Mineria de dades aplicada a la meteorologia (Febrer 2015)

Marc Tula – ls25787 Jordi Farràs – ls25887

***Abstract*—En aquest treball es presenta una explicació de les tècniques de mineria de dades aplicades a la meteorologia.**
Es veurà l'aplicació de clusterització amb Kmeans i amb xarxes neuronals (les SOM concretament).
Es veurà també com a partir d'aplicar regles a les agrupacions que es facin sobre als clusters es podran classificar les dades.

***Índex*—Durant el treball s'explicaran els següents punts: 1 –Introducció , 2- Aplicació , 3- El data set, 4- L'algorisme, 5- Procés, 6- Disseny Experimental, 7- Resultats 8-Conclusions 9-Referències**

I. INTRODUCCÓ

L'elecció d'aquest treball es deu entre altres aspectes a que un dels dos integrants del grup té un projecte de donar la volta al món en veler. En un trajecte d'aquestes dimensions amb un vaixell de 17 metres és vital tenir controlat tot canvi de meteorologia.

És per aquest motiu que aquest treball abordarà el tractament de les dades de meteorologia, com s'analitzen els diferents casos i sota quines condicions la màquina incorporada al vaixell avisa a la tripulació de que cal un canvi de rumb degut a que s'entrarà en una zona amb climatologia no convenient.

Entrant en matèria, la meteorologia és l'estudi científic interdisciplinari de l'atmosfera on s'observa els canvis en la temperatura, la pressió atmosfèrica, la humitat i la direcció del vent. En general, la temperatura, la pressió, mesuraments de vent i la humitat són variables que es mesuren amb un termòmetre, baròmetre, anemòmetre... però també hi ha altres mètodes de recollida de dades com els satèl·lits.

Universitat Ramon Llull La Salle (e-mail: tulaguardiolamarc@gmail.com-jfarras8@hotmail.com).

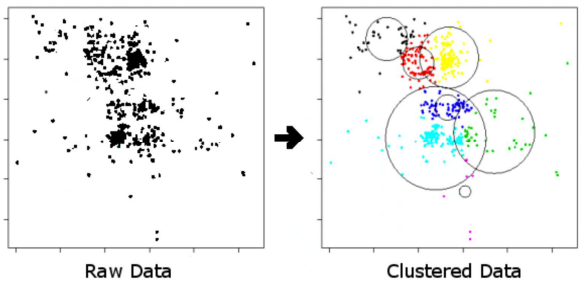
Existeixen diferents rutes on s'experimenten tot tipus de climatologies que poden posar en risc el trajecte i a la tripulació. És per això que vaixell porta incorporats diversos dispositius que analitzen en tot moment la climatologia propera.
- Radar per a detectar tant possibles col·lisions com canvis a la meteorologia propera.
- Ordinadors amb cartes nàutiques electròniques.
- GPS
- Comunicacions per satèl·lit

El radar que incorporen aquests vaixells és un sistema d'ones electromagnètiques per mesurar distàncies, altituds, direccions, velocitats i formacions meteorològiques.
Aquests radars es caracteritzen per tenir una molt bona precisió alhora de detectar possibles col·lisions o climatologies perilloses. No obstant això aquests radars són poc configurables (distància a analitzar, interfície visual, etc..).
Si el projecte s'acabés materialitzant en un producte al mercat s'intentaria millorar la interfície visual del radar, permetre personalitzar aspectes de configuració com la distància màxima o amb quina intensitat ha de sonar l'alarma, repeticions d'aquesta per seguretat ... amb una precisió igual o semblant a la dels radars actuals. A més també es miraria d'oferir d'un avís de quan es pot anar a més velocitat degut a que hi ha una ruta segura o de que cal disminuir la velocitat per entrada a ports o rutes amb velocitat màxima.

