (cioè ad esempio se voglio prevedere il tempo per il giorno G che presenta tra gli attributi del giorno precedente una temperatura dell'aria “temp_aria=media”, considero solamente le regole che non contemplano nelle premesse la temperatura dell'aria o che, se la contemplano, questa è “media”);

(d) con la stessa procedura Java ho scelto tra le regole compatibili quella che aveva il maggior indice di confidenza e, dalla conseguenza di questa regola, ho dedotto la previsione (se esistono più regole con la stessa confidenza, prendo quella che ha un maggior supporto).

Il punto (b) è quello più pesante da un punto di vista della computazione; rigenerare il set di regole da un archivio di 8 anni di una sola stazione, è un'operazione che, su un normale computer domestico, impiega qualche minuto, per questo ho analizzato in questo modo un solo anno di dati ottenendo i seguenti risultati.

TEMPERATURA MEDIA GIORNALIERA

	Molto basse	Basse	Medie	Alte	Molto alte	Altissime
Molto basse	2	7	0	0	0	0
Basse	5	78	14	0	0	0
Medie	0	18	70	9	0	0
Alte	0	0	7	81	12	3
Molto alte	0	0	0	15	20	7
Altissime	0	0	0	1	9	7

PRECIPITAZIONE TOTALE GIORNALIERA

	Assenti	Debole	Moderato	Forte	Molto Forte
Assenti	73	19	7	0	0
Debole	16	11	0	0	0
Moderato	10	5	1	0	0
Forte	2	2	2	0	0
Molto forte	1	2	0	0	0

VENTO MEDIO GIORNALIERO

	Debole	Moderato	Forte	Molto Forte
Debole	109	20	17	0
Moderato	42	26	33	1
Forte	18	23	37	1
Molto forte	19	4	33	0

Nella successiva tabella si riassume quindi la capacità predittiva dell'algoritmo basato sull'analisi delle regole.

	Istanze correttamente classificate	Kappa-statistic
Temperatura	67,1 %	0,62
Precipitazioni	56,3 %	0,13
Vento	44,9 %	0,21

Un'ulteriore possibilità è costituita dall'applicazione di algoritmi noti in letteratura e basati sull'analisi delle regole; tra questi merita attenzione l'algoritmo "**Decision table**"[14] che, nonostante la sua semplicità, fornisce dei risultati interessanti.

Nella sua forma elementare, data una lista di istanze classificate si scelgono quelle che trovano positivo riscontro con un'istanza da classificare (cioè che hanno attributi identici) interpretandole come regole e, tra questa lista, viene scelta la classe di appartenenza della maggioranza di queste istanze come classe dell'istanza da classificare.

Questo algoritmo è computazionalmente un poco più impegnativo nel momento in cui si intendesse implementarlo in un sistema real-time che deve prelevare i dati da un database, discretizzare un set di istanze numeriche per poi analizzare le regole e quindi applicare la decisione.

Per questo motivo è opportuno separare (come è stato fatto nel caso dell'applicazione dell'algoritmo Apriori) la fase di generazione delle regole che può essere fatta una tantum (ad

14 - Per approfondimenti: Ron Kohavi: The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, 1995.

esempio una volta al giorno), dalla fase di analisi delle stesse (fatta ogni volta a richiesta) in modo da pervenire ad un risultato speditivo (come è stato fatto con la procedura ad hoc).

Per completezza si riporta quindi anche la tabella di sintesi delle analisi svolte con l'algoritmo “Decision-table” implementato in WEKA.

Decision-table	**Istanze correttamente classificate**	**Kappa-statistic**
Temperatura	80%	0,74
Precipitazioni	71%	0,15
Vento	57%	0,37

2.3.3 - L'analisi delle giornate meteorologicamente vicine

Pur essendo disponibili in WEKA algoritmi di classificazione basati sull'analisi di prossimità e similarità (“Instance-based learning” capitolo 4.7 I.H.Witten e E. Frank [7]), si è ritenuto più opportuno implementare direttamente il semplice schema di analisi dei giorni vicini basato sul calcolo della distanza euclidea perché si ritiene possa essere di ausilio alla previsione per analogia.

Sì è ritenuto molto più interessante ricercare in archivio la/le giornata/e più vicine ad una presa in esame piuttosto che utilizzare l'analisi di prossimità per classificare le istanze in archivio.

Sono stati quindi utilizzati i valori numerici orari di ogni singola variabile meteorologica ed è stata realizzata una procedura che, a partire dai dati orari di una specifica giornata, ricercasse in archivio le 5 giornate meteorologicamente più vicine.

Prima però di decidere la metodologia con cui proporre all'utente il risultato dell'analisi è stato necessario valutare l'opportunità di confrontare due giorni di dati un parametro per volta o se confrontare due giornate considerando più variabili in modo congiunto.

In quest'ultimo caso è necessario procedere alla normalizzazione delle variabili come descritto nel capitolo 2.2.4.

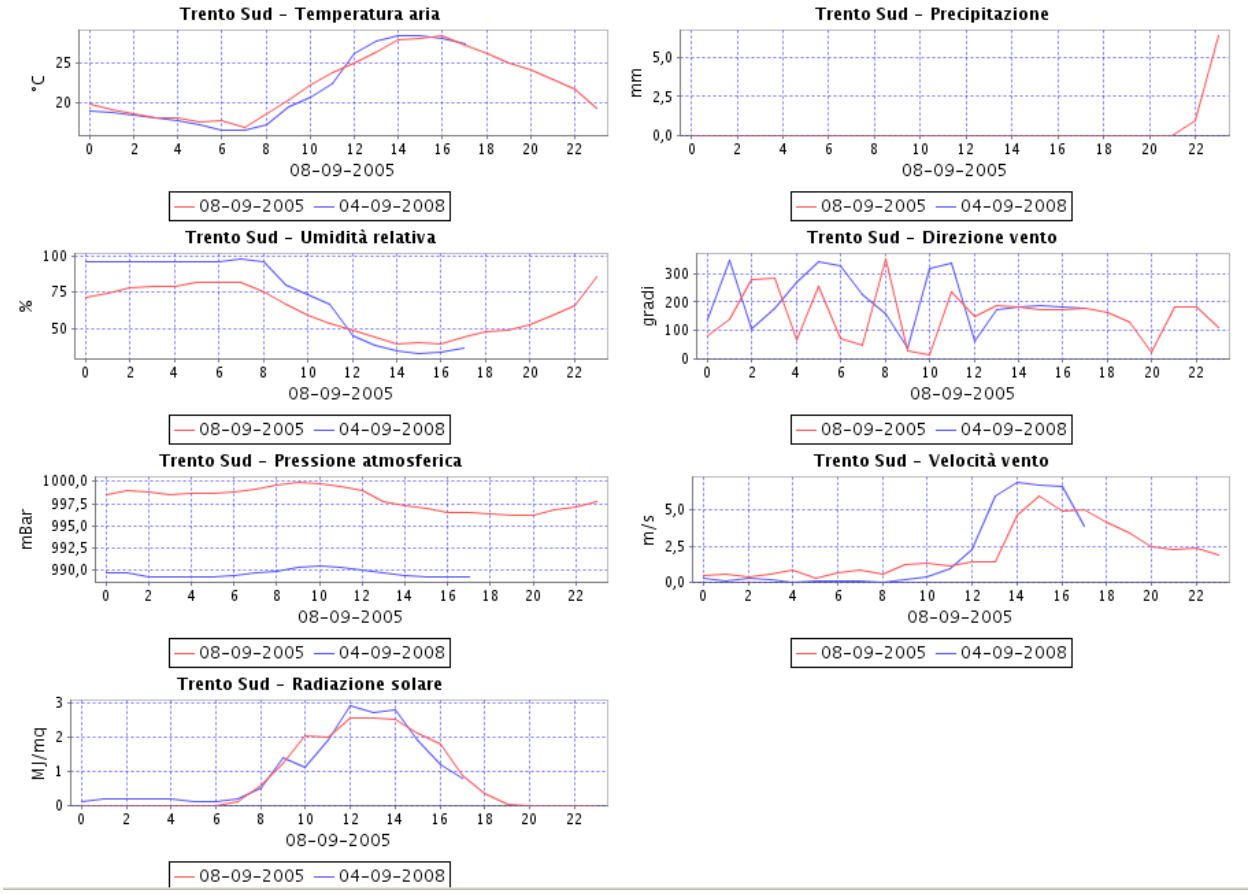


Figura 7: analisi di prossimità

Dalle prove fatte è risultato che solo in un numero limitato di casi l'analisi effettuata con l'apporto di tutte le variabili contribuisce a rintracciare in archivio giornate significativamente vicine. Otto anni di dati in archivio sono forse ancora un periodo temporalmente limitato, anche se la ricerca funziona correttamente, come si può vedere da quanto riportato in figura 7.

Nell'esempio si evidenzia infatti la capacità dell'algoritmo di valutare anche giornate non complete; in questo caso riguarda un'analisi svolta sulla stazione meteorologica di Trento Sud alle ore 17 del 4 settembre 2008 nella quale si evidenzia che nella giornata più vicina, l'8 settembre 2005, si sono poi verificate precipitazioni superiori a 5 mm di pioggia tra le ore 21 e le ore 24; tra le ore 18 e le ore 20 del 4 settembre 2008 (giorno in esame) si sono puntualmente verificate precipitazioni per un totale di 2 mm, confermando, perlomeno in questo esempio, una discreta capacità di supporto all'analisi predittiva per analogia.

Utilizzando invece la ricerca come singolo parametro si ottiene molto spesso l'accesso a giornate in cui c'è una forte analogia tra l'andamento temporale della variabile meteorologica nei due giorni in confronto, vedi l'esempio qui sotto riportato (figura 8).

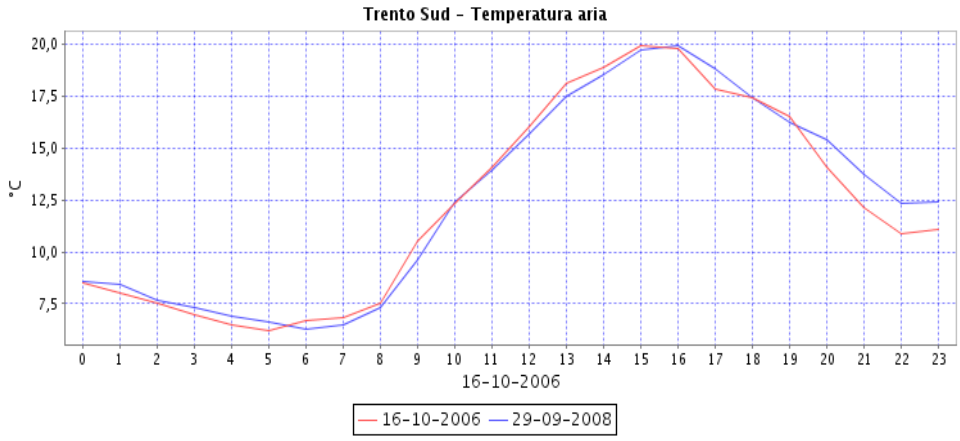


Figura 8: due giornate "vicine" per temperatura dell'aria

In questo caso l'elaborazione effettuata sui dati del 29 settembre 2008, mette in evidenza che, rispetto alla temperatura dell'aria, la giornata era più simile ad una di ottobre (16 ottobre 2006). In effetti nel mese di settembre 2008 si sono registrate temperature un po' più basse rispetto alla media.

Un'analisi di questo tipo, effettuata però su un periodo temporale più lungo (più giorni), potrebbe ad esempio essere di ausilio per valutare l'andamento della stagione.

