Università degli Studi di Roma "La Sapienza"
Corso di Laurea Magistrale in Informatica

Progetto del corso di Machine Learning

ANALISI DELLA POPOLAZIONE DEGLI INSETTI GRASS GRUBS E DELLA CONTAMINAZIONE DEI RACCOLTI IN NUOVA ZELANDA

a.a. 2013/2014

Docente
Prof.ssa P. Velardi
Assistente
dott. S. Faralli

Studenti
Sara De Cristofano
mat. 1210689
Emanuele Giarlini
mat. 1210359

Indice

Descrizione del problema

I **grass grubs** sono una delle specie di insetti più diffuse e voraci della Nuova Zelanda, e possono causare consistenti danni ai raccolti e ingenti perdite economiche per gli agricoltori.

Il danno avviene maggiormente nella fase di sviluppo e crescita delle larve, che si nutrono delle radici degli arbusti. Gli adulti tendono a riprodursi in zone dal clima asciutto ed in terreni umidi; la Nuova Zelanda, essendo un'isola dal clima temperato, è la zona ideale per la crescita di questi insetti.

L'**obiettivo** di questo progetto è capire come sia possibile prevenire lo sviluppo delle larve dei grass grubs.

Il primo passo è stato analizzare dati degli anni passati relativi alla piovosità ed ai raccolti in diverse zone della Nuova Zelanda. Quindi, utilizzando l'algoritmo **Apriori**, abbiamo generato gli itemset frequenti e da questi abbiamo dedotto delle regole associative che descrivono le condizioni favorevoli o meno allo sviluppo dei grass grubs.

Abbiamo scelto di implementare questo algoritmo di data mining in particolare poiché ci permette di tracciare i pattern frequenti nel nostro insieme di transazioni, che nel nostro caso corrisponde a ricercare le condizioni che portano alla proliferazione delle larve. Apriori, quindi, è una perfetta soluzione al nostro problema.

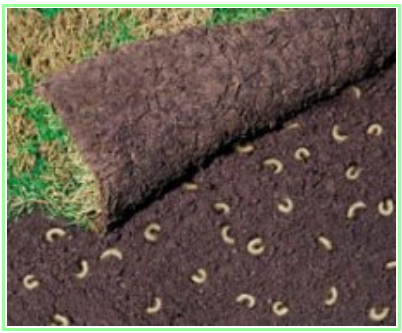
Infine, tramite il classificatore **Naïve Bayes**, è possibile inserire una nuova istanza ed osservare se questa genererà una percentuale di danni alta, media o bassa.

Anche in questo caso, per il nostro problema è di fondamentale importanza poter prevedere l'entità del danno per prevenirlo. Proprio per questo abbiamo scelto di implementare il classificatore Naïve Bayes.

Raccolta ed analisi dei dati

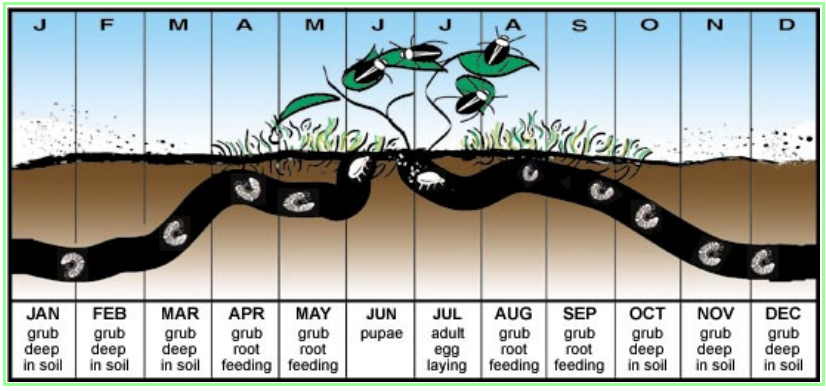
Per comprendere il problema è stato necessario innanzitutto conoscere il comportamento dei grass grubs, e come vanno ad intaccare i raccolti e la produzione agricola neozelandese **[1] [2] [4]**.

I **grass grubs** *(Costelytra zealandica)* sono degli insetti che si cibano delle radici di arbusti, pascoli e colture da campo. Si diffondono in territori piovosi ma dal clima asciutto, poiché le larve si sviluppano e crescono in terreni umidi.



Il **ciclo di vita** dei grass grubs è suddiviso in tre fasi:

1) in primavera le larve escono dal letargo e si nutrono delle radici dei raccolti. Quando si sono nutrite a sufficienza ritornano nel terreno e creano un bozzolo;
2) in estate dal bozzolo fuoriesce uno scarafaggio che si nutre delle foglie e dei fiori degli arbusti, e che poi depone le uova nel terreno;
3) in autunno le uova si schiudono consentendo alle nuove larve di cibarsi delle radici dei raccolti.