Els pronòstics del temps es fan mitjançant la recopilació de dades quantitatives sobre l'estat actual de

l'atmosfera però el principal problema que es planteja en aquesta predicció és que tracta amb grans volums d'informació que cal simplificar abans de fer l'anàlisi.

En aquest treball s'explicaran diverses tècniques que es poden aplicar al tractament dels grans volums de dades meteorològiques.



Exemple de clustering

II. APLICACIÓ

Caldrà fer **clustering** ja que és una eina poderosa que s'ha utilitzat en diverses tasques de previsió.
Caldrà també fer una classificació per poder discernir si es pot navegar o no en funció de les diferents agrupacions que hagi generat el Kmeans mitjançant el software Weka.
Per tal de poder fer els clusters caldrà que les diferents dades que es compararan siguin de les **mateixes estacions, data i hora.** Es presenten dues opcions per a fer les classes, utilitzar el coneixement dels experts o utilitzar regles específiques en aquest context que es puguin fixar, per cada tipus de característica que es vol analitzar: potència del vent, direcció, temperatura...

III. DATA SET

Cada continent gestiona les dades meteorològiques de la seva regió.

Concretament són **dades continues** ja que entre valor i valor hi ha infinits valors possibles i cal processar-les per a ser interpretades pel WEKA ja que hi ha un fitxer per cada tipus de característica que es vol analitzar : intensitat del vent, direcció, temperatura...

```
15-22 DATE : Date YYYYMMDD
24-28 FG   : Wind speed in 0.1 m/s
30-34 Q_FG : quality code for FG (0='valid'; 1='suspect'; 9='missing')

This is the blended series of station MAASTRICHT, NETHERLANDS (STAID: 168)
Blended and updated with sources:13661 906380
See files sources.txt and stations.txt for more info.

STAID, SOUID,    DATE,   FG, Q_FG
   168, 13661,19530101,   41,    0
   168, 13661,19530102,   36,    0
   168, 13661,19530103,   21,    0
   168, 13661,19530104,   51,    0
```

Exemple de fitxer recopilatori de dades d'intensitat del vent

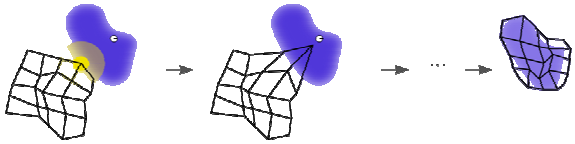
Es veu també que caldrà simplificar el projecte havent d'utilitzar dades històriques per evitar costos afegits com el fet d'haver de fer petites transaccions per cada consulta.

Els fitxers tenen diferents valors de qualitat, 0=vàlid, 1= Entra dintre el rang de valors possibles però no te sentit (Exemple: -2 graus de temperatura a Sevilla a l'estiu) i -9999 si el valor és erroni, es surt del rang vàlid. Dir també que al ser un **domini no confidencial** no hi haurà cap tipus de problema amb temes de legislació i protecció de dades.

IV. ALGORISMES

Caldrà comparar el rendiment de les agrupacions amb diferents algorismes:
Som:

Aquest és un dels models de xarxes neuronals més populars, el que és especialment adequat per a l'alta dimensió de visualització de dades, clustering i el modelatge.
S'utilitza un **aprenentatge no supervisat** per a la creació d'un conjunt de vectors prototips que representen les dades. El SOM es va introduir a les ciències meteorològiques i climàtiques de finals de 1990 com un mètode de reconeixement de dades.



.

Exemple d'aprenentatge mapes SOM

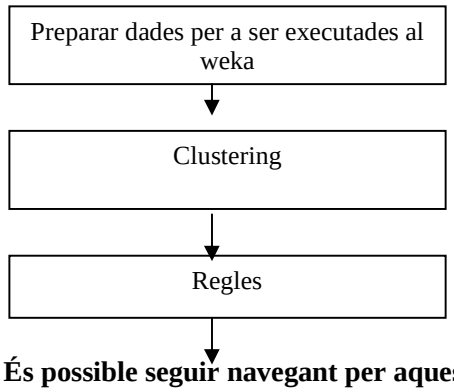
A causa de que els mapes preserven les relacions de veïnatge de les dades d'entrada, el SOM és una tècnica de preservació de la topologia i permet que aquests mapes auto-organitzats, amb diferents parts de la xarxa responguin similarment a certs patrons de l'entrada.