2.3.4 - Predizione di valori numerici attraverso la correlazione

L'analisi di regressione è un algoritmo implementato in WEKA ed ha delle ottime performance in termini di velocità di esecuzione. Per questo ho pensato di utilizzarlo per la previsione a brevissimo termine laddove la velocità di esecuzione diventa importante.

Come attributi (numerici) dell'algoritmo sono stati utilizzati tutti i valori di tutte le variabili meteorologiche nelle 6 ore precedenti a quella in esame. Si tratta cioè di utilizzare 42 attributi (6 ore x 7 variabili) e classificarli per 3 volte ad ognuna delle variabili che si intende prevedere: ad esempio per la previsione di temperatura applico l'algoritmo la prima volta per

prevedere la temperatura al tempo t1 = t0 + 1h, la seconda volta per prevedere la temperatura al tempo t2 = t0 + 2h, la terza volta per prevedere la temperatura al tempo t3 = t0 + 3h dove t0 è l'istante in cui ho la disponibilità dell'ultimo dato in archivio.

Con questo tipo di previsione si ottengono dei risultati abbastanza disomogenei in relazione al tipo di variabile; di seguito si riassumono gli indici medi dei valori di credibilità ottenuti dall'applicazione dell'analisi della correlazione.

	t1		t2		t3	
	Coefficiente di correlazione	Errore medio assoluto	Coefficiente di correlazione	Errore medio assoluto	Coefficiente di correlazione	Errore medio assoluto
temperatura media oraria	0,99	0,6 °C	0,99	1,4 °C	0,98	1,9 °C
umidità relativa media oraria	0,97	3,5%	0,91	6,5%	0,86	8,8%
pioggia totale oraria	0,73	0,5 mm	0,68	0,5 mm	0,62	0,5 mm
radiazione solare globale	0,88	0,07 MJ/mq	0,82	0,27 MJ/mq	0,81	0,45 MJ/mq
velocità vento media oraria	0,81	0,5 m/s	0,67	0,6 m/s	0,58	0,7 m/s
Press. Atmosf. media oraria	0,98	1,5 mBar	0,97	1,7 mBar	0,97	1,9 mBar

Dalla tabella possiamo notare in particolare un'elevata correlazione per la previsione di temperatura, umidità e pressione atmosferica.

2.3.5 - La classificazione statistica

Il classificatore che è stato maggiormente utilizzato in tutte le nuove sperimentazioni fatte con WEKA è il "Naive Bayes", un classificatore molto performante in termini di velocità di esecuzione che ci consente di avere un'idea immediata delle potenzialità predittive delle istanze in esame.

Le stesse istanze di dati di Trento Sud, discretizzate nello stesso modo come quelle utilizzate per l'analisi delle regole, sono state quindi utilizzate con questo algoritmo ottenendo i seguenti risultati per l'analisi delle temperature, delle precipitazioni e del vento .

TEMPERATURA MEDIA GIORNALIERA

	Molto basse	Basse	Medie	Alte	Molto alte	Altissime
Molto basse	71	46	0	0	0	1
Basse	40	611	79	1	0	1
Medie	0	84	516	77	0	0
Alte	0	0	101	554	116	4
Molto alte	0	0	0	80	271	60
Altissime	0	0	0	1	46	108

PRECIPITAZIONE TOTALE GIORNALIERA

	Assenti	Debole	Moderato	Forte	Molto Forte
Assenti	1755	176	19	8	36
Debole	263	143	15	11	40
Moderato	125	62	8	7	21
Forte	67	44	11	3	18
Molto forte	6	9	3	0	18

VENTO MEDIO GIORNALIERO

	Debole	Moderato	Forte	Molto Forte
Debole	714	137	35	21
Moderato	257	356	335	39
Forte	89	185	575	50
Molto forte	9	10	25	31

Si riassumono quindi in una tabella sintetica le capacità predittive dell'algoritmo Naive Bayes.

	Istanze correttamente classificate	Kappa-statistic
Temperatura	74,3 %	0,67
Precipitazioni	67,1%	0,24
Vento	58,4 %	0,29

2.3.6 - Una comparazione tra i vari modelli

Per un confronto si riporta un'analisi sintetica dei risultati dei vari metodi applicati. Si riporta anche l'analisi svolta sull'output grezzo delle previsioni del modello fisico-matematico di Reading dedotta da un insieme di istanze di 789 giorni a partire dall'1 gennaio 2007.

Per la comparazione delle elaborazioni su scala locale e delle elaborazioni del modello-fisico matematico dell'ECMWF (European Centre for Medium Range Weather Forecats) si deve però premettere questa considerazione: la comparazione su base giornaliera dei dati è l'unica che si poteva svolgere allo stato attuale, visto e considerato che le elaborazioni locali sono state fatte su base giornaliera. Quindi anche i dati del modello fisico-matematico sono stati aggregati su base giornaliera anche se in realtà il modello elabora delle previsioni molto più dettagliate nel tempo (step temporali disponibili su base almeno tri-oraria) e fornisce inoltre la predizione di un elevato numero di altre variabili.

Non deve quindi nemmeno sfioraci l'idea che le elaborazioni locali possano sostituire quelle del modello. Un modello locale semmai potrebbe essere utile per disporre di analisi speditive e complementari disponibili in qualsiasi momento della giornata, visto e considerato che, l'output del modello fisico-matematico è disponibile dopo circa 8 ore dalla sua elaborazione e che il modello esegue per due sole volte nell'arco della giornata: alle ore 00 ed alle ore 12.

	BAYES		DECISION-TABLE		READING	
	Istanze correttamente classificate	**Kappa-statistic**	**Istanze correttamente classificate**	**Kappa-statistic**	**Istanze correttamente classificate**	**Kappa-statistic**
Temperatura	74,3 %	0,67	80%	0,74	//	//
Precipitazioni	67,1%	0,24	71%	0,15	59,4 %	0,31
Velocità vento	58,4 %	0,29	57%	0,37	38,5 %	0,14

***Nota**: in questa tabella è stato escluso il confronto con il modello di Reading per le temperature poiché necessiterebbe di un'attività di correzione dell'output diretto del modello per ricondurre la previsione alla quota di analisi, molto diversa dalla quota del modello.*

Da questa comparazione si evidenzia una peculiarità dei modelli basati sull'analisi meteorologica locale: i modelli locali prevedono con maggior credibilità la velocità media giornaliera del vento. Un risultato sicuramente non sorprendente, ma che conferma una seppur minima capacità predittiva.

Ovviamente anche in questo caso non si cerca di mettere in discussione il modello di Reading che, per quanto riguarda il vento, ci fornisce un dettaglio temporale maggiore ed anche la variabile “direzione del vento”, ma si intende sostenere il fatto che con i semplici dati forniti da una sola stazione meteorologica e senza l'applicazione di complessi calcoli di tipo fisico-

matematico si riescono ad ottenere dei risultati significativi che, qualora fossero integrati anche con altre informazioni, possono essere di ausilio alla previsione locale.

Un modello di questo tipo, basato anche sul contributo delle analisi automatiche delle informazioni locali, può quindi fornire un supporto continuativo e speditivo che sia complementare agli altri strumenti di analisi.

2.4 - La rappresentazione delle conoscenze

Ritenendo che gli algoritmi sopra proposti possano essere utili, questi devono poter essere anche agevolmente utilizzabili; il modo quindi come si elaborano e si espongono i risultati determina anche l'usabilità del sistema. Nel nostro caso specifico, il brevissimo intervallo di tempo che intercorre tra l'arrivo dei dati, che giungono in tempo quasi reale, e l'analisi, che deve essere fatta nel più breve tempo possibile per essere utile, impone di progettare un sistema con un elevato grado di "usabilità". Per questo la modalità di rappresentazione delle conoscenze assume un grande rilievo.

2.4.1 - La rappresentazione delle regole

Per rappresentare le regole si propone una rappresentazione tabellare secondo la figura 9.

Regole delle temperature compatibili per il 25-05-2008.

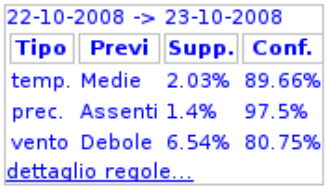
premesse												conseguenza		
mese	media temp.aria	tendenza temp.aria	media press.	tendenza press.	media umidità rel.	tendenza umidità	media vento GG	totale precip. GG	vento preval. h.12-h.20	vento preval. h.21-h.23	totale radiaz. GG	temp. previste giorno succ.	supp. %	conf. %
mag	Alte	*	*	*	*	*	*	Deboli	*	*	*	Alte	1.63	95.5
mag	Alte	*	*	*	*	*	*	*	ForteS	*	*	Alte	1.48	95.0
mag	Alte	*	*	*	*	*	*	Deboli	*	*	Alta	Alte	1.22	93.9
mag	Alte	*	Alta	*	*	*	*	*	ForteS	*	*	Alte	1.15	93.5
mag	Alte	*	Alta	*	*	*	*	Deboli	*	*	*	Alte	1.11	93.3
mag	Alte	*	*	Aumento	*	*	*	*	*	*	*	Alte	1.48	92.5
mag	Alte	*	*	*	*	Costante	*	*	*	*	*	Alte	2.63	90.1

Figura 9: Rappresentazione tabellare delle regole

L'intestazione della tabella riporta il nome degli attributi delle premesse (in blu) ed il nome dell'attributo della conseguenza (in rosso) assieme all'indicazione delle colonne utilizzate

per gli indici di supporto e confidenza. Ogni riga rappresenta una regola; ogni cella riporta il valore assunto dall'attributo (il cui nome è indicato nell'intestazione di colonna) che interviene nella regola, un "*" indica che la regola vale in qualsiasi condizione si trovi l'attributo indicato nell'intestazione di colonna.

Queste regole dovranno essere filtrate in modo tale da consentire di analizzare solo le regole compatibili con una certa giornata in esame, inoltre, anche dopo questo filtro, le regole possono essere ancora moltissime, per cui è utile pervenire alla tabella di figura 9 solo dopo che l'utente intenda effettuare un'analisi più approfondita delle regole.


22-10-2008 -> 23-10-2008

Tipo	Previ	Supp.	Conf.
temp.	Medie	2.03%	89.66%
prec.	Assenti	1.4%	97.5%
vento	Debole	6.54%	80.75%

dettaglio regole...

Figura 10: rappresentazione sintetica delle regole.

In prima battuta è meglio quindi esporre all'utente una sintesi delle regole ottenuta recuperando tra tutte le regole compatibili con la giornata in esame solo quella che ha il maggior indice di confidenza e, se ci sono più regole con lo stesso indice di confidenza, quella che ha il maggior supporto.

Si riporta quindi in figura 10 una modalità di rappresentazione sintetica delle regole con maggior confidenza applicabili ad una specifica giornata. Da qui sarà poi possibile accedere all'elenco dettagliato delle regole applicabili.

Nella rappresentazione si espone la data di elaborazione della regola e la data per cui la regola è applicabile; quindi una esposizione degli indici di supporto e confidenza per le tre regole applicabili: temperature, precipitazioni e vento.

2.4.2 - La rappresentazione delle giornate meteorologicamente vicine

Per rappresentare le giornate meteorologicamente vicine, il metodo è quello di visualizzare contemporaneamente su uno stesso grafico cartesiano sia la serie temporale della giornata in esame, sia la serie temporale della giornata vicina così come riportato di seguito in figura 11.