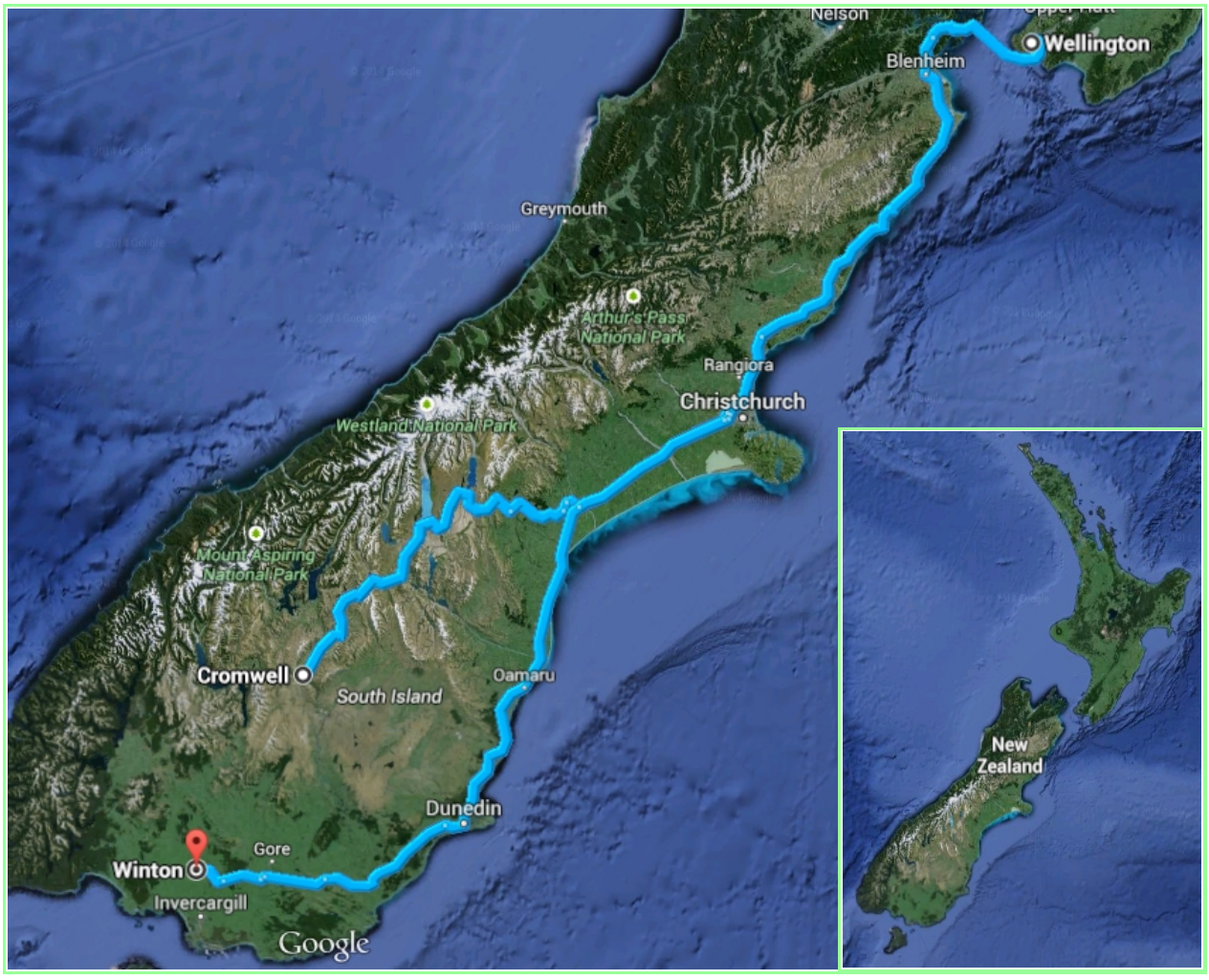


I **danni principali** vengono effettuati, quindi, durante la prima e terza fase, in cui le larve si sviluppano e si cibano delle radici degli arbusti.

Una delle caratteristiche di questi insetti è che, se le condizioni climatiche sono favorevoli, restano nella stessa zona anche per più stagioni consecutive (con picchi di 4 - 6 anni), creando delle vere e proprie colonie che compromettono la produzione raccolto dopo raccolto.

Considerando queste caratteristiche dei grass grubs, la nostra analisi si è quindi focalizzata su come poter **prevenire lo sviluppo delle larve** e come poter **prevedere i danni** ad un particolare raccolto conoscendo la piovosità e le caratteristiche del territorio.

A questo punto siamo passati ad analizzare le **caratteristiche territoriali** e climatiche della Nuova Zelanda.
Questa nazione ha clima simile a quello italiano poiché si trova alla stessa latitudine, ma essendo un'isola è predisposta ad avere molte più precipitazioni. Il territorio è prevalentemente collinare, con zone pianeggianti, molte delle quali sono riserve naturali protette.



Per avere un campione più rappresentativo possibile, abbiamo scelto tre cittadine situate in tre zone territoriali diverse:

- **Cromwell** (zona collinare)
- **Wellington** (zona costiera)
- **Winton** (zona pianeggiante)

Iscrivendoci al sito del **National Climate Database** della nuova Zelanda **[3]** siamo riusciti a risalire alle precipitazioni degli ultimi 30 anni (dal 1980 al 2013). Scegliendo le stazioni meteorologiche corrispondenti alle zone territoriali appena citate, abbiamo ottenuto i dati necessari alla modellazione del nostro problema, che analizzeremo in dettaglio nel paragrafo successivo.

A questo punto abbiamo definito il **tipo di raccolto** da considerare, scegliendo tra quelli più diffusi in Nuova Zelanda **[1]** e quelli più attaccati dai grass grubs. Abbiamo quindi scelto frutta, tabacco, legumi e cereali.

Nella fase di ricerca abbiamo inoltre appreso **[5] [6] [7]** che gli agricoltori spesso si affidano a pesticidi contenenti fosforo per combattere i grass grubs. Questa sostanza, se usata nel periodo autunnale o al momento della semina, rende le radici delle piante più robuste e resistenti all'attacco dei parassiti. Ci è sembrato, quindi, un parametro importante da valutare nell'analisi del quantitativo di danni causato dalle larve.

Modellazione del problema e creazione del dataset

Andiamo a vedere, nei dettagli, come ci siamo mossi per la creazione del dataset.

In base all'analisi fatta in precedenza, il dataset avrà le seguenti features:

- **piovosità** – millimetri di pioggia caduti nel periodo di osservazione; può assumere i valori *alta*, *media*, *bassa*;
- **tipo di raccolto** – può assumere i valori *frutta*, *tabacco*, *legumi*, *cereali*;
- **tipo di terreno** – zona in cui si ha la coltivazione; può assumere i valori *costa*, *collina*, *pianura*;
- **trattamento al fosforo** – indica se il terreno è stato trattato con antiparassitari; può assumere i valori *si* o *no*;
- **danno** (*classificazione*) – indica la percentuale di danno c dai grass grubs; può assumere i valori *alta*, *media*, *bassa*.

Dopo aver individuato le features è stato necessario capire quanto grande dovesse essere il nostro dataset. Usando la **computational learning theory** ed avendo uno spazio delle ipotesi finito, si ha che $|H|\cdot e^{-\varepsilon m} \leq \delta$ da cui si deriva che $m \geq \frac{1}{\varepsilon}(\ln|H| + \ln\frac{1}{\delta})$.

Quindi ci si vuole assicurare che, con probabilità δ, il nostro insieme di campionamento abbia errore minore di ε. Affinché questo sia vero, il nostro dataset deve contenere almeno m istanze.

Applicando questa teoria al nostro caso di studio, otteniamo che:

- $|H| = 3^{72} + 1$
- $\varepsilon = 0.1$
- $\delta = 95\%$

e di conseguenza $m \geq \frac{1}{0,1}(\ln(3^{72}+1) + \ln\frac{1}{0,05}) = 10 \cdot 79,1 + 2,99 = 793,99$.

Per avere una buona approssimazione, quindi, il nostro dataset dovrà contenere almeno 794 istanze.

Il dato più semplice da reperire è stata la **piovosità**, grazie ai dati messi a disposizione dal database climatico neozelandese. Dovendo ottenere almeno 794 istanze, abbiamo effettuato due campionamenti per anno (dal 1980 al 2013) per zona (Cromwell, Wellington, Winton). In questo modo abbiamo ottenuto 2 campionamenti · 34 anni · 3 zone = 204 istanze. Ad ognuna di esse

saranno associati i quattro diversi tipi di raccolto (come vedremo in seguito), quindi otterremo $204 \cdot 4 = 816$ istanze, sufficienti ad avere una buona approssimazione.

Abbiamo ricercato le stazioni meteorologiche corrispondenti alle tre cittadine scelte come campione:

- *Cromwell (identificativo stazioni 5529, 26381, 5524)*
- *Wellington (identificativo stazione 3445)*
- *Winton (identificativo stazione 5768)*

e possiamo vedere nell'immagine seguente il risultato della query al database:

The National Climate Database

NIWA Taihoro Nukurangi

Station information:

Name	Agent Number	Network Number	Latitude (dec.deg)	Longitude (dec.deg)	Height (m)	Position Precision	Observing Authority
Cromwell Ews	26381	I59013	-45.03392	169.19550	213	H	NIWA
Wellington Aero	3445	E14387	-41.322	174.804	4	G	N/A
Cromwell Sub Stn	5524	I59012	-45.062	169.189	213	G	N/A
Cromwell 2	5529	I59024	-45.035	169.195	213	G	N/A
Winton 2	5768	I68133	-46.157	168.328	44	G	N/A

Note: Position precision types are: "W" = based on whole minutes, "T" = estimated to tenth minute, "G" = derived from gridref, "E" = error cases derived from gridref, "H" = based on GPS readings (NZGD49), "D" = by definition i.e. grid points. ***For more info***

Back to Database Query Form

Statistics codes in this query are:

Code	Description	Units
00	Total Rainfall	Mm

Note: Statistics calculations are based on Local-Time.
Monthly extremes are recorded on the Local-Time day of the month.
Annual extremes are recorded in the Local-Time month of the year.