Es provarà també d'utilitzar un dels referents del clustering, el **k-means** proposant diferents k i també l'algorisme **x-means** el quan proposa una k a partir del input.
Pel que fa a les regles s'ha estat investigant l'algorisme **apriori** que es fa servir, per exemple, per a saber a partir del que s'ha comprat en un supermercat quins productes es compren junts així com també a bases de dades mèdiques per descobrir si ocorren comunament malalties

entre grups de persones. Finalment s'ha trobat més adient fer unes regles ad-hoc.

V. PROCÉS

Primer de tot caldrà fer una **recopilació de dades.** Per al context explicat de la navegació, a priori podria semblar que les dades rellevants per a poder dur a terme el projecte seran la intensitat del vent(m/s), la direcció d'aquest(graus), el nivell del mar(en hp) i el nivell de precipitació(mm). Tot i així **al no ser experts en la matèria** s'inclouen altres dades que ofereixen les estacions que puguin ser reveladores de nou coneixement, com ara el cobriment dels núvols(octas) o la humitat(%).
Com era extremadament complicat fer dades de tot el món s'ha acabat simplificant i s'ha analitzat tot a partir de la mateixa estació(ID. 0080032) situada en un estret de Groenlàndia.
Un cop trobats els fitxers es farà una petita modificació per tal d'intentar adaptar les dades a format WEKA, és a dir, en un únic fitxer arff.
En el weka es faran proves de clustering amb els diferents algorismes exposats anteriorment i posteriorment s'aplicaran regles per a classificar.
Caldrà estudiar la possibilitat d'usar un predictor de regles i posteriorment vindrà la part més important ,extreure conclusions i assegurar que les agrupacions tenen sentit.



És possible seguir navegant per aquest rumb?

VI. DISSENY EXPERIMENTAL

Un cop s'hagin generat els clusters amb els diferents algorismes i testejat quin aplica millor en aquest domini, caldrà veure com els nous casos que entrin serien classificats en aquest sistema classificador.
Al no tenir dades classificades seria una tasca molt complexa i només es podria mirar als nous casos que entren, la distància que sigui menor a un dels **centroids dels clusters**.

Pel que fa a les regles, es faran ad-hoc i amb les SOM es provarà de posar un valor 2 d'altura i d'amplada que són els que venen per defecte.

VII. RESULTATS

S'han parsejat els fitxers fonts per tal d'adaptar-los a un nou fitxer de Weka mitjançant les comandes **cat, awk i grep** del terminal per tal de quedar-se amb les columnes desitjades:

```
@attribute intensity real
@attribute direction real
@attribute seaLevel real
@attribute pluja real
@attribute humitat real
@attribute cloudCover real
@data
    98,   20,10168 ,   22 ,   84 ,    8
    38,  340,10104 ,   70 ,   87 ,    8
   113,   30,10068 ,   16 ,   96 ,    8
    38,    0,10020 ,    0 ,   62 ,    4
```

Concretament s'han agafat **500 casos**.