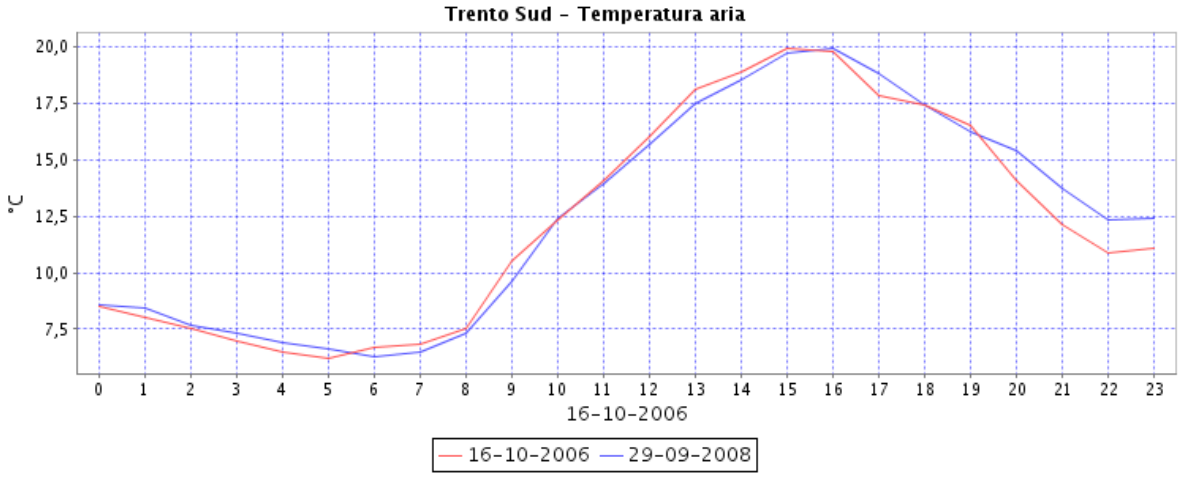


Figura 11: Rappresentazione cartesiana della giornata in esame (in rosso) e della giornata più simile (in blu) .

Anche in questo caso, prima di pervenire a questa rappresentazione, l'utente dovrà poter consultare un elenco delle giornate più vicine da cui attingere in base ad un indice di prossimità (la distanza euclidea); si dovrà quindi esporre, analogamente a quanto fatto per le regole, una tabella da cui scegliere la giornata vicina desiderata come riportato in figura 12.

data	vicino	dist.
03-11-2008	03-11-2008	0.0
03-11-2008	27-10-2008	2.3
03-11-2008	05-10-2005	4.6
03-11-2008	27-03-2006	4.8
03-11-2008	11-04-2008	5.0

Figura 12: Scelta di uno tra i 5 giorni più vicini.

2.4.3 - La rappresentazione delle previsioni a brevissima scadenza

Anche per la previsione a brevissimo termine dell'andamento di una variabile meteorologica ci si serve di una rappresentazione cartesiana.

Tale tipo di visualizzazione ha significato solo nella visualizzazione dei dati in tempo reale, cioè nella giornata in corso; per questo si propone una rappresentazione delle registrazioni orarie con una linea di una colorazione (rossa), subito seguita dall'andamento previsto con il

metodo regressione multilineare che viene rappresentato con una colorazione diversa (blu).

Un esempio è riportato nella seguente immagine di figura 13.

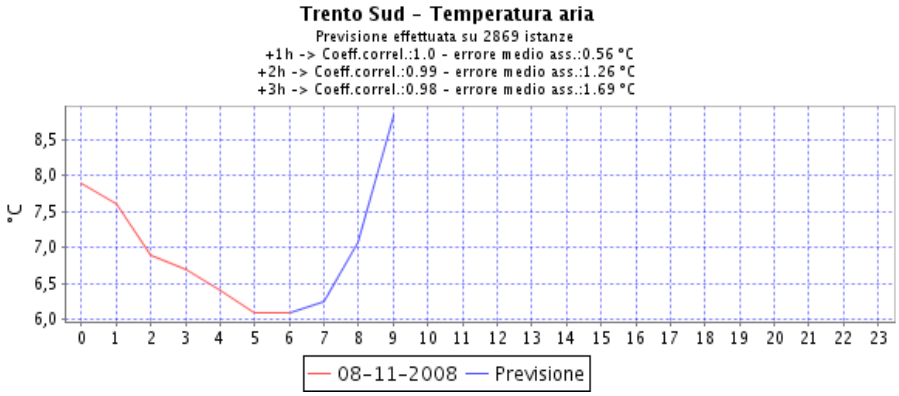


Figura 13: Rappresentazione della previsione fino a + 3 ore di una variabile meteorologica

2.4.4 - La rappresentazione integrata delle analisi

Per estrarre in modo rapido e flessibile i pattern e le analogie dei dati, si intendono utilizzare delle metodologie di rappresentazione tipicamente reperibili negli applicativi di business intelligence (BI). Uno degli scopi del presente lavoro è per l'appunto anche quello di proporre una modalità di rappresentazione dei dati meteorologici sullo stile dei prodotti di business intelligence finalizzando quindi il progetto software alla possibilità di un'analisi rapida e flessibile dei dati in situazioni di previsione a brevissima scadenza quando il tempo a disposizione dell'analisita è limitato.

Nei capitoli successivi si esporrà quindi la progettazione e l'implementazione di un prototipo software in grado di supportare le analisi meteorologiche locali.

Capitolo 3 - La progettazione di un sistema di analisi meteo locale

Nel presente capitolo si espone la progettazione di un prototipo di sistema integrato per l'analisi dei dati meteorologici riportando i requisiti che un tale sistema deve avere e la progettazione di alto livello che sovrintende le successive fasi di analisi più dettagliata ed implementazione.

Dalle premesse fatte si è ipotizzato che il sistema possa essere un primo passo verso una futura integrazione con altre sorgenti di informazioni meteorologiche locali, cioè di un data warehouse per l'analisi, riferibili ad un territorio a scala regionale secondo un schema come quello di figura 14.

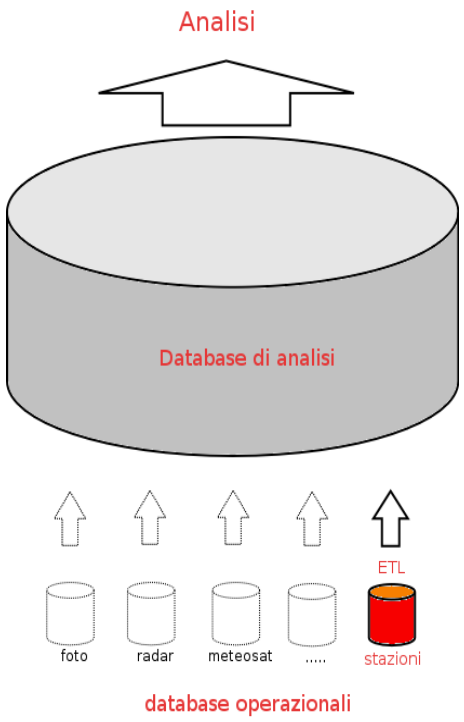


Figura 14: schema di integrazione delle fonti

In questa prospettiva possiamo pensare di progettare un prototipo di "data mart"[15], cioè una parte di un futuro data warehouse che verrebbe costruito con una strategia di tipo bottom-up, cioè partendo dal basso.

L'approccio bottom-up è anche consigliato perchè comporta generalmente meno rischi di insuccesso o, se l'insuccesso è il risultato finale, si evitano inutili sprechi di risorse.

Nella progettazione del sistema, dobbiamo sempre tener presenti quattro punti di vista fondamentali:

- una vista top down, che permette di selezionare le informazioni rilevanti da analizzare;
- una vista delle sorgenti dei dati, cioè un'analisi attraverso i diagrammi entità-relazione dei dati di partenza;
- una vista secondo lo schema del data ware house, che mette in rilievo i fatti centrali delle

15 - Un data mart è un sottoinsieme del data warehouse che soddisfa una specifica esigenza di analisi più limitata.

tabelle, le dimensioni di analisi, le sommarizzazioni da precalcolare, il periodo temporale da mantenere;

- una vista sulle query da restituire, cioè il punto di vista dell'utente finale che spesso è scarsamente interessato ai dettagli tecnici di implementazione.

Si procede quindi con le fasi di analisi dei requisiti, disegno del sistema, implementazione, test e, infine, il rilascio.

Come processo di sviluppo è opportuno ipotizzare un approccio a spirale, preferibile ad un classico approccio a cascata, perchè permette tempi più ristretti di feedback dagli utenti finali. In questo specifico caso il primo risultato a cui si perviene è un prototipo di applicativo software “aperto”, nel senso che non si tratta di un prototipo usa e getta, ma di un prototipo che si trova in una fase di primo stadio evolutivo del prodotto finale.

Rispetto al processo di sviluppo a cascata, quello a spirale svolge le attività classiche di analisi, disegno, implementazione, test e rilascio in tempi brevi cercando di effettuare di volta in volta piccoli passi ed avere immediatamente un feedback dal cliente per far emergere prima possibile i frequenti problemi derivanti da incomprensione tra esperti di dominio e sviluppatori.

In questo capitolo si descriverà la progettazione del sistema mentre nel capitolo successivo si svilupperà invece la progettazione di dettaglio, delle procedure di estrazione, trasformazione dei dati dal/dai database operazionali e quelle di caricamento degli stessi nel sistema per finire con le procedure di estrazione dei data dal database di analisi finalizzate alla presentazione dei dati all'utente.

3.1 - I requisiti del sistema

Gli attori del sistema sono fondamentalmente due: il previsore meteorologico regionale titolato ad effettuare analisi sui dati e l'attore “nascosto” che possiamo chiamare ETL, cioè l'insieme dei programmi di estrazione, trasformazione e caricamento che, attraverso procedure schedulate o ad evento alimentano continuamente il database di analisi. Non si esclude peraltro che si possa configurare un altro attore: il validatore che interviene con funzionalità di editing dei dati per correggere manualmente errori che non è possibile individuare automaticamente.

Il risultato a cui si intende arrivare è un sistema integrato adatto a supportare l'analisi meteorologica locale. In sostanza un'interfaccia utente sull'esempio dei prodotti di “business intelligence” adatta ad analizzare in modo rapido e flessibile una vasta mole di dati.

3.1.2 - Requisiti funzionali

Il sistema viene progettato a partire dai seguenti requisiti funzionali:

I. **omogeneità di gestione e rappresentazione:** i dati dovranno essere gestiti e rappresentati in modo omogeneo rispetto alla tipologia di stazione;

II. **scelta della stazione**: il sistema dovrà consentire di scegliere una stazione meteorologica attraverso un elenco oppure da una mappa regionale che ne faciliti la localizzazione territoriale; ai fini dell'analisi, i soli attributi importanti della stazione meteorologica locale sono il nome della località, la quota, le coordinate geografiche;

III. **scelta di una variabile meteorologica**: il sistema dovrà consentire di scegliere una variabile meteorologica principale su cui effettuare l'analisi e per la quale dovrà essere possibile visualizzarne sia l'andamento temporale su un grafico cartesiano sia la registrazione numerica; inoltre dovrà essere possibile anche un'agevole rappresentazione contemporanea di una seconda variabile meteorologica per confrontarla con quella principale;

IV. **scelta della data di analisi**: il sistema dovrà garantire l'accesso immediato alla giornata odierna, quindi dovrà essere possibile spostarsi temporalmente in modo agevole, ed aggregare i dati su periodi temporali diversi (giorno, settimana, mese, anno);

V. **accesso agli strumenti di estrazione di conoscenza**: dovrà essere agevolato l'accesso alle similarità degli andamenti delle variabili meteorologiche, l'accesso alle regole applicabili alla giornata in esame, l'accesso alla previsione dell'andamento fino a più tre ore;

VI. **statistiche**: dovranno essere rappresentate informazioni statistiche basate su tutti gli anni in archivio delle variabili vento, temperatura, precipitazione;

3.1.2 - Requisiti non funzionali

Tra i requisiti non funzionali meritevoli di attenzione, c'è la scalabilità del sistema. Il sistema, dovrà essere costruito per moduli tali da separare le procedure di accesso ai dati dalla logica di elaborazione e da quella di rappresentazione.