Stats: Combined

Station	Year	Stats Code	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Orig	Rel
3445	1980	00	86.7	29.6	147.7	133.2	79.7	180.9	76.4	100.6	66.9	67.4	117.6	86.5	1173.2	D	-
5524	1980	00	64.5	28.9	56.5	52.8	23.0	75.0	26.8	90.7	41.4	20.3	29.9	28.6	538.4	D	-
5768	1980	00	172.4	80.8	75.3	29.5	98.3	74.0	96.8	160.5	51.6	53.2	122.8	27.0	1042.2	D	-
3445	1981	00	17.2	5.7	47.2	43.7	202.3	124.8	138.3	99.8	52.0	147.3	85.3	49.7	1013.3	D	-
5524	1981	00	7.4	21.0	117.7	23.6	10.7	47.5	35.0	8.0	20.8	26.0	22.8	40.7	381.2	D	-
5768	1981	00	34.6	59.9	110.1	107.9	61.4	74.9	75.3	68.8	131.1	49.4	44.8	133.8	952.0	D	-
3445	1982	00	9.5	43.8	27.0	58.1	50.2	154.3	76.6	47.4	83.0	50.9	55.6	94.4	750.8	D	-
5524	1982	00	37.7	31.2	24.8	12.5	84.8	25.3	16.2	38.0	1.6	28.1	67.8	43.8	411.8	D	-
5768	1982	00	145.2	95.6	72.6	27.2	120.2	47.1	89.1	82.8	40.6	125.2	82.2	140.2	1078.2	D	

Abbiamo diviso ogni annualità in **due semestri**: il primo da maggio ad ottobre ed il secondo da novembre ad aprile. Questa scelta è derivata dalle osservazioni sul comportamento delle larve, che come abbiamo visto causano i danni maggiori in autunno ed in primavera (fase uno e tre del loro sviluppo). Per ogni semestre abbiamo considerato (in modo complessivo comprendendo tutte le zone) il quantitativo minore di pioggia caduta ed il maggiore, ed abbiamo calcolato il valore medio per semestre per anno.

Stats: Combined

Station	Year	Stats Code	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Orig	Rel
3445	1980	00	86.7	29.6	147.7	133.2	79.7	180.9	76.4	100.6	66.9	67.4	117.6	86.5	1173.2	D	-
5524	1980	00	64.5	28.9	56.5	52.8	23.0	75.0	26.8	90.7	41.4	20.3	29.9	28.6	538.4	D	-
5768	1980	00	172.4	80.8	75.3	29.5	98.3	74.0	96.8	160.5	51.6	53.2	122.8	27.0	1042.2	D	-
3445	1981	00	17.2	5.7	47.2	43.7	202.3	124.8	138.3	99.8	52.0	147.3	85.3	49.7	1013.3	D	-
5524	1981	00	7.4	21.0	117.7	23.6	10.7	47.5	35.0	8.0	20.8	26.0	22.8	40.7	381.2	D	-
5768	1981	00	34.6	59.9	110.1	107.9	61.4	74.9	75.3	68.8	131.1	49.4	44.8	133.8	952.0	D	-
3445	1982	00	9.5	43.8	27.0	58.1	50.2	154.3	76.6	47.4	83.0	50.9	55.6	94.4	750.8	D	-
5524	1982	00	37.7	31.2	24.8	12.5	84.8	25.3	16.2	38.0	1.6	28.1	67.8	43.8	411.8	D	-

I risultati sono stati discretizzati basandoci ancora una volta sui minimi e massimi ottenuti, arrivando ad avere gli intervalli di piovosità finali:

- **piovosità bassa**: da 43 mm a 80 mm
- **piovosità media**: da 80 mm a 123 mm
- **piovosità alta**: da 123 mm a 152 mm

Il resto del dataset è generato in modo **automatico** dall'algoritmo implementato in Java. La generazione avviene assegnando ai valori di ogni feature un **peso**: maggiore è il peso, maggiore è la probabilità che quel valore generi o subisca danni. Analizziamo questa strategia in dettaglio.

Ai valori della feature **piovosità** sono stati assegnati i seguenti pesi:

```
PESO_PIOGGIA_ALTO = 3;
PESO_PIOGGIA_MEDIO = 9;
PESO_PIOGGIA_BASSO = 6;
```

Come abbiamo visto, le larve si sviluppano in terreni umidi e climi asciutti. La piovosità media è quindi una condizione ideale affinché il suolo sia bagnato ma l'aria non ristagni. Analogamente, la piovosità alta determina troppa umidità, mentre la piovosità bassa potrebbe non bagnare a sufficienza il terreno.

Ai valori della feature **tipo di raccolto** sono stati assegnati i seguenti pesi:

```
PESO_RACCOLTO_CEREALI = 6;
PESO_RACCOLTO_FRUTTA = 2;
PESO_RACCOLTO_TABACCO = 4;
PESO_RACCOLTO_LEGUMI = 6;
```

La motivazione è che i raccolti danneggiati maggiormente dai grass grubs sono legumi e cereali,

poiché sono arbusti dalle radici sottili facilmente consumabili dalle larve. Viceversa, gli alberi da frutto sono i meno facili da attaccare poiché hanno una struttura più resistente. Il tabacco è a sua volta un tipo di pianta debole, ma tipicamente è coltivato in strutture chiuse, quindi più protetto rispetto a legumi e cereali.

Ai valori della feature **tipo di terreno** sono stati assegnati i seguenti pesi:

```
PESO_TERRENO_COSTA = 6;
PESO_TERRENO_COLLINA = 4;
PESO_TERRENO_PIANURA = 2;
```

Come visto in precedenza, i grass grubs si sviluppano meglio in zone dal clima asciutto, ed in terreni bagnati. Delle tre zone scelte, la costa è sicuramente la più asciutta, mentre la pianura, nelle zone interne, è quella più soggetta a ristagni di umidità nell'aria.

La feature **trattamento al fosforo** è stata assegnata in modo casuale ad ogni istanza poiché non abbiamo riscontrato preferenze negli agricoltori per l'uso di pesticidi in zone o raccolti particolari. I valori della feature sulla percentuale di danno influiscono nel modo seguente **[5] [6] [7]**:

`PESO_FOSFORO = 60;` (se c'è stato il trattamento il danno al raccolto si riduce del 60%)

`PESO_FOSFORO_PIOGGIA = 40;` (se c'è stato il trattamento ma ha piovuto molto, ovvero il trattamento non ha avuto modo di attecchire, il danno al raccolto si riduce solo del 40%).