A continuació es mostraran els resultats d'executar el fitxer usant **Xmeans** com algorisme en el Weka. Com es veu ha retornat 4 clusters del total de 6 atributs entrats.

```
=======
Cutoff factor                 : 0.5
=======

Cluster centers               : 4 centers

Cluster 0
        50.94736842105263 15.263157894736842 10153.868421052632 10024.631578947368 72.15789473684211 89.55263157894737
Cluster 1
        38.57597173144876 25.26501766784452 10076.144876325088 35.685512367491164 -59.08480565371025 -29.314487632508833
Cluster 2
        88.31746031746032 279.5238095238095 10102.380952380952 9955.587301587302 51.492063492063494 79.4920634920635
Cluster 3
        63.8448275862069 283.1034482758621 10093.318965517241 32.63793103448276 -17.870689655172413 3.836206896551724

Distortion: 117.151167
BIC-Value : 1401.491122

Time taken to build model (full training data) : 0.06 seconds
```

En la captura anterior es pot observar la sortida ja clusteritzada amb el Xmeans. S'observa la informació obtinguda per cada cluster (mitjana). Així doncs, en el primer cluster ens està donant que la **mitjana** d'intensitat del vent és de 50.94 m/s, 15.23 graus de direcció, 10153 hp (hectopascals) de pressió de nivell de mar ,10024 mm de pluja, una humitat de 72 (en 1%) i ocupació de núvols de 89,5 octas.

Es veu també que com a sortida hi han 38 instàncies al cluster 0, 282 al cluster 1,63 al cluster 2 i 117 al cluster 3. Dir també que poden aparèixer alguns valors negatius degut a que dins les mostres hi havia alguns valors amb -9999 que no s'han pogut eliminar i si ha quedat un valor negatiu significa que majoritàriament hi havien valors baixos.

Amb tot això els centroides dels clusters serien:

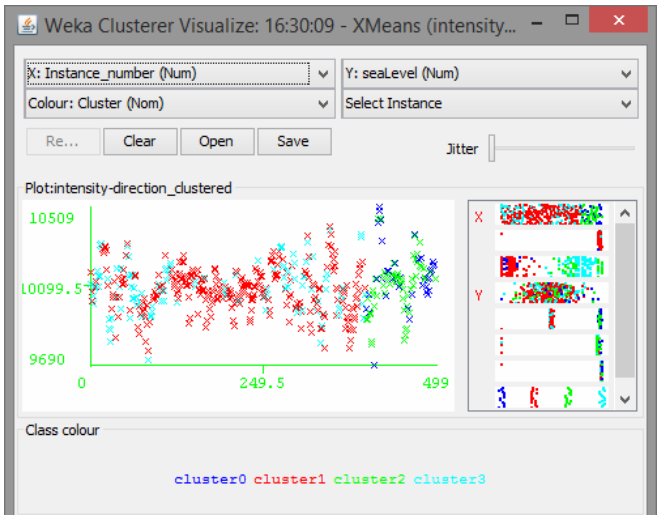
CLUSTER 0;(3401,07)fort vent del nord, pluja alta, molt núvol.
CLUSTER 1:(1674) calmat, poc núvol.
CLUSTER 2:(3475)vent extrem del oest, pluja forta, molt núvol.
CLUSTER 3:(10478,56)fort vent oest, pluja moderada, cel obert.

Es veu com les agrupacions tenen sentit ja que per exemple quan hi ha poc núvol hi ha poca precipitació. També es pot veure com a la zona en que esta situada l'estació hi ha perill en navegació amb direcció del vent nord i oest. Ha estat important posar dades que en un inici no es consideraven importants com la quantitat de núvols ja que permet classificar els cassos ràpidament. com es veurà en les regles.
També es veu com abunda el mal temps degut a la zona geogràfica de l'estació.

Si per exemple entrés el cas:

- " 125, 300, 10004 , 0 , 54 , 1"