Per le potenzialità di elaborazione, l'elemento centrale del sistema è il database che supporta tutto il carico di computazione delle aggregazioni sui dati e che deve pertanto essere robusto, affidabile ed anch'esso scalabile.

Il sistema dovrà essere orientato ad un'accesso alle analisi attraverso un web browser.

3.2 - Progettazione concettuale, logica e fisica

Per comprendere le attività da svolgere dobbiamo analizzare i dati dei possibili database operativi delle stazioni meteorologiche che potranno fornire i dati al sistema di analisi. Da questo punto di vista esistono due principali tipologie di oggetti con cui ci si confronta: un'anagrafica delle stazioni ed un insieme di registrazioni di parametri meteorologici riferiti ad una certa località indicata appunto in questa anagrafica.

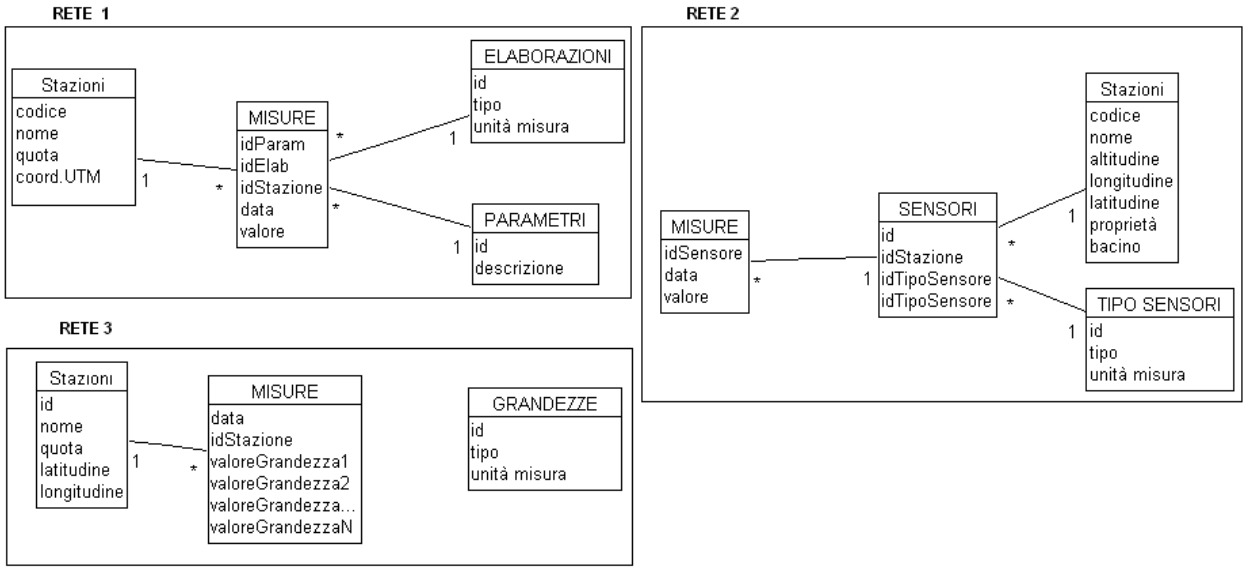


Figura 15: Schema concettuale delle fonti

Lo schema concettuale delle sorgenti di figura 15, conferma l'estrema disomogeneità

delle fonti che richiede quindi un'attività di semplificazione e riconciliazione delle fonti. Anche considerando le frequenze di acquisizione delle misure, alcune elaborazioni vengono fatte su base oraria, altre con una frequenza sub-oraria fino al limite di 5 minuti per alcune tipologie di parametri.

Analizzando il sistema da un punto di vista delle analisi che si intendono svolgere, occorre quindi pensare ai fatti, alle dimensioni di analisi ed alle informazioni rilevanti da gestire.

I “fatti” principali (o concetti) da analizzare sono: l'evoluzione temporale della temperatura dell'aria, del vento e delle precipitazioni di una rete di stazioni meteorologiche. Le dimensioni di analisi sono il tempo e lo spazio. Lo spazio è identificato dagli attributi della stazione (quota, coordinate), che hanno limiti ben definiti. Per il tempo visto che si tratta di una dimensione di analisi continuamente crescente (quindi potenzialmente illimitata) e considerato che una misura meteorologica non è tale senza una marca temporale, si ritiene di doverlo gestire nella stesso oggetto che descrive le misure dei fatti anziché nelle tabelle che esprimono la dimensione di analisi. Lo schema concettuale risulta quindi notevolmente semplificato.

Per un sistema ottimizzato per l'analisi che sia omogeneo, si intende mantenere quindi in archivio il dato orario delle variabili meteorologiche principali ed alcune limitate informazioni anagrafiche della stazione, ottenendo il seguente schema concettuale:

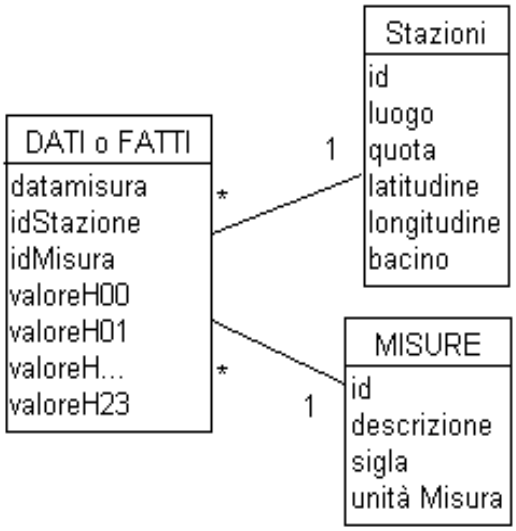


Figura 16: Schema concettuale database di analisi

Nell'oggetto che serve per analizzare i “fatti” vengono memorizzate tutte le misurazioni orarie delle variabili meteorologiche principali insieme con la marca temporale, con un indice che descrive il tipo di misura e con l'identificativo della stazione a cui si riferiscono.

Dalle misurazioni orarie si possono poi dedurre variabili derivate e/o aggregate tipo la tendenza, la medie o il totale da utilizzare per la costruzione delle istanze giornaliere di addestramento.

L'implementazione del sistema si appoggia ad un modello logico dei dati di tipo relazionale, per la cui implementazione fisica si è scelto di appoggiarsi al database PostgreSQL[16], un robusto database che offre diversi vantaggi tra i quali, per questa specifica applicazione, possono essere annoverati la caratteristica di essere orientato agli oggetti, per un possibile futuro utilizzo con applicazioni G.I.S. (Geographic Information System) e la disponibilità di un linguaggio nativo PL/psSQL che rende più performante l'utilizzo intensivo delle query.

La logica con cui è stata implementata la base di dati, ricalca il modello concettuale: per i tre oggetti dello schema concettuale sono state definite le tre relazioni principali, dati, misure, stazioni.

- DATI(datamisura, idstazione, idmisura,h00,h01,...,h23);
- MISURE(id,descrizione,unitàMisura,sigla);
- STAZIONI (id, luogo,quota, latitudine, longitudine, bacino);

Le query più complesse che il sistema deve restituire, riguardano le elaborazioni per la predisposizione delle istanze di addestramento da utilizzare negli algoritmi di apprendimento “Apriori”, “Decision Table”, “Analisi di regressione”.

Fornirò un esempio di come viene restituita dal database la vista che serve a rendere disponibili agli algoritmi di apprendimento le istanze per la generazione delle regole del vento. Per le altre istanze il metodo utilizzato è molto simile.

In questo caso si tratta di restituire un insieme di istanze discretizzate su base

16 - PostgreSQL è un potente database relazionale OpenSource di fascia enterprise: vedere http://www.postgresql.org/

giornaliera secondo una relazione del tipo:

- ISTANZE_GIORNALIERE_VENTO (mese, media_ta, tendenza_ta, media_bar, tendenza_bar, media_rh, tendenza_rh, totale_rr, media_vv, cls_vento_preval_12_20, cls_vento_preval_21_23, totale_rad, media_vv_domani);

Gli attributi “media_<x>” (media giornaliera della variabile x), tendenza_<x> (tendenza giornaliera della variabile x) e “cls_vento_preval_<x_y>” (classe del vento prevalente tra le ore x e le ore y), che devono essere attributi discreti, sono ottenuti a partire dai valori orari delle variabili meteorologiche tipicamente espressi invece in valori decimali; quindi, sono state implementate direttamente delle funzioni in linguaggio nativo PL/psSQL (vedi un esempio sotto riportato) che restituiscono la classe numerica di appartenenza la quale a sua volta, attraverso una relazione con la rispettiva tabella di decodifica, restituisce il valore discreto della classe di appartenenza.

```
CREATE OR REPLACE FUNCTION cls_rad(n real)
  RETURNS smallint AS
$BODY$
-- ritorna la classe di appartenenza della radiazione solare totale giornaliera
DECLARE
classe smallint ;
BEGIN
   If (n < 4) Then
      classe = 1; -- bassa
   elsif (n >= 4 and n < 12) Then
      classe = 2; -- media
   elsif (n >= 12 and n < 24) Then
      classe = 3; -- alta
   elsif (n >= 24) Then
      classe = 4; -- molto alta
   else
      -- se non è classificabile assegno valore medio
      classe=3;
   end if;
  RETURN classe;
END;
$BODY$
LANGUAGE 'plpgsql' STABLE
```

Per arrivare a costruire l'intera istanza del tipo cercato per il vento è necessario

preventivamente definire delle viste di appoggio per ognuna delle variabili meteorologiche; ad esempio per estrarre la classe di appartenenza della tendenza della temperatura giornaliera dell'aria utilizzo la seguente interrogazione (espressa con un formalismo pseudo relational-algebra):

$$\text{tendenza\_ta} = [\text{CLASSE\_TEND\_TA}] \bowtie \mathit{clsTendTA}\ (\mathit{tendenza}(\sigma_{\text{idMisura=1 ^ idStazione =32}}\ [\text{DATI}]))$$

cioè eseguo una selezione sulla tabella dati per estrarre solo i dati delle misure di temperatura e della stazione di interesse ed applico successivamente le funzioni implementate con il linguaggio PL/psSQL per il calcolo della tendenza (coefficiente angolare della retta di regressione) e della classe di appartenenza (da un numero decimale ritorna un'intero), quindi eseguo un join con la tabella delle classi di tendenza della temperatura per ottenere il valore discreto.

Predispongo poi ulteriori viste di appoggio per ottenere allo stesso modo tutti gli altri attributi discreti ed unisco le viste con un operazione di join finale che le mette in relazione attraverso la data della misura. Analoga cosa faccio per ottenere la discretizzazione della classe di appartenenza dell'istanza di addestramento, tenendo presente che devo ottenere però una vista dove la data da utilizzare per il join è quella del giorno precedente dal momento che voglio prevedere cosa farà il giorno successivo.

Le viste che restituiscono istanze così create sia per la previsione del vento che della temperatura media che della pioggia, saranno poi utilizzate dalle procedure applicative che sfruttano gli algoritmi di apprendimento.