Il calcolo della percentuale di danno, quindi, avviene nel modo seguente:

Percentuale di danno = peso_piovosità + peso_raccolto + peso_terreno

Questo valore viene poi abbassato del 40% o del 60% se è stato fatto il trattamento al fosforo, e riportato in centesimi.

La **classificazione** avviene quindi nel modo seguente:

- se la percentuale di danno è minore del 35% → il danno al raccolto è *basso*
- se la percentuale di danno è tra il 35% e il 55% → il danno al raccolto è *medio*
- se la percentuale di danno è maggiore del 55% → il danno al raccolto è *alto*

Siccome la feature *trattamento al fosforo* è assegnata casualmente, ad ogni esecuzione del programma verrà generato un dataset diverso, ma comunque coerente con le caratteristiche analizzate finora.

Implementazione in Java

Il programma che abbiamo implementato si divide in tre parti principali:

1) creazione del dataset
2) esecuzione dell'algoritmo Apriori
3) classificazione di una nuova istanza con Naïve Baies

Creazione del dataset

Il dataset viene creato utilizzando la classe `Generator.java`.

Il punto di partenza è l'inserimento manuale della piovosità per ogni anno: questa è memorizzata in tre `LinkedList` per il primo semestre (piovosità alta, media, bassa) e tra per il secondo semestre. Quindi, per ogni valore presente nelle liste, viene generata una istanza avente le feature descritte in precedenza.

A questo punto il dataset è stato generato; ogni istanza viene inserita nella `LinkedList datasetClassification` che verrà utilizzata per le operazioni successive, mentre per chiarezza una copia del dataset viene salvata nel file `dataset_classification.csv`.

Esecuzione dell'algoritmo Apriori

La `LinkedList datasetClassification` creata nella fase precedente rappresenta delle istanze, mentre l'algoritmo Apriori lavora su transazioni. Il contenuto della lista viene quindi trasformato e salvato nel file `database_classification.dat`; in questo file il valore di ogni feature è codificato con un intero diverso (da 1 a 15, si veda il paragrafo *Esempio di esecuzione dello script* per una descrizione dettagliata), in modo da essere adatto per la creazione dei frequent itemset.

La classe `Apriori.java` prende quindi in input il file `database_classification.dat` e il supporto minimo degli itemset, che abbiamo fissato a 20. Il funzionamento dell'algoritmo mappa esattamente lo pseudocodice visto a lezione: vengono generati ed analizzati mano a mano gli insiemi candidati di dimensione k, e vengono eliminati quelli con supporto insufficiente. L'avanzamento dell'algoritmo ed i risultati intermedi vengono salvati nel file `apriori_log.txt`. Alla fine si ottiene l'elenco degli itemset frequenti (aventi supporto almeno 20): questi vengono salvati nel file `apriori_result.txt`.

Da questi itemset poi, manualmente, dovranno essere ricavate le association rules.

Classificazione di una nuova istanza con Naïve Baies

La classe `NaiveBayes.java` consente di classificare una nuova istanza basandosi sul dataset creato.

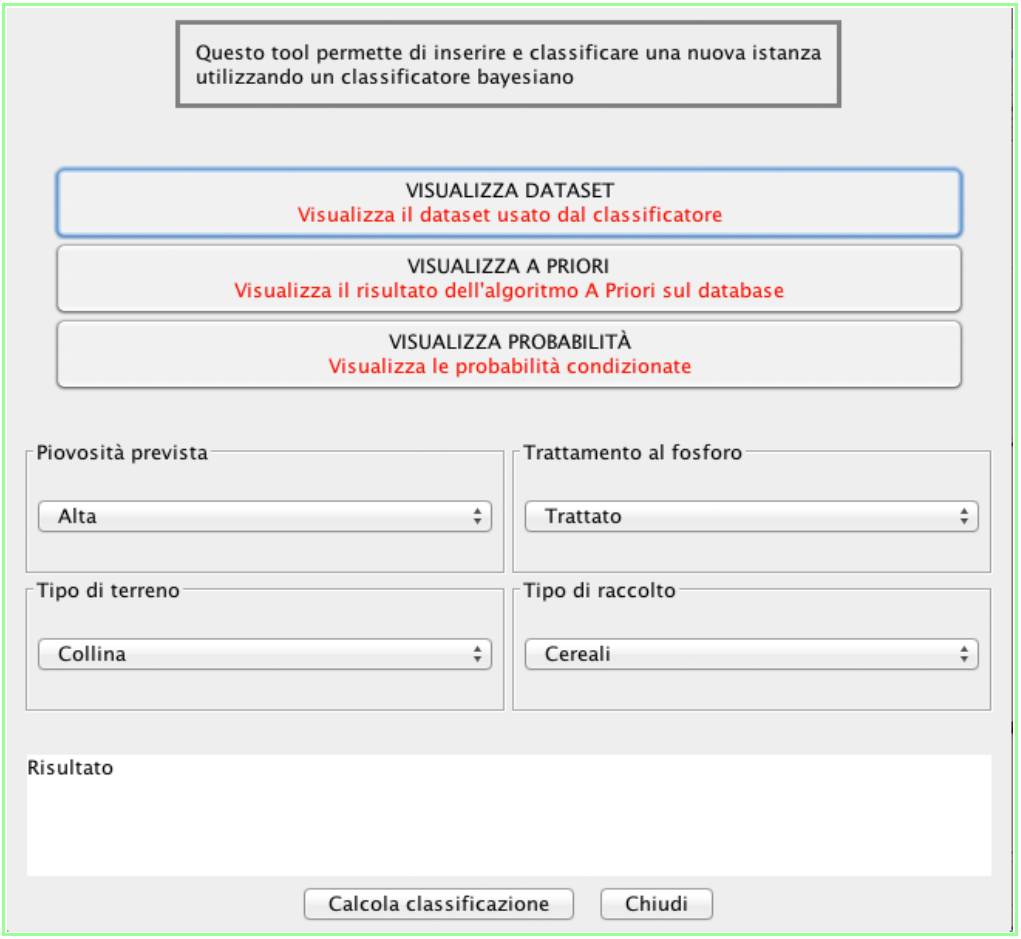
Innanzitutto sono state calcolate le probabilità condizionate dei valori di ogni feature per ognuna delle tre classi, e queste probabilità sono memorizzate nella `HashMap probMap` (e per chiarezza salvate nel file `prob_map.txt`). Quindi, tramite una semplice interfaccia grafica, è possibile modificare i valori di piovosità, tipo di raccolto, tipo di terreno e trattamento al fosforo, creando così una nuova istanza.

Analizzando le nuove features, il programma calcola la probabilità dell'istanza di appartenere alle tre classi, scegliendo la maggiore come classificazione finale. In output verrà visualizzata la classificazione della nuova istanza, ovvero se questa avrà con maggiore probabilità un danno alto, medio o basso.

Esempio di esecuzione dello script

Mostriamo ora un esempio di esecuzione del programma.