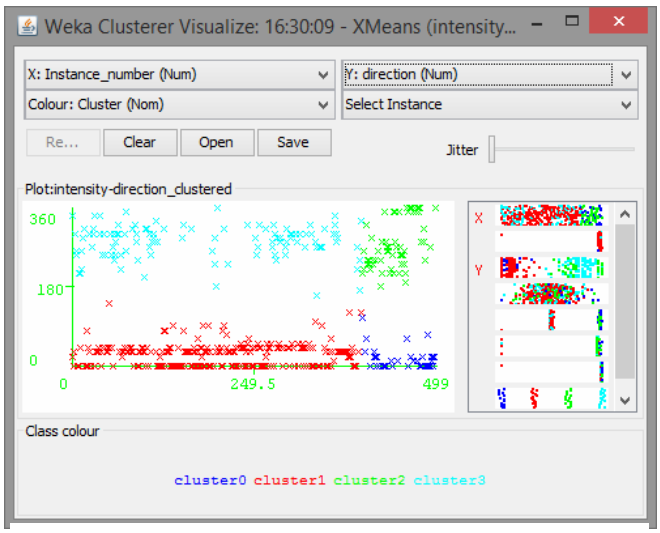
Fent la mitjana (10484 / 6 =1747,3) s'agruparia en el cluster1 degut a que amb el conjunt de característiques és el que té la distancia més pròxima tot i que es podrien fer servir altres criteris.



Clustering amb eix Y de nivell del mar Xmeans

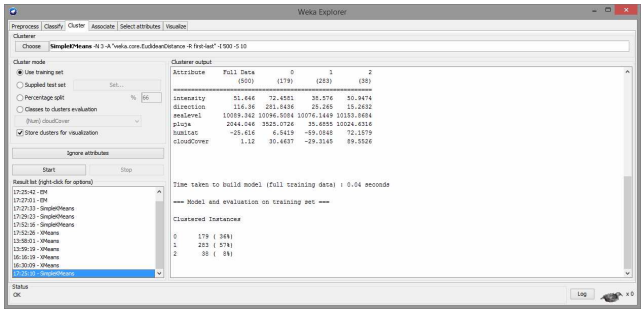
Analitzant amb Xmeans l'eix x el nombre d'instàncies i a l'eix Y el nivell del mar es pot apreciar a la captura superior com obtenim un resultat que no té els clusters ben cohesionats amb molta dispersió entre ells degut a que és una característica poc rellevant en la classificació.

D'igual forma passa amb la humitat tal i com es preveia a l'hora de triar quines dades es farien servir.

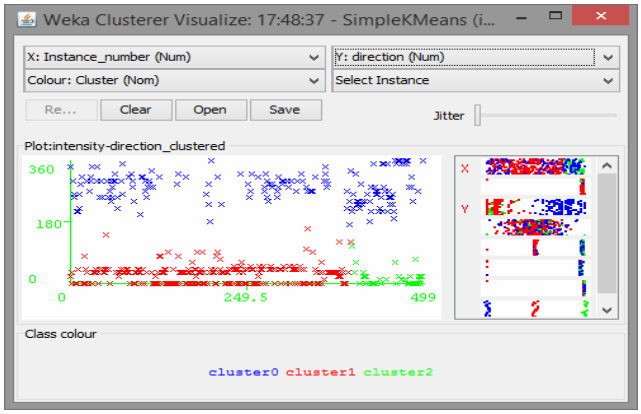


Clustering amb eix Y de direcció del vent amb x-means

La direcció del vent en canvi es clau per determinar el cluster tant amb Xmeans com amb Kmeans. En aquest últim, es fixa el nombre de clusters a 3 per tal de veure el comportament de l'algorisme si té menys opcions on agrupar les dades. A continuació es mostren els resultats d'aplicar el Kmeans sobre el fitxer amb 3 clusters.



A la captura següent es pot observar com no s'altera que els vents vinguin en direcció nord o oest. En aquest cas es classifiquen 179 instàncies al cluster 0, 283 al cluster 1 i 38 al cluster 2.