3.3 - Componenti del sistema

Il sistema organizzato con una modularità distinta in tre livelli: livello di accesso ai dati, livello di controllo/elaborazione e livello di presentazione.

Lo schema dei componenti è quello qui sotto riportato:

- il livello di accesso ai dati si distingue a sua volta in accesso al database di

analisi gestito da una classe specializzata e accesso ai vari database sorgente (per ora questa parte sarà implementata solo per una singola tipologia di stazioni, ma è opportuno che l'architettura finale sia prevista anche per altre fonti come indicato);

- il livello di controllo si interfaccia con la libreria WEKA, con la libreria JFreeChart (per la generazione di rappresentazioni di serie temporali) e con le classi di interfacciamento al database di analisi e restituisce al livello di presentazione i dati in formato XML o sotto forma di grafici di serie temporali in formato PNG;

- il livello di presentazione consiste principalmente in interfacce html, jsp e codice javascript gestite da un servlet container (in questo caso Apache Tomcat).

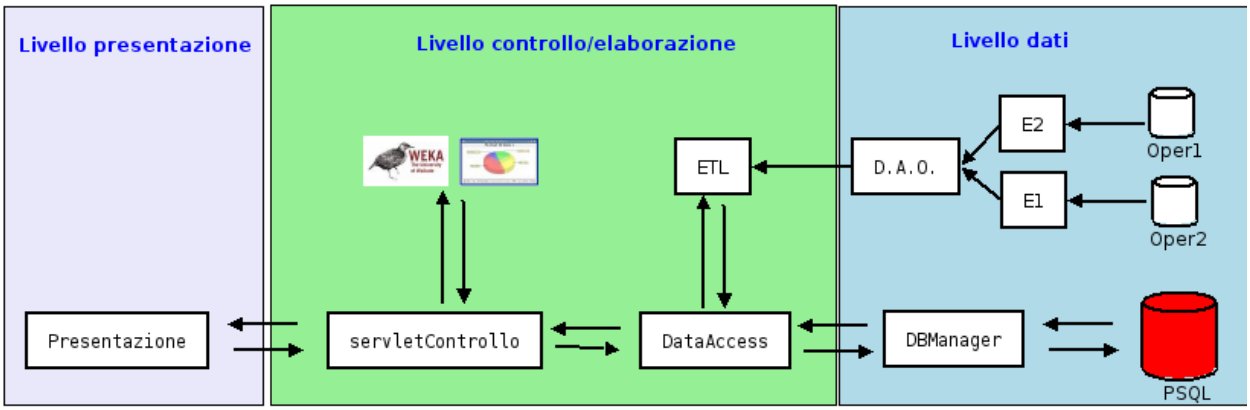


Figura 17: Schema dei componenti

Lo schema di figura 17 rappresenta la progettazione dei componenti ad alto livello; ogni componente sarà costituito da un insieme di classi specializzate in varie operazioni coerenti per il livello di appartenenza e per funzionalità del componente.

Nello sviluppo del prototipo applicativo, viene privilegiata l'attività di realizzazione del flusso di presentazione dei risultati all'utente, mentre le attività ETL saranno semplificate e limitate al solo caricamento in modalità batch dei dati nel database di analisi dopo averli prelevati da una sorgente.

Capitolo 4 - L'architettura

Per definire l'architettura del sistema è opportuno circoscrivere con maggior dettaglio le modalità di interazione dell'utente con il sistema, lo stato in cui si può trovare il sistema e le collaborazioni tra le classi software che interagiscono nel sistema. Per questo si analizzeranno nei prossimi sotto capitoli i casi d'uso, gli stati e le classi del sistema.

4.1 - I casi d'uso

A partire dai requisiti delineati nel precedente capitolo, si specifica con maggior dettaglio i casi d'uso del sistema, in modo da delineare con un approfondimento maggiore le possibili interazioni con il sistema.

Si identificano quindi i seguenti casi d'uso:

UC1 – lo scheduler di sistema attiva il caricamento dei dati			
Requisiti	Input	Output	Descrizione
I	-	File di log delle attività schedulate	Ad intervalli fissi (es. ogni ora) lo scheduler di sistema attiva una procedura che interroga le fonti, preleva i dati dalle fonti, li trasforma nel modo adatto per il database di analisi e li inserisce nel database omogeneizzando in tal modo la modalità di gestione e analisi dei dati.

UC2 – lo scheduler di sistema si attiva per la rigenerazione delle regole			
Requisiti	Input	Output	Descrizione
I	-	1 file di testo per ogni stazione contenente le regole costruite dall'algoritmo “Apriori”	La rigenerazione delle regole è un'operazione che richiede un discreto tempo di computazione e non ha significato farlo ad ogni richiesta dell'utente, anche perchè sono regole giornaliere. All'attivazione del timer di sistema (ogni giorno nelle ore notturne ad esempio), viene lanciata l'esecuzione dell'algoritmo “Apriori” che rigenera l'insieme delle regole.

UC3 – il previsore sceglie una nuova stazione			
Requisiti	Input	Output	Descrizione
II	Codice della stazione (data analisi, sigla della prima variabile da analizzare, sigla seconda variabile per confronto, tipo di aggregazione temporale vengono invece dedotte dallo stato del sistema)	Il sistema aggiorna i grafici e le tabelle di sintesi nell'interfaccia utente in modo coerente con i dati di input.	Il previsore sceglie una nuova stazione attraverso un elenco di stazioni disponibili (cosicché l'attore può individuarla per attributi nominali) oppure attraverso una mappa geografica (cosicché l'attore può meglio individuarla come localizzazione geografica); una volta ottenuto il codice stazione, il sistema dovrà: aggiornare la parte statistica dell'interfaccia (qualora necessario); aggiornare gli strumenti di accesso alle conoscenze acquisite (vedi UC6, UC7); aggiornare i grafici delle serie temporali della variabile meteorologica principale e secondaria (vedi anche caso d'uso UC4) secondo anche la data attuale di analisi (vedi anche caso d'uso UC5) e del tipo di aggregazione temporale;

UC4 – il previsore sceglie una nuova variabile meteorologica			
Requisiti	Input	Output	Descrizione
III	Sigla della prima variabile meteorologica o sigla della seconda variabile meteorologica (stazione, data analisi, tipo di aggregazione e altra variabile meteorologica vengono invece dedotte dallo stato del sistema)	Il sistema aggiorna i grafici e le tabelle di sintesi nell'interfaccia utente in modo coerente con i dati di input.	Il previsore effettua la scelta della variabile attraverso un pulsante dell'interfaccia; Il sistema reagisce aggiornando il grafico della serie temporale corrispondente (prima o seconda variabile) coerentemente con stazione, data e periodo di aggregazione dedotti dallo stato del sistema; se la variabile è scelta tra vento, pioggia e temperatura, viene aggiornata anche la corrispondente parte statistica; il sistema aggiornerà anche la tabella delle giornate vicine coerentemente con la nuova variabile scelta (vedi UC6)

UC5 – il previsore cambia data dell'analisi			
Requisiti	Input	Output	Descrizione
IV	Data del sistema (stazione, sigla della prima variabile meteorologica, sigla della seconda variabile meteorologica, tipo di aggregazione vengono invece dedotte dallo stato del sistema)	Il sistema aggiorna i grafici e le tabelle di sintesi nell'interfaccia utente in modo coerente con i dati di input.	Quando il sistema viene avviato, si configura in modo predefinito sulla data odierna dedotta dall'orologio hardware; il previsore può scegliere la data da un calendario oppure può spostarsi di un intervallo costante (es. un giorno) sul periodo immediatamente precedente o successivo alla data di cui allo stato del sistema; il sistema reagisce aggiornando i grafici della serie temporale secondo la data scelta coerentemente con stazione, variabili meteo plottate e periodo di aggregazione dedotti dallo stato del sistema; il sistema aggiorna anche gli strumenti di accesso alle conoscenze acquisite (vedi UC6, UC7, UC8);

UC6 – il previsore sceglie un giorno tra quelli vicini alla data in esame			
Requisiti	Input	Output	Descrizione
V	Data di confronto e Stato del sistema (data analisi, prima e seconda variabile, stazione e tipo di aggregazione)	Il sistema aggiorna la rappresentazione grafica della prima variabile sovrapponendo l'andamento della giornata scelta tra quelle più vicine.	Il sistema presenta una tabelle delle 5 giornate più vicine rispetto alla stazione, alla data ed alla variabile meteorologica in esame; questa tabella è stata in precedenza aggiornata da uno tra i casi d'uso UC3, UC4 o UC5; da questa tabella il previsore può scegliere uno di questi giorni e conseguentemente il sistema aggiorna il grafico della prima variabile esposta sovrapponendo l'andamento del giorno vicino per un agevole confronto;

UC7 – il previsore sceglie di approfondire le regole individuate			
Requisiti	Input	Output	Descrizione
V	Stazione, giornata in esame	Una rappresentazione tabellare delle regole sullo stile di quanto indicato nel paragrafo 2.4.1	In un piccolo riquadro, il sistema presenta un'interfaccia sintetica della migliore regola trovata rispettivamente per la temperatura, per le precipitazioni e per il vento che viene aggiornata quando si attiva uno tra i casi d'uso

			UC3 e UC5; se il previsore richiede un'analisi più approfondita, il sistema presenta una tabella sullo stile di quanto descritto nel paragrafo 2.4.1.

UC8 - il previsore sceglie di fare una previsione fino a +3 ore			
Requisiti	Input	Output	Descrizione
V	Data del sistema, prima variabile in esame;	Andamento grafico della prima variabile a cui viene aggiunto, con altro colore, la previsione a + 3h sullo stile di quanto indicato nel paragrafo 2.4.3	Condizione: l'accesso alla previsione a + 3h è condizionato dal fatto che la data di analisi sia la data odierna; attraverso un pulsante il previsore può attivare l'elaborazione corrispondente; il sistema avvisa della necessità di una breve attesa visto l'impegno di qualche secondo per pervenire alla previsione;

UC9 – il previsore sceglie di aggregare i dati su periodi temporali diversi			
Requisiti	Input	Output	Descrizione
IV	Tipo di aggregazione (data di analisi, stazione, variabile etc... vengono dedotte dallo stato del sistema);	Visualizzazione grafica dell'andamento della variabile meteorologica aggregata su periodi diversi (settimana, mese , anno, intero archivio)	Il sistema presenta una serie di pulsanti per l'attivazione delle aggregazioni temporali; l'utente sceglie il pulsante e si ottiene l'output descritto;

4.2 - Gli stati dell'interfaccia utente

A partire dai casi d'uso indicati si può notare come l'interfaccia utente del sistema, si possa trovare in alcuni stati ben definibili dalla combinazione dei seguenti parametri:

- la stazione su cui si focalizza l'analisi (codice numerico della stazione);
- la prima delle due variabili meteo da analizzare (un valore tra ta,rh,rr,rs,vv,dv,bar);
- la seconda delle due variabili meteo da analizzare (come sopra);
- la data su cui si focalizza l'analisi;
- il tipo di raggruppamento temporale, che corrisponde al numero di giorni precedenti alla

data di analisi (1 gg,2 gg, 1 settimana,1 mese, 1 anno o tutto l'archivio);

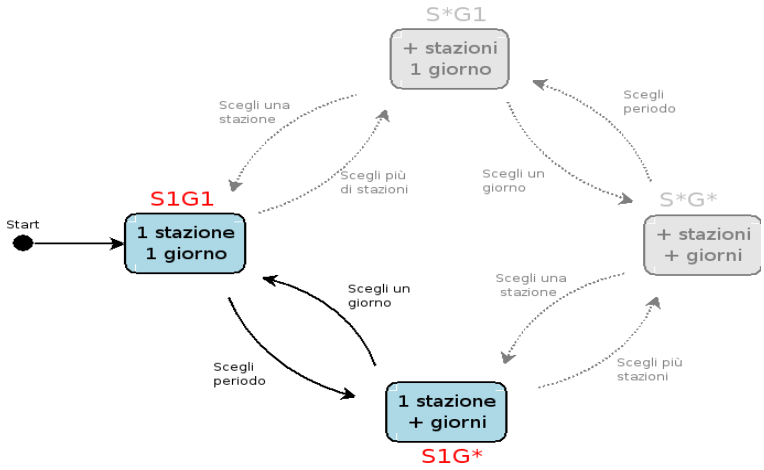


Figura 18: Stati dell'interfaccia utente.