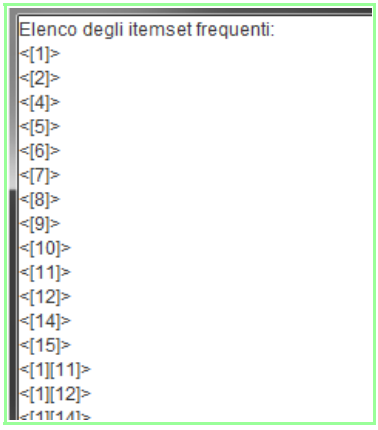
Avviando l'eseguibile *progettoML.jar* si apre la seguente interfaccia:



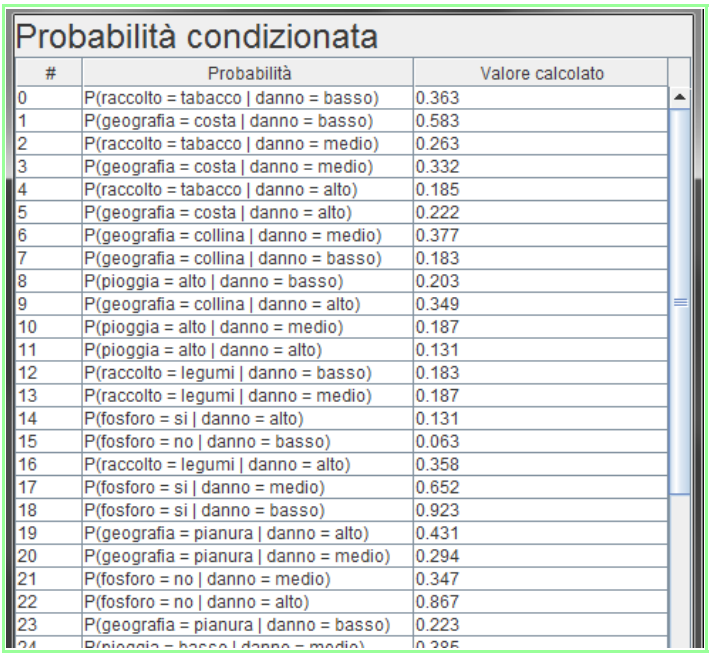
Cliccando sul pulsante *Visualizza Dataset* si apre il file *dataset_classification.csv* contenente il dataset generato dal programma:

#	Anno	Piovosità	Zona	Raccolto	Fosforo	Danno
0	Anno	Piovosità	Zona	Raccolto	Fosforo	Danno
1	1980	media	costa	cereali	si	medio
2	1980	media	costa	frutta	no	alto
3	1980	media	costa	tabacco	no	alto
4	1980	media	costa	legumi	si	medio
5	1980	media	collina	cereali	si	medio
6	1980	media	collina	frutta	si	medio
7	1980	media	collina	tabacco	no	alto
8	1980	media	collina	legumi	no	alto
9	1980	media	pianura	cereali	si	alto
10	1980	media	pianura	frutta	si	medio
11	1980	media	pianura	tabacco	si	medio

Cliccando sul pulsante *Visualizza Apriori* si apre il file *a_priori_result.txt* contenente l'elenco degli itemset frequenti trovati dall'algoritmo:



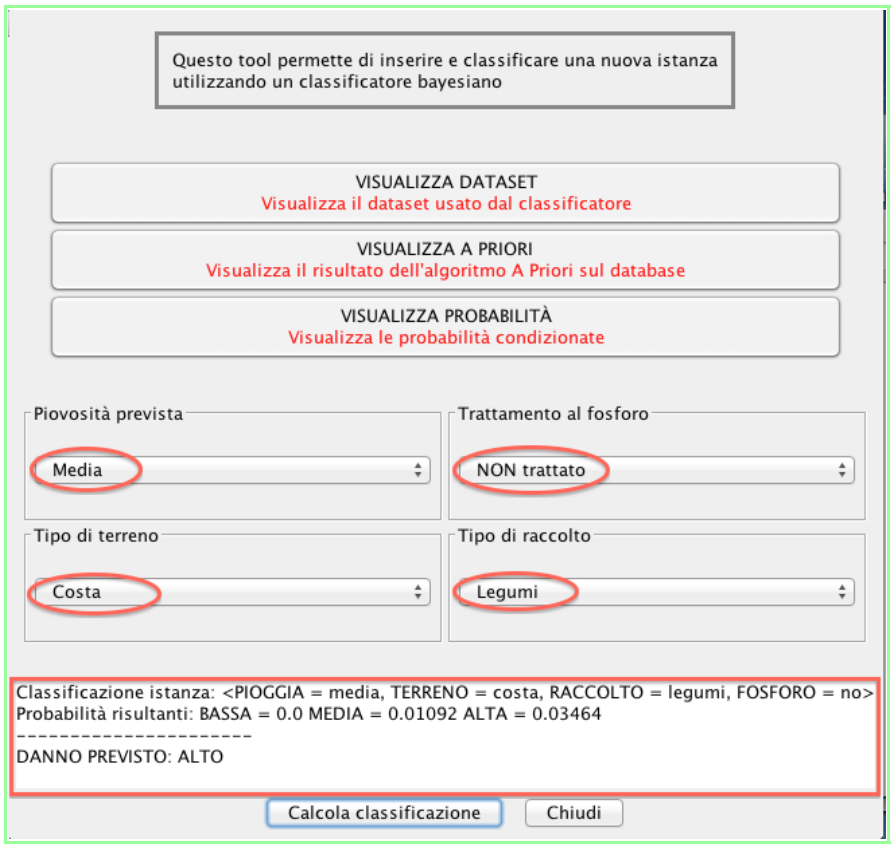
Cliccando sul pulsante *Visualizza Probabilità* viene aperto il file *prob_condizionate.csv* contenente le probabilità condizionate utilizzate per classificare la nuova istanza:



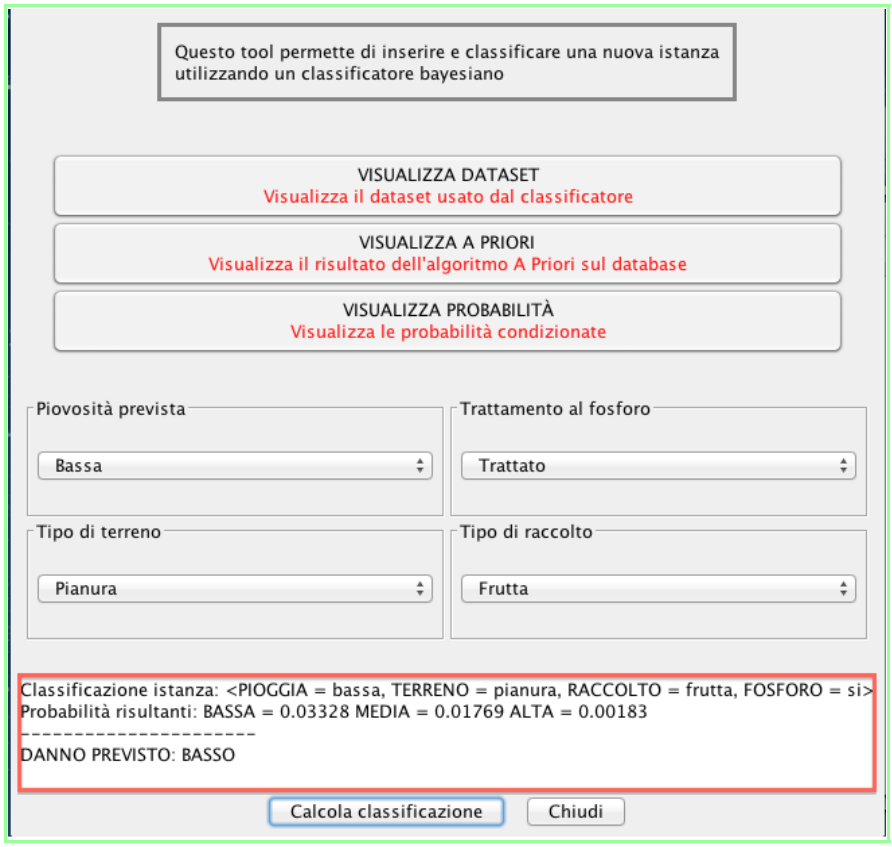
Probabilità condizionata

#	Probabilità	Valore calcolato
0	P(raccolto = tabacco | danno = basso)	0.363
1	P(geografia = costa | danno = basso)	0.583
2	P(raccolto = tabacco | danno = medio)	0.263
3	P(geografia = costa | danno = medio)	0.332
4	P(raccolto = tabacco | danno = alto)	0.185
5	P(geografia = costa | danno = alto)	0.222
6	P(geografia = collina | danno = medio)	0.377
7	P(geografia = collina | danno = basso)	0.183
8	P(pioggia = alto | danno = basso)	0.203
9	P(geografia = collina | danno = alto)	0.349
10	P(pioggia = alto | danno = medio)	0.187
11	P(pioggia = alto | danno = alto)	0.131
12	P(raccolto = legumi | danno = basso)	0.183
13	P(raccolto = legumi | danno = medio)	0.187
14	P(fosforo = si | danno = alto)	0.131
15	P(fosforo = no | danno = basso)	0.063
16	P(raccolto = legumi | danno = alto)	0.358
17	P(fosforo = si | danno = medio)	0.652
18	P(fosforo = si | danno = basso)	0.923
19	P(geografia = pianura | danno = alto)	0.431
20	P(geografia = pianura | danno = medio)	0.294
21	P(fosforo = no | danno = medio)	0.347
22	P(fosforo = no | danno = alto)	0.867
23	P(geografia = pianura | danno = basso)	0.223

La parte inferiore dell'interfaccia consente di creare una nuova istanza scegliendo i valori delle diverse features dai menù a tendina; cliccando poi sul pulsante *Calcola Classificazione,* nell'area di testo verranno visualizzati i valori delle probabilità dell'istanza di appartenere alle tre classi *Danno Alto*, *Danno Medio* o *Danno Basso*, ed il danno finale previsto (la probabilità con valore maggiore):



Esempio di una seconda classificazione:



Come detto in precedenza, l'algoritmo Apriori calcola soltanto i frequent itemset. Da questi, le associations rules devono essere ricavate manualmente.

Mostriamone un esempio.

Consideriamo i seguenti itemset (ricordiamo che il supporto usato è pari a 20):

<[1]>, <[2]>, <[4]>, <[5]>, <[6]>, <[7]>, <[8]>, <[9]>, <[10]>, <[11]>, <[12]>, <[14]>, <[15]>

<[1][11]>, <[1][12]>, <[1][14]>, <[2][11]>, <[2][12]>, <[2][14]>, <[2][15]>, <[4][14]>, <[11][15]>, <[12][14]>

<[2][11][15]>, <[2][12][14]>

Ogni feature del database è codificata con un numero intero, quindi abbiamo che:

```
piovosità bassa = 1 , piovosità media = 2 , piovosità alta = 3
costa = 4 , collina = 5 , pianura = 6
cereali = 7 , frutta = 8 , tabacco = 9 , legumi = 10
non trattato = 11 , trattato = 12
danno basso = 13 , danno medio = 14 , danno alto = 15
```

Tralasciando gli itemset unari, possiamo generare le seguenti regole:

1 → 11	11 → 1
1 → 12	12 → 1
1 → 14	14 → 1
2 → 11	11 → 2
2 → 12	12 → 2
2 → 14	14 → 2
2 → 15	15 → 2
4 → 14	14 → 4
6 → 15	15 → 6
11 → 15	15 → 11
12 → 14	14 → 12
2 → 11 15	11 15 → 2
11 → 2 15	2 15 → 11
15 → 2 11	2 11 → 15
2 → 12 14	12 14 → 2
12 → 2 14	2 14 → 12
14 → 2 12	2 12 → 14

Si può subito osservare che le feature più presenti sono quelle riguardanti il **danno**, la **piovosità** ed il **trattamento al fosforo**, quindi ci aspettiamo una correlazione tra questi elementi.

Il calcolo della confidenza non è stato riportato in questo lavoro, possiamo però fare delle considerazioni generali sulle regole appena trascritte.

Abbiamo evidenziato quelle regole che implicano in qualche modo la presenza di un danno, che è la cosa che più ci interessa.

Osserviamo che un danno medio si ha con piovosità bassa o media, nelle coltivazioni costiere e se si tratta il terreno con il fosforo.

Un danno alto invece viene osservato con piovosità medie e se non si tratta il terreno con il fosforo.

Queste considerazioni sono coerenti con le ipotesi fatte durante lo studio del problema.