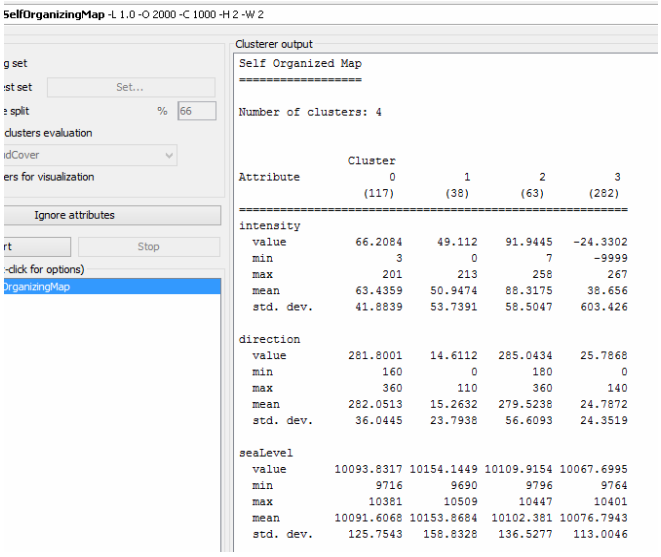


Clustering amb eix Y de direcció del vent amb 3-means

Ara bàsicament segueix havent un cluster per el bon temps (cluster 1) i s'hi posen 283 casos a diferència d'abans on al cluster 1 s'hi posaven 282.

Per tant, la principal diferència rau en agrupar els casos en condicions de mal temps ja que ara no existeix la possibilitat d'avisar únicament de pluja i no de ventada ja que els dos clusters restants, tenen les dos coses i majoritàriament s'agrupen en el 1r dels dos (de mal temps).
Amb 3 clusters també veiem com a partir de 160 graus sempre classifica igual totes les instàncies a diferència d'abans, que les posava al cluster 2 i al cluster3.

També s'ha provat de comparar el rendiment amb un altre tipus de clustering no supervisat, les **SOM**.




```
Self Organized Map
==================

Number of clusters: 4


                Cluster
Attribute            0          1          2          3
                 (117)       (38)       (63)       (282)
=============================================================
intensity
  value          66.2084    49.112    91.9445   -24.3302
  min                  3         0          7       -9999
  max                201       213        258        267
  mean           63.4359   50.9474    88.3175     38.656
  std. dev.      41.8839   53.7391    58.5047    603.426

direction
  value         281.8001   14.6112   285.0434    25.7868
  min                160         0        180          0
  max                360       110        360        140
  mean           282.0513   15.2632   279.5238    24.7872
  std. dev.       36.0445   23.7938    56.6093    24.3519

seaLevel
  value       10093.8317 10154.1449 10109.9154 10067.6995
  min               9716       9690       9796       9764
  max              10381      10509      10447      10401
  mean        10091.6068 10153.8684 10102.381 10076.7943
  std. dev.     125.7543   158.8328   136.5277   113.0046
```

Es pot veure que com amb alçada 2 ha fet 4 clusters on per cada un, desglossa cada característica amb el seu valor amb diferent informació com el màxim, mínim o desviació estàndard.

Concretament pel que fa a la intensitat del vent, per exemple, es veu com es fa un cluster on abunden dades errònies que tenia l'estació a diferencia dels algorismes vistos anteriorment.
Amb el valor extremadament elevat de la desviació estàndard ja s'avisa que hi ha molta variació.

Dir també que en el cas de fer alçada 4 l'algorisme proposa 5 clústers (veure captura següent).

```
Scheme:       weka.clusterers.SelfOrganizingMap -L 1.0 -O 2000 -C 1000 -H 3 -W 2
Relation:     intensity-direction
Instances:    500
Attributes:   6
              intensity
              direction
              seaLevel
              pluja
              humitat
              cloudCover
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


Self Organized Map
==================

Number of clusters: 6


                Cluster
Attribute            0          1          2          3          4          5
                 (37)       (265)       (30)       (20)       (34)       (114)
=============================================================================
intensity
  value          49.2477   -25.1913    67.0525    41.7071   107.5461    66.9246
  min                  0       -9999          7          8          7          3
  max                213         267        132        158        258        201
  mean           51.1622    38.5019    66.6333       40.9   106.1176    64.0526
  std. dev.      54.4638    622.4798   32.6523    36.9849    69.2972    41.9922
```