Le combinazioni possibili sono molte, ma ciò che principalmente differenzia gli stati è l'aggregazione temporale (cioè la dimensione principale di analisi) e l'aggregazione spaziale (cioè l'altra dimensione di analisi per il momento solo ipotizzabile), che consenta di aggregare le misure meteorologiche di più stazioni (ad esempio per quote omogenee o per bacini o per zone).

Definiamo quindi i quattro stati seguenti secondo lo schema di figura 18 e ad ognuno di essi assegniamo un'etichetta. Per ora alcuni stati sono interdetti perché non ancora implementati.

Tenendo presente il disegno dei componenti ad alto livello e la definizione dei casi d'uso, si arriva all'implementazione del codice passando attraverso la definizione delle classi specializzate nelle varie operazioni.

Per i casi d'uso 5, 6, 7 e 8 devo definire le classi e le interazione tra le stesse; per questo nel prossimo sotto-capitolo, presento alcune classi di esempio e la sequenza delle operazioni che si intendono implementare.

4.3 - Le classi e le interazioni

Il caso d'uso numero 5 ad esempio (cambia data), si attiva quando l'attore sceglie di

cambiare la data di analisi. A seguito dell'evento ricevuto dall'interfaccia utente, il sistema riconosce il suo stato deducendolo ad esempio da una variabile di sessione web oppure da un cookie e, attraverso le informazioni di stato (data, stazione, variabile meteo da graficare, tipo di aggregazione temporale), attiva una procedura lato server (una servlet Java) che a sua volta si avvale di un Java Bean specializzato nel recuperare i dati dal database e nel restituirli alla servlet come tipologia di dati indipendente dalla sorgente (una classe Java che implementa un vettore dinamico). La servlet a sua volta, utilizzando una libreria specializzata nella generazione di grafici di serie temporali (JFreeChart), utilizza i dati per costruire e restituire al browser del client un immagine grafica che riflette l'andamento temporale della variabile.

Essendo cambiata a questo punto la data di rappresentazione l'interfaccia dovrà cambiare, in modo molto simile alla procedura sopra indicata, anche la rappresentazione dell'andamento temporale della seconda variabile meteo, la tabella con i dati numerici della variabile, la tabella con i giorni meteorologicamente vicini, la tabelle sintetica delle regole.

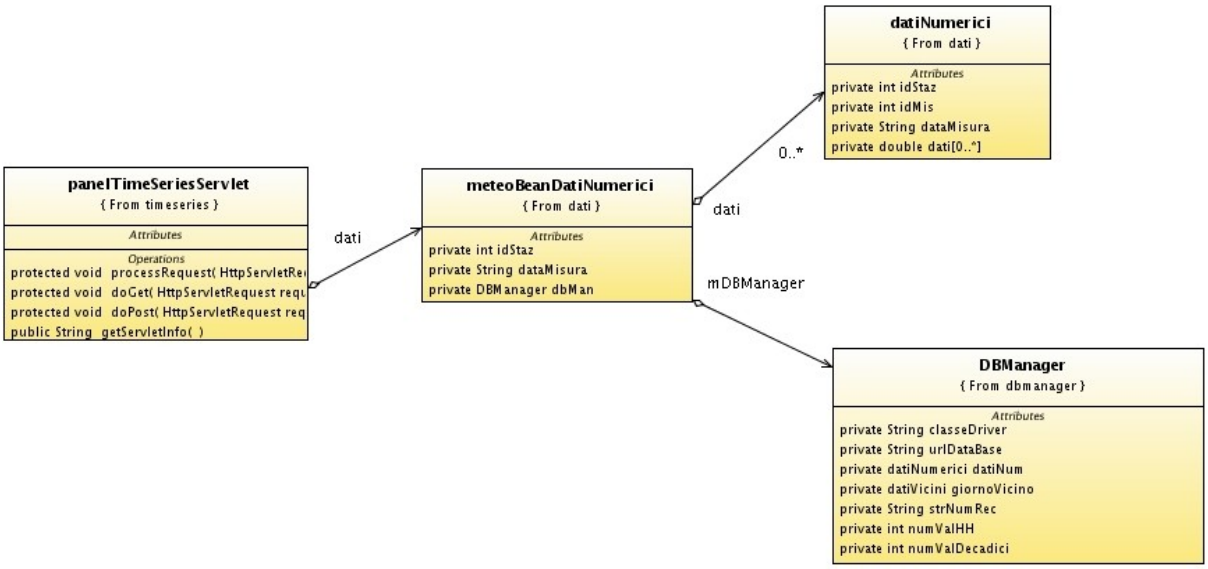


Figura 19: alcune classi di esempio che implementano i casi d'uso 5 e 6

Nell'immagine di figura 19, viene riportato lo schema semplificato di alcune classi che sono coinvolte nella restituzione di un grafico di una serie temporale.

Per rappresentare in modo più efficace la sequenza delle operazioni eseguite, la stessa operazione di restituzione del grafico della serie temporale, viene schematizzata con il diagramma sequenziale di figura 20. In modo molto simile sono state progettate ed implementate le altre classi relative alla completa gestione di tutti i casi d'uso descritti in precedenza.

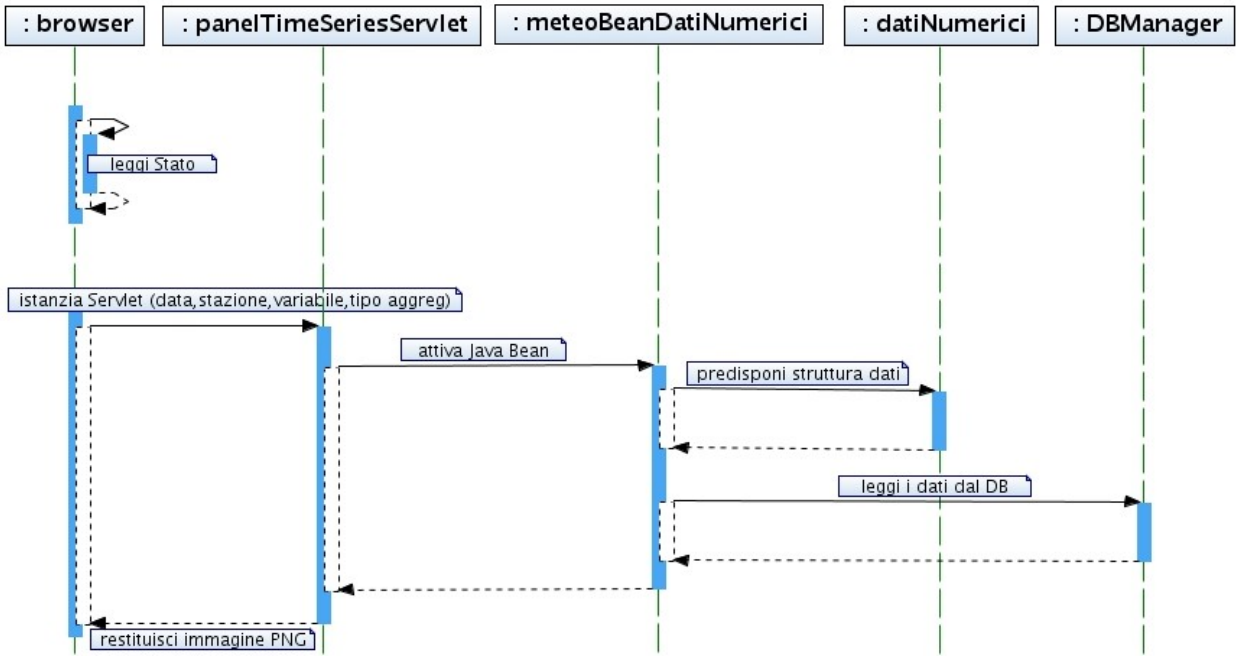


Figura 20: Sequenza delle operazioni

Capitolo 5 - MEKA, il prototipo sviluppato

Nel presente capitolo si descrive come è stato implementato il prototipo software e come lo stesso può essere utilizzato dagli operatori.

Il progetto si è concretizzato con l'implementazione di un prototipo di un applicativo software che integra in un'unica interfaccia le principali funzionalità di analisi ritenute idonee e passate al vaglio della progettazione di cui al capitolo 3 ed al capitolo 4.

Il software è stato denominato MEKA (Meteo Environment for Knowledge Analysis) per la sua derivazione dal più generale strumento di analisi WEKA dal quale sono state ereditate le principali funzionalità di analisi.

Il software può eseguire in un ambiente distribuito, costituito da sistemi client dotati di browser di navigazione e da uno o più componenti fisici lato server (servers); per un'eventuale scalabilità del sistema può infatti essere opportuno suddividere il carico sul lato server installando il database su una postazione hardware ed il web server su un'altra postazione.

Per la parte server l'implementazione del software è stata fatta in linguaggio Java utilizzando le modalità di programmazione lato server (Java Server Pages e Servlet Java) che estendono le funzionalità dello stesso.

Per la parte client sono state realizzate pagine HTML e JSP che utilizzano fogli di stile a cascata (CSS) ed è stato utilizzato il linguaggio di script lato client Javascript. La gestione delle tabelle viene fatto sfruttando le potenzialità di Javascript di elaborare file XML, i veri contenitori delle informazioni, a sua volta prelevati tramite servlet dalle componenti interne del sistema (JavaBean e classi di accesso ai dati).

Il lato server è supportato dall'ambiente “Apache Tomcat 6.0”, il contenitore delle applicazioni e delle tecnologie Java per il Web, quindi il software funziona su server web e, a seconda della raggiungibilità di questo server, può essere disponibile sia su una rete intranet che

su internet.

Oltre a WEKA è stata utilizzata anche la libreria JFreeChart[17], che mette a disposizione un'ampia gamma di classi Java per la rappresentazioni di grafici di serie temporali di vario tipo.

Il database che contiene tutti i dati è, come è stato citato in precedenza, postgreSQL 8.2.9.

La tecnologia utilizzata è disponibile sia per il sistema operativo Windows che Linux e quindi il sistema, pur essendo stato sviluppato in quest'ultimo ambiente esegue in entrambi.

Allo stato attuale, il prototipo è per il momento installato in un unica postazione di lavoro equipaggiata sia con il database, che con il web-server che con il browser-client.

L'interfaccia principale viene richiamata aprendo un browser e digitando l'indirizzo del server web come segue: “http://<server-name>/Web/meka-intro.jsp”.

Nel sotto capitolo che segue vengono ora descritte le funzionalità principali.

5.1 - Descrizione dell'utilizzo

In figura 21 è stata riportata la schermata principale che appare all'avvio del sistema con l'indicazione (in rosso) della finalità a cui sono state destinate le varie aree dell'interfaccia.

