Universidade do Minho
Mestrado em Engenharia Informática

Perfil Sistemas Inteligentes
Aprendizagem e Extração de Conhecimento

Projeto de Extração de Conhecimento

WEKA

janeiro 2015

Figura 1: Ana Margarida Ferreira Cruz, pg27747

Figura 2: Isabel Maria Ferreira Cruz, pg27746

Figura 3: Serafim Miguel da Costa Pinto, pg28506

Resumo

No âmbito da Unidade Curricular de Aprendizagem e Extração de Conhecimento, do Perfil Sistemas Inteligentes do Mestrado em Engenharia Informática, este relatório pretende apresentar o Projeto de Extração de Conhecimento, adotando uma metodologia para a extração de conhecimento, realizado pelo grupo, bem como as decisões que foram tomadas, as suas etapas e dificuldades.

Neste trabalho foi utilizada a ferramenta para extração de conhecimento WEKA recorrendo à linguagem de programação Java, e um *dataset* que contém informação relativa a incêndios florestais no parque de Montesinho, situado no Nordeste Transmontano em Portugal.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

Após os diversos conhecimentos adquiridos através das aulas práticas da unidade curricular Aprendizagem e Extração de Conhecimento, este trabalho surge como o passo seguinte dessa aprendizagem de modo a colocar em prática esses mesmos conhecimentos.

Atualmente, verifica-se uma enorme oferta de dados na *Internet* através de fontes de dados abundantes em diversas áreas de conhecimento, tais como o comércio, a ciência e a sociedade. Muitas vezes, estes dados são de difícil análise para se obterem conclusões e/ou prever resultados futuros. Através da disponibilização de ferramentas automáticas, é possível analisar estes dados através do pré-processamento, da filtragem, da classificação, e de outros métodos de extração de conhecimento. Com esta análise, é possível encontrar resultados, determinar quais os resultados da análise que se apresentam como os mais relevantes, e interpretá-los de modo a decidir sobre a forma como estes poderão ser úteis no contexto dos dados analisados.

1.2 Caso de Estudo

Uma vez que os conhecimentos mencionados anteriormente se referem à preparação de dados e à extração de conhecimento, o primeiro passo consistiu na definição do caso de estudo. Assim, ele incidirá sobre os dados relativos aos incêndios florestais, mais concretamente no parque de Montesinho em Portugal, situado no Nordeste Transmontano. Este *dataset* foi retirado do repositório do WEKA, onde estão disponíveis vários *datasets* para o seu conhecimento ser extraído e analisado. Este caso de estudo pareceu ao grupo que se adequa ao tipo de trabalho que se pretende desenvolver uma vez que devido à área em que se insere, os dados recolhidos nem sempre estão completos e consistentes e nem todos os dados serão úteis.

Por fim, um facto importante que influenciou o grupo na escolha deste caso

de estudo, é a sua integração na área da Segurança, mais concretamente os Incêndios Florestais. Pois, cada vez mais em Portugal tem sido um problema grave para várias entidades ao longo dos anos. E portanto, a extração de conhecimento poderá ter um valor importante nesta área, com o objetivo de prevenir ou atenuar este problema que é considerado completamente real.

1.3 Objetivos

Tendo consciência que nos tempos atuais toda a informação é guardada em dados, e existem dados sobre qualquer temática, surge a necessidade de obter conhecimento útil dessa informação. Assim, o grande objetivo do grupo, para além de provar que consegue aplicar os conhecimentos obtidos nas aulas, é ser capaz de extrair conhecimento desconhecido e importante de um dado conjunto de dados. Os objetivos, em termos globais neste projeto, serão o desenvolvimento da capacidade de trabalho nesta área e a produção de resultados positivos e interessantes, assim como a aquisição de técnicas que permitam resolver variados problemas, mais concretamente nesta área dos Incêndios, na qual se insere este segundo trabalho prático.

1.4 Ferramentas

Para a realização da segunda parte do trabalho de grupo, foi necessário instalar a ferramenta WEKA[1]. Esta ferramenta tem como objetivo agregar algoritmos provenientes de diferentes abordagens/paradigmas na sub-área da inteligência artificial dedicada ao estudo da aprendizagem por parte das máquinas [1].

A linguagem Java caracteriza-se por ser uma programação orientada a objetos. O WEKA disponibiliza uma API em Java bastante poderosa e flexível que permite a sua integração em qualquer tipo de sistema baseado em Java. Esta API fornece um conjunto de classes e métodos de forma a utilizar todas as funcionalidades do WEKA.

1.5 Estrutura do Relatório

Numa primeira fase, é apresentado no relatório o Resumo deste trabalho prático, onde é descrito o problema e a justificação da sua solução.

De seguida é apresentada a Introdução, onde é exposto o caso de estudo e onde é feita uma pequena contextualização do problema, bem como os objetivos a alcançar com este segundo trabalho prático.