Amb aquest últim cas al haver més clusters es veu com al nostre context anava millor tenir classificat en menys perquè calia avisar només en condicions de perill i no situacions mixtes. Referent als mapes SOM (o Kohonen) s'ha constatat com el **Weka no és capaç de generar-los.**

Dit tot això, en el cas de quedar-nos amb el Xmeans que proposa 4 clusters , unes **regles** ad-hoc a aplicar per a determinar si es pot navegar serien:

- **si** (direcció >= 0 & direcció<=20 || direcció >260 **llavors si** (intensitat > 30)
  llavors sona_alarma_perill()
- **si** (volum_núvols >70) **llavors** alarma_pluja
- **si** (pluja >900) **llavors** alarma_pluja
- **si** (direcció_actual != direcció) **llavors** notificació_augment_velocitat

Cal aclarir que per a imposar aquestes regles s'ha tingut en compte la opinió de diversos experts. Segons aquests experts navegar amb una intensitat de vent superior als 45 m/s (80 nusos) pot constituir un perill. Amb aquesta intensitat de vent per molt ben encarat que estigui el vaixell contra el vent hi ha perill de trencar veles, trencar màstil (encara que sigui de fibra de carboni) i en cas d'intentar navegar amb les veles desplegades hi ha perill

real de bolcar. A part de la intensitat, també cal tenir en compte la direcció del vent ja que de vegades, no es possible situar el vaixell en una posició còmode respecte el vent i en aquest cas si ens ve de cantó i no per proa (davant) o popa (darrera) pot esdevenir un perill.

S'ha investigat que els velers de vela lleugera (7 metres) pateixen amb vents superiors a 17-18 nusos. En aquest cas, però , el veler esta preparat per navegar còmodament amb 50 – 60 nusos, de 60 - 80 serà difícil no mullar-se i d' aquí cap endavant serà perillós navegar.

Un cop constituïdes les regles podríem tenir un sistema amb avisos diferents (treballant amb dades actuals i futures):

-Una primera alarma que avisés d'impossibilitat de navegar (turmenta + vent).

- Una segona alarma que avisés només en cas de forta pluja que, tal i com s'ha vist amb l'extracció d'informació no te perquè esta relacionat amb una gran ventada.

-Una notificació que permetés informar a la tripulació que es possible augmentar velocitat al no haver perill de turmenta.

-Una notificació que permetés informar a la tripulació que cal disminuir la velocitat ja que la ruta en la que ens trobem te marcada una velocitat màxima (entrada a port, canal estret, cala, etc..)

- Avisar en cas de que es detecti que el vaixell pot calar (tocar el terra) degut a que ens acostem a una zona amb poca profunditat (nivell del mar).

Comentar que un cop el sistema treballi amb un input real (dades usades a les estacions meteorològiques per predir el temps) l'alarma saltaria al detectar que amb el rumb actual ens apropem a unes condicions poc òptimes o si per cas d'error, no ha estat capaç de predir-ho i es posa a ploure o s'aixequen condicions poc òptimes sobre el vaixell.

(Dins del zip es troben les taules generades pel Weka).

VIII. CONCLUSIONS

Un cop acabat aquest treball s'ha pogut constatar que la part més complicada ha estat pensar exactament perquè s'utilitzarien les dades que s'analitzaran i quines són les que poden fer més servei.

Comentar també que tot i que en aquest treball s'han presentat "proves de joguina", aquests poden representar una aproximació al que seria realment un entorn de predicció en temps real que permetés decidir si s'ha de canviar de ruta de navegació o no.

Ha costat força adaptar les dades al weka però finalment s'ha vist com ha valgut la pena invertir temps en posar totes les dades possibles que oferien les estacions per tal de després adonar-se d'informació entre elles que estava oculta com el fet de que amb volum de núvols ja es determini tempesta.