17 - JfreeChart è una libreria grafica liberamente scaricabile da http://www.jfree.org

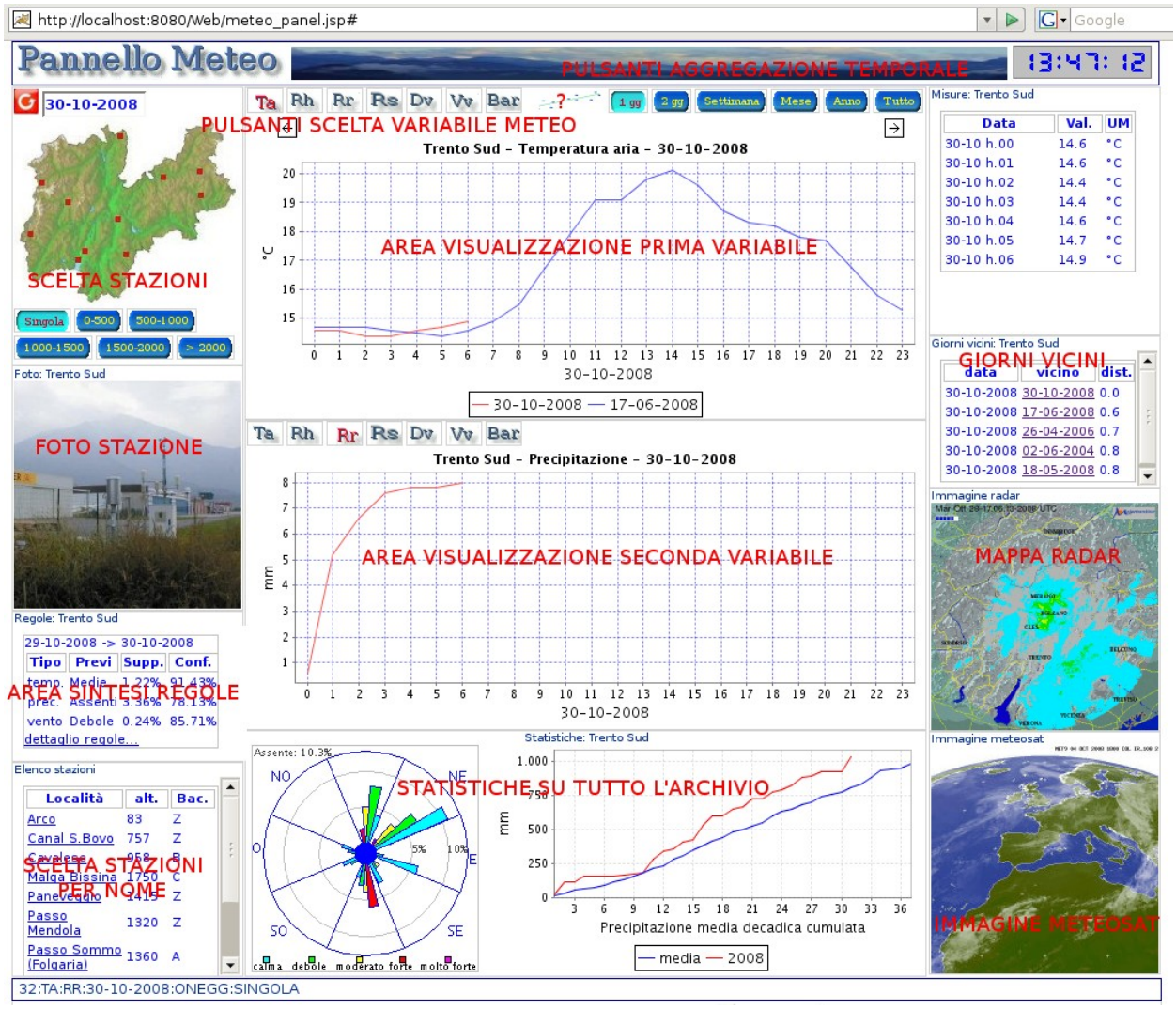


Figura 21: Schermata principale

Ognuna di queste aree ha un comportamento autonomo in relazione al tipo di interazione dell'utente e si auto aggiorna nel momento in cui lo stato del sistema cambia (cambio di data, cambio di stazione, cambio di variabile, cambio del periodo di aggregazione).

I principali elementi attivi del sistema, gli elementi cioè che determinano una modifica dello stato, sono i seguenti:

- due pulsantiere (una per ogni serie temporale) da cui scegliere le variabili da analizzare;

- una pulsantiera per scegliere il periodo di aggregazione temporale;

- un calendario da cui scegliere la data;

- dei pulsanti di navigazione temporale per spostarsi rispettivamente in avanti o indietro nel tempo di un numero di giorni pari all'intervallo di aggregazione scelto;

- un pulsante da cui lanciare l'esecuzione dell'analisi di regressione per la previsione fino a +3 ore che si attiva solo nelle condizioni di trovarsi ad esaminare la giornata odierna;

- altri elementi attivi di facile intuizione nelle sue funzionalità, sono:
  - le mappe radar e satellitari per le quali con un click del puntatore mouse sull'oggetto si ottiene un ingrandimento della mappa;
  - un elenco tabellare delle stazioni (da cui scegliere la stazione per nome);
  - una mappa delle stazioni, da cui scegliere la stazione (scelta per posizione geografica) su cui è possibile anche effettuare un ingrandimento della mappa;
  - collegamenti ipertestuali nelle tabelle giorni vicini per un facile accesso al giorno vicino della variabile in esame;
  - collegamento ipertestuale per accesso immediato all'analisi di tutte le regole.

Capitolo 6 - I test svolti e le valutazioni

Sul prototipo software sono stati realizzati dei test volti a verificarne la correttezza dell'implementazione ed i tempi di esecuzione. Vengono quindi riportati i risultati ottenuti e le valutazioni conseguenti.

Per la valutazione del prototipo sono stati svolti alcuni test. In particolare, vista la fase prototipale dell'applicazione, sono stati svolti dei collaudi di tipo black-box concentrandosi quindi esclusivamente sul rispetto e sulla valutazione dei requisiti funzionali dell'applicazione.

I test di questo tipo, a differenza di ciò che si potrebbe fare con un collaudo di tipo white-box, non esauriscono quindi tutti i possibili cammini che il programma potrebbe fare, ma consentono tipicamente di rilevare errori di funzionalità mancanti, errori di interfaccia utente, errori di strutturazione o di accesso ai dati, errori di inizializzazione e di terminazione.

Sono stati quindi riportati i casi d'uso di cui al capitolo 4.1 verificandone la correttezza funzionale e registrando i tempi di esecuzione del componente eseguito in modo indipendente (laddove il componente lavora in modo indipendente) o all'interno del sistema integrato (qualora la sua azione venga svolta all'interno dell'interfaccia integrata) in modo tale da verificarne le capacità in condizioni operative.

I test sono stati svolti su un calcolatore personale portatile IBM T42 dotato di processore Intel Centrino con una frequenza di clock di 1.7 Ghz, equipaggiato con 1 GB di memoria RAM e fornito di sistema operativo Linux – Ubuntu 7.10 kernel 2.6.22-15-generic.

Il sistema è stato reso operativo sullo stesso calcolatore sia per quanto riguarda l'interfaccia utente (il client web Firefox 2.0), che per il servlet container e web server Java (Apache Tomcat 6.0), che per il motore di database (postgreSQL 8.2.9).

I test, eseguiti su una sola postazione, non tengono ovviamente conto della necessità di scambiare i dati via rete qualora il sistema sia distribuito su più componenti hardware ed inoltre non tengono conto del potenziale carico dovuto ad un numero elevato di utenti.

Per il primo dei due problemi è ovviamente opportuno dislocare il sistema su un server solo se effettivamente si dispone di una struttura di rete adeguata (il sistema è stato anche testato nelle sue funzionalità tra due postazioni portatili che comunicano tra loro in modalità wireless su una mini rete locale riscontrando tempi di risposta analoghi a quelli svolti sulla postazione locale).

Per il secondo problema non è stato sviluppato alcun test, tuttavia visto che la procedura è rivolta ad un numero limitato di utenti che sono interessati all'analisi meteo, un'eventuale accesso può essere eventualmente protetto e limitato ad un numero di utenti pari a quello sostenibile dal sistema qualora si riscontrassero delle difficoltà operative.

Questo tipo di test sono stati svolti in modo autonomo e indipendente rispetto all'utenza e sono descritti nel successivo sotto capitolo 6.1.

Il sistema è stato inoltre valutato da un potenziale utente che ha rilasciato un suo commento riportato di seguito nel sotto capitolo 6.2.

6.1 - Test

Nella tabella conclusiva di questa sezione in figura 22 è riportata la sintesi dei test svolti. Nell'elenco è facile constatare per molti di essi la corrispondenza della descrizione con i casi d'uso elencati; nella stessa tabella sono stati riportati anche i tempi di esecuzione alle condizioni che si espongono di seguito.

Alcuni di questi test sono stati svolti in modo indipendente dall'interfaccia utente; ad esempio non ha alcun significato il test del componente integrato nell'interfaccia per la rigenerazione una tantum delle regole, che viene svolto automaticamente dal sistema fuori dal contesto delle operazioni dell'utente; per la maggior parte degli altri casi, il componente è stato invece collaudato direttamente all'interno dell'interfaccia utente insieme agli altri componenti della stessa.

Per il test è stata utilizzata una stazione meteorologica fittizia caricando i dati casuali di circa 20 anni.

Di seguito si espongono le modalità con cui sono stati svolti i singoli test:

1. Per il primo test, che corrisponde all'attività circoscritta nel caso d'uso UC1, è stata usata una procedura che ad intervalli regolari richiama una pagina web, scarica i dati di una stazione e li inserisce nel database; è stato riportato il tempo di esecuzione peggiore ottenuto da una serie di prove ripetute;

2. Per il secondo test, che corrisponde all'attività di cui al caso d'uso UC2, ho predisposto una procedura batch che, nelle intenzioni di progetto, dovrebbe eseguire una volta al giorno per rigenereare le regole delle stazioni; questa procedura è la più onerosa in termini di tempi di esecuzione; per il test è stato istruito l'algoritmo “Apriori” in modo da estrarre un massimo di 10.000 regole che abbiano una confidenza minima del 60% ed un supporto minimo dell 0.1 % cioè estraggo anche regole che coinvolgono anche solamente poche istanze; i tempi esposti riguardano poi una sola tipologia di regole (quelle del vento), mentre operativamente la procedura deve eseguire altre due volte per ogni stazione (per le regole di precipitazione e di temperatura);

3. Per il terzo test (non compreso in nessun caso d'uso), è stato opportuno verificare anche i tempi di attivazione del sistema; infatti, una volta attivata la pagina principale del sito web, il sistema deve aggiornare tutti gli oggetti dell'interfaccia secondo una vista predefinita e pertanto impiega qualche secondo per essere completamente aggiornata l'interfaccia utente;

4. Il quarto test (caso d'uso UC3), riguarda la scelta di una stazione diversa all'interno dell'interfaccia del sistema; al cambio di stazione il sistema deve aggiornare alcuni oggetti dell'interfaccia ed è quindi sottoposto ad un certo carico;

5. La scelta di una variabile meteorologica diversa da quella visualizzata (caso d'uso UC4), ha ripercussioni solo su un numero più limitato di oggetti dell'interfaccia che devono essere aggiornati; solo nelle condizioni in cui devono essere aggiornate le statistiche generali (scelta della variabile temperatura aria o precipitazione) il sistema è sottoposto ad un carico leggermente superiore e questo è il caso di est;