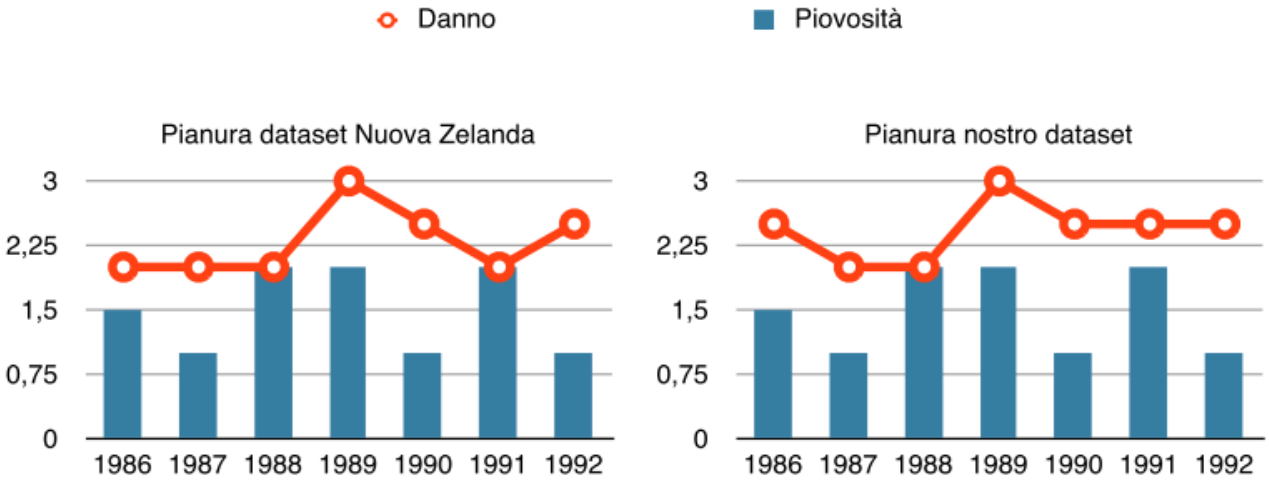
Risultati e osservazioni

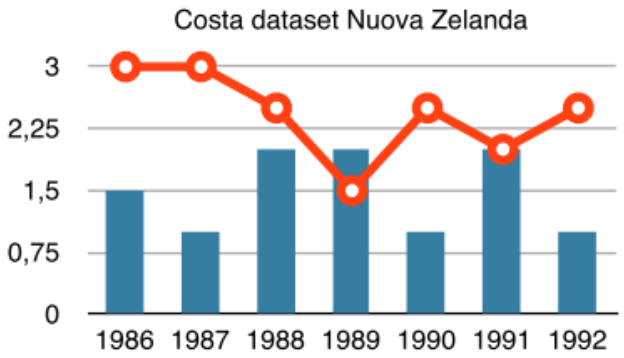
Al fine di verificare la **correttezza** del nostro algoritmo abbiamo contattato diverse **università neozelandesi** chiedendo ai dipartimenti di biologia, ecologia e agricoltura materiale relativo a studi passati che potesse darci un riscontro sulla verosimilitudine dei dati da noi creati.

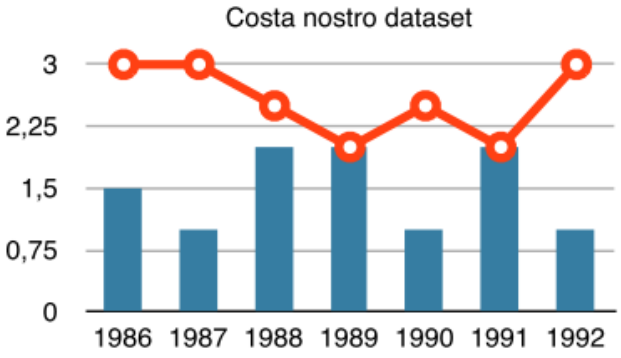
La **Lincoln University**, la **Waikato University** e la **University of Otago** ci hanno cordialmente risposto inviandoci diversi articoli scientifici **[8] [9] [10] [11] [12] [13]**.

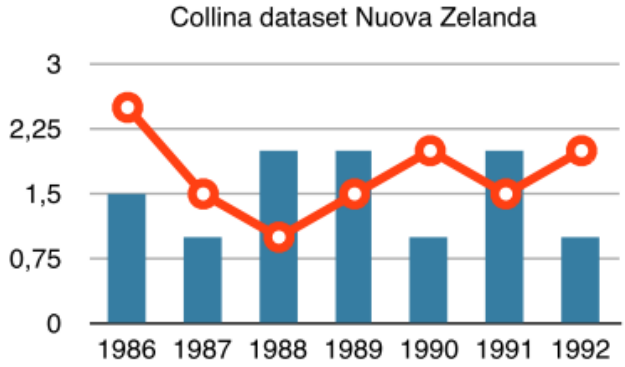
In particolare, il prof. R. J. Townsend della Lincoln University ci ha inviato un dataset (in allegato) che evidenzia il danno effettuato dai grass grubs in terreni collinari, costieri e pianeggianti dal 1986 al 1992. Abbiamo messo in relazione questi dati con la piovosità registrata nel database climatico neozelandese negli stessi anni.

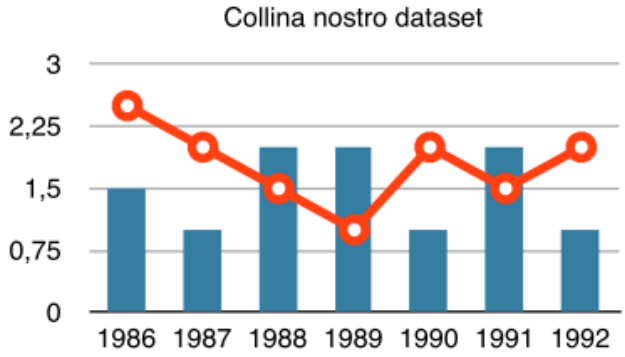
Abbiamo modificato i **pesi** del nostro algoritmo affinchè il dataset generato si adattasse ai dati forniti dall'Università di Lincoln. Di seguito vediamo dei grafici che mostrano le differenza tra la nostra classificazione e quella neozelandese divisa per territorio (abbiamo assegnato una scala di valori al danno: `alto = 3`, `medio-alto = 2.5`, `medio = 2`, `medio-basso = 1.5`, `basso = 1`).









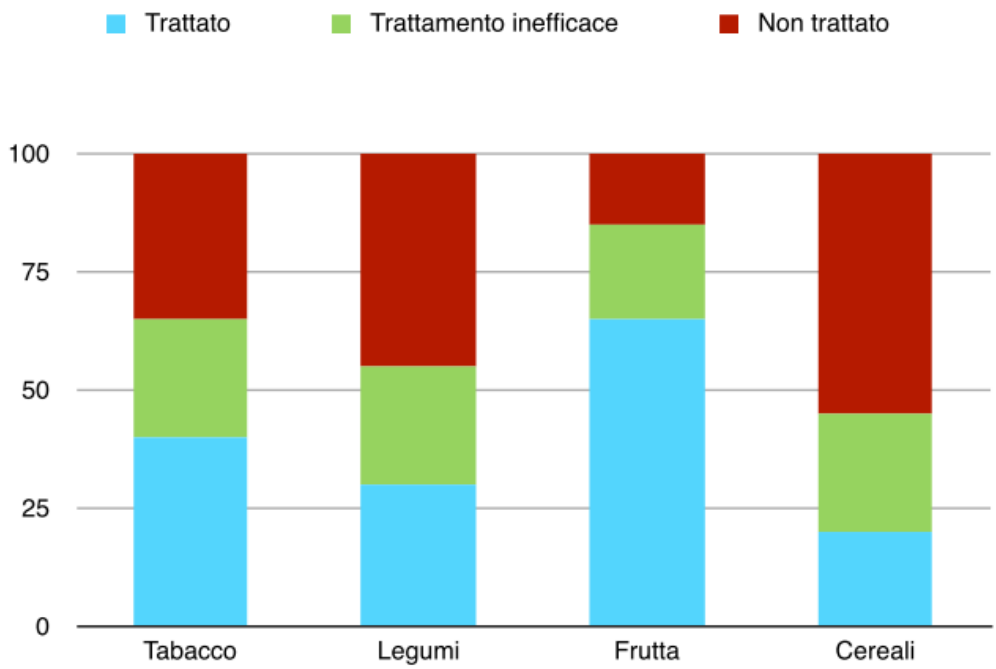


Osserviamo come le classificazioni non sono identiche, ma molto simili. Ci siamo ritenuti soddisfatti del risultato, ed abbiamo quindi esteso la classificazione anche agli anni dal 1980 al 1985 e dal 1993 al 2013.

Le altre features da noi considerate (tipo di raccolto e trattamento al fosforo) sono state aggiunte secondo i criteri descritti nei vari articoli analizzati.

Aggiungiamo inoltre che, una volta creato il dataset completo, lo abbiamo inviato alle università precedentemente citate per un riscontro. Ancora una volta il prof. R. J. Townsend dell'Università di Lincoln ci ha risposto dicendo che le assunzioni fatte possono essere considerate abbastanza veritiere.