S'ha vist també com el Xmeans és el que ha donat millor resultats degut a que ja analitza el millor valor per a que els clusters que es generin estiguin el màxim cohesionats possible.

Dir també que ha costat molt poder executar les SOM en el WEKA degut a que no es troben de forma nativa i ha calgut afegir externament un paquet per línia de comandes. També ha calgut canviar la versió que es tenia instal·lada per tal de que pogués funcionar.

S'han recercat alternatives per treballar amb les SOM i només s'ha trobat un complement per Matlab, però al no tenir-lo instal·lat es va trobar més adient afegir el paquet al Weka.

Per concloure, dir que aquest treball, tot i que ha estat feixuc en algunes parts ens ha servit per a saber utilitzar i triar quines són les eines que aplicaven al context d'aquest domini. Creiem sincerament que ens ha servit per reforçar coneixements que no havien acabat de quedar clars de forma teòrica.

Ha estat una bona manera d'assimilar conceptes de l'assignatura i a més ha servit també perquè en Marc tingui més coneixements dels que disposa actualment en aquest sector i quines opcions de millora hi haurien en cas que es volgués arribar a materialitzar aquest projecte.

Amb tot això, per concloure l'assignatura, fent una visió general de tot el que s'ha vist **podem afinar més en definir que és exactament la mineria de dades** on diríem que és aquell camp que estudia grans volums d'informació de domini totalment variat en que mitjanat diferents tècniques d'intel·ligència artificial, aprenentatge automàtic i estadística (el Bayes vist a classe per exemple) és capaç de trobar certs patrons i informació oculta a simple vista.

Creiem que ara tenim un coneixement bàsic de la majoria de tècniques que es poden aplicar i que seriem capaços d'aprofundir en algunes en concret si fos necessari en la nostre carrera laboral o fins i tot ens podrà fer entendre millor coses que fins ara no veiem com es feien, com el fet de trobar correlacions amb les compres que es fan a un supermercat. Per últim fer menció que creiem que ha estat interessant veure com

redactar correctament els treballs així com familiaritzar-nos amb papers professionals.

IX. REFERÈNCIES

MacQueen, J. “Some methods for classification and analysis of multivariate observations”. University of California Press, Berkeley, Calif., 1967.

Pelleg, D. and Moore, A. “X-means: Extending K-means with Efficient Estimation of the Number of Clusters”. Pages 727-734 San Francisco, CA, USA

Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, 100–108. JSTOR 2346830

R. Agrawal & R. Srikant
"Fast Algorithms for Mining Association Rules"
Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), pp 487-499
Morgan Kaufmann Publishers Inc. San Francisco, CA, USA
ISBN:1-55860-153-8
(1994)

Kohonen Maps in Weka, John Salatas,2009
http://weka.sourceforge.net/packageMetaData/SelfOrganizingMap/index.html

European weather data
http://eca.knmi.nl/dailydata/predefinedseries.php

Mineria de dades-Clustering(slides)

(Elisabet Golobardes i Ribé) La Salle, 2014.

Mineria de dades-Association rules(slides)

(Elisabet Golobardes i Ribé) La Salle, 2014.

Experts:

- Maties Pons Mir, membre del club nàutic es castell i del club nàutic de mahó. Campió de les illes Balears amb vela lleugera (làser, 5 metres).

https://members.sailing.org/biog.php?unique=1423072866.4255&includeref=membbiog&memberid=100012&js=1

http://www.sailracer.org/sailor_results.asp?sailor=Maties%20PONS%20MIR

- Enric Pons Mir, membre del club nàutic es Castell. 26è classifcat al mundial de vela lleugera (làser, 5 metres) 2014.

http://www.federacionbalearvela.org/es/default/regatistas/ficharegatista/id_sailor/1117

Universitat Ramon Llull La Salle (e-mail: tulaguardiolamarc@gmail.com-jfarras8@hotmail.com).