6. La scelta di cambiare la data di analisi (caso d'uso UC5), che può essere fatta sia attraverso il calendario che attraverso le frecce di navigazione, determina anch'essa l'aggiornamento di alcuni oggetti dell'interfaccia ;

7. La scelta di un giorno vicino (caso d'uso UC6) è un'operazione alla quale il sistema reagisce abbastanza rapidamente essendo necessario solo il recupero di due giorni di dati di una variabile meteo per aggiornare il grafico relativo alla sola prima variabile visualizzata;

8. Per l'accesso ad un'analisi dettagliata delle regole (caso d'uso UC7), il sistema attiva una pagina web separata in cui elenca tutte le regole compatibili con la giornata in esame e quindi la risposta dipende molto da quante regole compatibili ci sono; i tempi riportati riguardano quindi un caso medio;

9. Per la visualizzazione della previsione a +3 h dopo aver scelto una variabile meteo e trovandosi nella giornata odierna, si clicca sul corrispondente pulsante ottenendo la modifica del grafico della prima variabile temporale in un tempo mediamente uguale a quello riportato indipendentemente dalla variabile considerata;

10. Per il testo di aggregazione dei dati (caso d'uso UC9) è stata scelta la condizione peggiore e cioè l'aggregazione di tutto l'archivio della stazione fittizia;

numero test	Descrizione prova	Rispetto funzionalità	Tempo di esecuzione (s)
1	Estrazione e caricamento dei dati	OK	2s
2	Rigenerazione regole una tantum	OK	335s
3	Inizializzazione dell'interfaccia utente	OK	22s
4	Scelta di una stazione	OK	3s
5	Scelta di una variabile meteorologica	OK	2s
6	Cambio di data	OK	3s
7	Scelta di un giorno vicino	OK	1s
8	Scelta analisi dettagliata regole	OK	4s
9	Scelta previsione a +3h	OK	10s

Figura 22: Test del sistema

6.2 - Valutazioni

La valutazione generale del software è sempre affidata all'utente finale del sistema, poiché solo dal costante utilizzo dello strumento e dal conseguente feedback dato allo sviluppatore può essere percepita la buona riuscita dell'applicativo.

Diversamente dagli applicativi di tipo gestionale, nel caso specifico di un software destinato all'analisi com'è MEKA diventa ancora più decisivo il parere dell'utente.

Si riporta di seguito il commento espresso dall'ingegner Emanuele Eccel, ricercatore presso l'unità di Agrometeorologia e Climatologia dell'Istituto Agrario di S.Michele All'Adige e previsore presso il servizio meteorologico della Provincia Autonoma di Trento (Meteotrentino):

“Il software MEKA si propone all’utente con un’accattivante interfaccia grafica per la ricerca e visualizzazione delle informazioni prodotte. In un’unica schermata è possibile visualizzare gli ultimi dati meteo della stazione selezionata ed effettuare quindi una ricerca delle giornate meteorologicamente più “vicine” a quella in esame, con una quantificazione dell’indice di somiglianza. Il software consente perciò di ottenere indicazioni, valutandone l’affidabilità sull’evoluzione successiva delle principali grandezze meteorologiche. Di particolare interesse la possibilità di ottenere un’indicazione del nowcasting per le successive tre ore, desunto dalla regressione multilineare con le grandezze misurate nelle ore precedenti. L’applicazione risulta un utile corredo alle informazioni normalmente disponibili ad un previsore meteorologico, sia in fase di analisi che di previsione. Le informazioni vanno naturalmente ad integrare quelle ad uso prognostico che vengono fornite dal modello meteorologico operativo e possono costituire uno strumento di approfondimento (con dettaglio orario) e downscaling (in quanto calibrate su singole stazioni) della previsione a più larga scala.”.

Conclusioni e prospettive future

Il presente lavoro, sorto dalla consapevolezza di disporre di un patrimonio di informazioni meteorologiche importanti, come quello fornito dalle reti di stazioni meteorologiche regionali, ha evidenziato che all'interno del settore delle previsioni del tempo esiste l'opportunità di sfruttare questa risorsa per addestrare classici schemi di apprendimento automatico basati sulla classificazione delle informazioni al fine di disporre di uno strumento di analisi complementare a quelli usualmente utilizzati.

Dopo aver analizzando i dati di due sole stazioni meteorologiche per un periodo temporale limitato, è stato costruito un prototipo di applicativo software che consente di rappresentare delle caratteristiche ricorrenti rintracciabili nei dati meteorologici in modo agevole, quali le giornate meteorologicamente “vicine” e le regole dedotte dall'insieme di dati; in questa maniera è sembrato conveniente proporre una strategia di previsione basata sul confronto per analogia con situazioni del passato.

E' stato anche possibile implementare un semplice schema di apprendimento lineare per predirre i valori delle successive 3 ore rispetto alla data attuale di analisi di alcune variabili meteorologiche. Se applicato operativamente il sistema potrebbe quindi consentire al previsore meteorologico di prendere delle decisioni supportato da uno strumento di analisi oggettiva.

Alcuni degli algoritmi di apprendimento analizzati (analisi regressione e giorni vicini) sono meglio performanti di altri in termini di velocità di esecuzione, per cui, nell'implementazione del prototipo del sistema di analisi quasi real-time, sono stati preferiti rispetto a schemi di apprendimento più credibili (decision table, bayes); per l'algoritmo “Apriori”, è stata distinta la fase operativamente più pesante dell'algoritmo (generazione delle regole) dall'attività di analisi delle regole utili (esposizione delle regole).

L'applicazione in campo meteorologico di questi semplici schemi di apprendimento lineari e/o per analogia potrebbe tuttavia essere solo un primo passo per lo sviluppo di un sistema di analisi integrato che disponga anche di schemi di apprendimento più complessi.

L'applicativo software potrebbe infatti integrare altre informazioni di cui si avvale il previsore meteorologico nelle attività di previsione a brevissimo termine (massimo 24 ore), come ad esempio le immagini satellitari, le immagini radar, le foto delle condizioni del tempo trasmesse dalle videocamere installate sulle stazioni.

Il sistema potrebbe fare un ulteriore salto di qualità se tutti i dati forniti da vari strumenti di misura e osservazione (satellite, radar, foto, etc ...) potessero essere analizzati in modo integrato, ma questo è un aspetto molto più complesso perché coinvolge un notevole numero di fonti informative eterogenee non solo per la modalità con cui sono archiviati i dati ma anche rispetto a ciò che rappresentano.

Un elemento che potrebbe invece migliorare lo stato attuale del prototipo senza ulteriori necessità di modifiche nell'attuale architettura fisica, è invece la possibilità di far confluire nel sistema le stazioni di una rete a scala regionale in modo da utilizzare i dati che non entrano nei circuiti di analisi internazionali e che sono quindi sfruttati solo in modo parziale. In questo modo si possono ipotizzare aggregazioni e analisi su un numero notevole di stazioni.

In conclusione, si può affermare che c'è spazio quindi per gestire una modellistica locale di previsione a brevissimo termine che non si basi solo sull'analisi fisico-matematica, ma anche sulla capacità di classificare ed analizzare le informazioni rilevate al suolo.

Il prototipo software implementato, consente inoltre anche ad una persona non esperta in meteorologia di analizzare regole meteorologiche dedotte dalla discretizzazione dei dati delle variabili meteorologiche (tipicamente dati numerici) per interpretare in modo quasi naturale le relazioni esistenti tra i vari parametri meteorologici.

Questo aspetto si è rilevato potenzialmente interessante per fornire una risposta ai ragazzi delle scuole primarie che, negli incontri di formazione con gli esperti di meteorologia, chiedono spesso, in modo assai perspicace, dell'esistenza di algoritmi, regole o procedure per prevedere il tempo a partire dalle osservazioni. Le regole possono essere una risposta, seppur semplice, facilmente interpretabile.

Un risultato non secondario del presente lavoro è anche quello di aver messo in

evidenza la necessità di seguire anche nella gestione dei dati meteorologici una linea di tendenza comune nel campo della scienza e della tecnologia dell'informazione: lo sviluppo del data mining come strumento che consente ad un'organizzazione di crescere mantenendo una solida base strutturata di conoscenze dal quale attingere per la ricerca di utili analogie con le situazioni del passato.

In questo modo anche in meteorologia, come già accade in altri campi e come accadrà sempre più spesso in futuro in molti campi di applicazione, il processo decisionale si prefigura come un percorso ben definito di estrazione di conoscenza dalle enormi quantità di informazioni disponibili le quali difficilmente possono essere interpretate correttamente dall'uomo.

Bibliografia

[1] M.Fowler (2000), *"UML Distilled - prima edizione italiana"*, Milano, Addison Wesley Longmann.

[2] H.Garcia-Molina, J.D.Ullman, J.Widom (2002), *"Database systems : the complete book"*, International ed., Upper Saddle River N.J., Prentice-Hall.

[3] P. Giudici (2005), *"Data mining: metodi informatici, statistici e applicazioni"*, 2nd ed., McGraw-Hill.

[4] M.Golfarelli, S.Rizzi (2002), ***"****Data warehouse : teoria e pratica della progettazione"*, *Milano*, McGraw-Hill Companies.

[5] J.Han, M.Kamber (2006), *"Data mining : concepts and techniques"*, 2nd ed., San Francisco, Morgan Kaufmann.

[6] G.Kappenberger e J.Kerkman (1997), *"Il tempo in montagna - manuale di meteorologia alpina"*, Bologna, Zanichelli.

[7] I.H.Witten e E. Frank (2005), *"Data Mining: Practical Machine Learning Tools and Techniques"* , 2nd ed., San Francisco, Morgan Kaufmann.

[8] *"PostgreSQL 8.1.4 Documentation"* - URL: http://www.postgresql.org/docs/

[9] *"Sito ufficiale WEKA"*, http://www.cs.waikato.ac.nz/ml/weka/

[10] Ch. Häberli, D. Tombros, *"MCH-DWH: the MeteoSwiss Data Warehouse System"*, http://www.meteoschweiz.admin.ch/web/en/research/projects/mchdwh.html

[11] Il sito web di V. Villasmunta, *"Corso basico di meteorologia"*, http://www.villasmunta.it/

[12] Il sito web di Meteotrentino (Provincia Autonoma di Trento) - http://www.meteotrentino.it

[13] Il sito web dell'Unità Operativa di Agrometeorologia e Climatologia dell'Istituto Agrario di S.Michele All'Adige – http://meteo.iasma.it/meteo/

Ringraziamenti

Al termine di questo lavoro voglio ringraziare l'ingegner Emanuele Eccel dell'Istituto Agrario di S.Michele All'Adige dal quale ho ricevuto utili suggerimenti, la dottoressa Marta Pendesini, meteorologo presso Meteotrentino, con cui mi sono confrontato, l'Unità Operativa di Agrometeorologia e Climatologia dell'Istituto Agrario di S.Michele All'Adige che ha messo a disposizione i dati per l'analisi delle stazioni di Trento Sud e Arco, l'Ufficio Previsioni e Organizzazione della Provincia Autonoma di Trento che ha fornito i dati di output del modello di previsione numerica e dal quale ho appreso le conoscenze meteorologiche di base.

Desidero anche ringraziare tutti i miei colleghi di lavoro, in particolare il gruppo che si occupa di meteorologia, il gruppo che si occupa della rete di rilevamento dati ed il mitico Mariano del "gruppo informatici".

Infine un grazie a miei genitori, ai miei familiari ed a tutti gli amici che mi hanno incoraggiato.

Trento, 17 dicembre 2008.