Una interessante **osservazione** derivante dall'esecuzione dell'algoritmo è la seguente.



Il grafico mostra la **percentuale di danno** raggiungibile in base al tipo di coltura e al trattamento al fosforo.

La parte azzurra indica la percentuale massima di danno raggiungibile in seguito ad un trattamento al fosforo andato a buon fine. La parte verde indica la percentuale massima di danno raggiungibile in seguito ad un trattamento effettuato in un periodo molto piovoso (come abbiamo specificato in precedenza, se piove molto il trattamento ha scarsa efficacia). La parte rossa indica la situazione peggiore, ovvero il danno massimo che si ottiene quando la coltura non viene trattata.

Riportiamo alcuni test effettuati usando il nostro algoritmo per classificare le seguenti istanze:

Coltura = **frutta**, Terreno = **costa**, Trattamento = **NO**, Piovosità = **alta**
→ **danno** = **ALTO**
Coltura = **frutta**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **alta**
→ **danno = BASSO**
Coltura = **frutta**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **media**
→ **danno = MEDIO**

Sia dal grafico che dalle classificazioni si evince che la **frutta** è la coltura più resistente, poiché è molto difficile ottenere un danno alto (lo otteniamo soltanto nella configurazione peggiore, ovvero la prima istanza nell'esempio). Notiamo inoltre che, con la sola aggiunta del trattamento, il danno

viene ridotto drasticamente. La terza istanza mostra uno dei pochi casi in cui si ottiene un danno medio, poiché è molto più probabile ricadere all'interno della zona azzurra del grafico e quindi avere un danno basso.

Analizziamo ora i **cereali**.

Coltura = **cereali**, Terreno = **costa**, Trattamento = **NO**, Piovosità = **media**
→ **danno** = **ALTO**
Coltura = **cereali**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **media**
→ **danno = MEDIO**

In questo caso la prima istanza rappresenta il caso peggiore in assoluto per il tipo di coltura, ed infatti si ha un danno alto. Notiamo che neanche in seguito ad un trattamento il danno riesce ad abbassarsi del tutto, restando nella zona verde del grafico.

Risultati analoghi si ottengono con i **legumi**:

Coltura = **legumi**, Terreno = **costa**, Trattamento = **NO**, Piovosità = **media**
→ **danno** = **ALTO**
Coltura = **legumi**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **media**
→ **danno = MEDIO**
Coltura = **legumi**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **bassa**
→ **danno = MEDIO**

In questo caso, anche se la piovosità è minore e presenta una condizione sfavorevole allo sviluppo delle larve, il danno resta medio nonostante il trattamento.

Infine analizziamo il **tabacco**.

Coltura = **tabacco**, Terreno = **costa**, Trattamento = **NO**, Piovosità = **media**
→ **danno** = **ALTO**
Coltura = **tabacco**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **media**
→ **danno = MEDIO**

Coltura = **tabacco**, Terreno = **costa**, Trattamento = **SI**, Piovosità = **basso**

→ **danno = BASSO**

Questa coltura si pone tra i legumi e la frutta in termini di resistenza alle infestazioni. Infatti notiamo che trattando il terreno il danno da alto diventa medio; se però si ha una condizione climatica sfavorevole allo sviluppo delle larve, il danno si abbassa ulteriormente.

Conclusioni e sviluppi futuri

La parte più interessante del progetto è stata la creazione del dataset, un aspetto che a lezione e negli homework non avevamo affrontato.
Abbiamo potuto osservare quanto sia importante il processo di ricerca e analisi dei dati, che non sempre sono coerenti o utilizzabili ai fini del progetto che si sta realizzando.
Altro aspetto non trascurabile è la dimensione del dataset: calcoli più attenti ci hanno portato a modificare i primi prototipi del programma poiché non contenevano abbastanza istanze, ed ottenevamo classificazioni molto approssimative.

Il risultato ottenuto può essere considerato soddisfacente: grazie anche al risconto con i dati delle università neozelandesi, il nostro programma classifica in modo coerente le nuove istanze, e inoltre è stato possibile determinare le cause primarie dello sviluppo dei grass grubs.
Gli obiettivi del progetto sono quindi stati raggiunti.

Possiamo evidenziare i seguenti punti come **sviluppi futuri**:

- poiché il fosforo è considerato un deterrente allo sviluppo delle larve, sarebbe interessante introdurre nell'insieme delle features la composizione chimica del terreno per analizzare il fenomeno in zone in cui il fosforo è già naturalmente presente;
- considerare il tasso di assorbimento sia di acqua che di fosforo del terreno;
- in alcuni studi si è evidenziato come il surriscaldamento globale abbia portato ad un aumento dello sviluppo delle larve nei terreni. Sarebbe interessante tener conto di questo trend nelle analisi dei prossimi 50 – 100 anni;
- una ottimizzazione computazionale potrebbe consistere nell'implementare l'algoritmo *FP-Growth* invece che Apriori per l'estrazione dei frequent itemset, che come abbiamo visto a lezione risulta essere più efficiente;
- una ulteriore miglioria potrebbe essere il calcolo automatico delle association rules e della relativa confidenza dopo aver generato i frequent itemsets.

Infine, vogliamo ringraziare le università neozelandesi di Lincoln, Otago e Waikato per la tempestiva e gentilissima collaborazione, e per la condivisione di dati, ricerche e articoli scientifici.

Bibliografia e fonti

[1] www.wikipedia.org

[2] R. East, P. D. King, R.N. Watson - *Population studies of grass grub and black beatle* - New Zealand Journal of Ecology 1981

[3] http://cliflo.niwa.co.nz/

[4] R. East, W. M. Kain - *Prediction of grass grub populations* - New Zealand Entomologist, 1982

[5] http://www.grass-man.com/faqs.html

[6] http://www.greenviewfertilizer.com/articles/fertilizer-facts

[7] http://www.omafra.gov.on.ca/english/crops/facts/08-017.htm

[8] Yeates GW 1991. *Impact of historical changes in land use on the soil fauna* - New Zealand Journal of Ecology 15: 99-106.

[9] Jackson TA 1990. *Biological control of grass grub in Canterbury*. Proceedings of the New Zealand Grassland Association 52: 217-220.

[10] McLennan JA, Pottinger RP 1976. *Mortality of grass grub, Costelytra zealandica (White), and earthworms (Lumbricidae) during autumn cultivation*. New Zealand Journal of Agricultural Research 19: 257-263.

[11] Stewart KM, Archibald RD 1987. *The effects of pasture management on population density and diseases of porina (Lepidoptera: Hepialidae)*. New Zealand Journal of Experimental Agriculture 15: 375-379.

[12] Stewart, K.M. 1986. Control of grass grub (Costelytra zealandica) by cultivation in spring or summer. New Zealand Journal of Agricultural Research 14: 83-87

[13] Lefort MC, Worner SP, De Romans S, Armstrong K, Glare TR, Boyer S 2014. *Invasion success of a scarab beetle within its native range: host range expansion vs. host-shift*. PeerJ 2:e262; DOI 10.7717/peerj.262: e262.