Seguidamente surge o Desenvolvimento onde serão apresentados os capítulos que explicam o que foi feito durante o trabalho, as decisões tomadas pelo grupo e os resultados obtidos.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Posteriormente é apresentada a Conclusão onde o grupo faz uma apreciação crítica sobre o trabalho prático, apontando os seus pontos fortes e fracos.

Por fim a Bibliografia é apresentada após a Conclusão, e nela é apresentada a lista das fontes bibliográficas consultadas durante a realização deste trabalho.

Capítulo 2

Desenvolvimento

2.1 Conjunto de dados: Incêndios Florestais

O grupo usou um *dataset* público (tópicos [8] e [9] da Bibliografia) que se encontra disponível para investigação. Os detalhes estão descritos em [Cortez e Morais, 2007]:

> *P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: http://www.dsi.uminho.pt/∼pcortez/fires.pdf*

Através da extração de conhecimento sobre os dados analisados, esperamos encontrar regras podem conduzir a uma maior prevenção dos incêndios florestais, isto é, identificar situações propícias à sua ocorrência.

2.1.1 Descrição do conjunto de dados

O *dataset* analisado contém 517 instâncias e 13 atributos. É de referir que vários dos atributos podem estar correlacionados, portanto, faz sentido aplicar algum tipo de seleção de recursos.

2.1.1.1 Atributos

No que diz respeito à informação dos atributos, estes são:

1. X - coordenada espacial do eixo x dentro do mapa do parque de Montesinho;
2. Y - coordenada espacial do eixo y dentro do mapa do parque de Montesinho;

3. month - mês do ano;
4. day - dia da semana;
5. FFMC - índice FFMC do sistema FWI;
6. DMC - índice DMC do sistema FWI;
7. DC - índice DC do sistema FWI;
8. ISI - índice ISI do sistema FWI;
9. temp - temperatura em graus Celsius;
10. RH - humidade relativa em percentagem;
11. wind - velocidade do vento em km/h;
12. rain - chuva em mm/m^2;
13. area - área ardida da floresta (em ha).

A sigla FWI significa o índice de clima de incêndios (*Fire Weather Index* - FWI). Neste contexto:

- FFMC (*Fine Fuel Moisture Code*) é uma classificação numérica do teor de humidade na superfície de terra e de outros combustíveis finos;
- DMC (*Duff Moisture Code*) é uma classificação numérica do teor de humidade média das camadas orgânicas vagamente compactadas de profundidade moderada;
- DC (*Drought Code*) é uma classificação numérica do teor de humidade das camadas profundas, compactas e orgânicas;
- ISI (*Initial Spread Index*) indica que a taxa de incêndio se irá propagar nas suas fases iniciais. Este índice é calculado a partir da classificação do índice FFMC e do fator vento.

2.1.2 Pré-processamento

De início, os atributos "rain" e "area" indicavam respetivamente a chuva exterior em mm/m^2 e a área ardida da floresta em ha. O grupo decidiu alterar o significado destes atributos para a existência ou não de chuva e área ardida, de modo a facilitar e clarificar o processo de extração de conhecimento.

- rain - existência ou inexixtência de chuva;
- area - existência ou inexistência de área ardida.

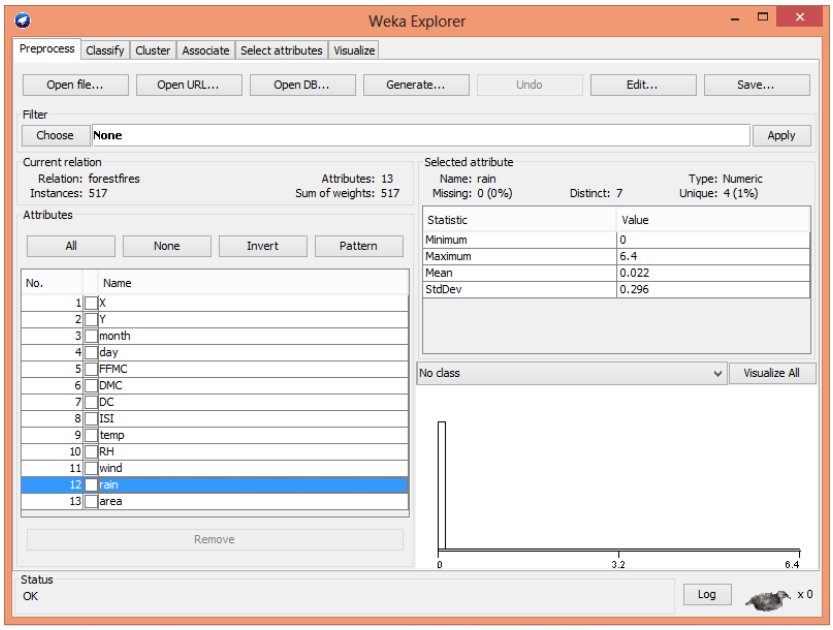

Figura 2.1: Atributo “rain” do *dataset* original

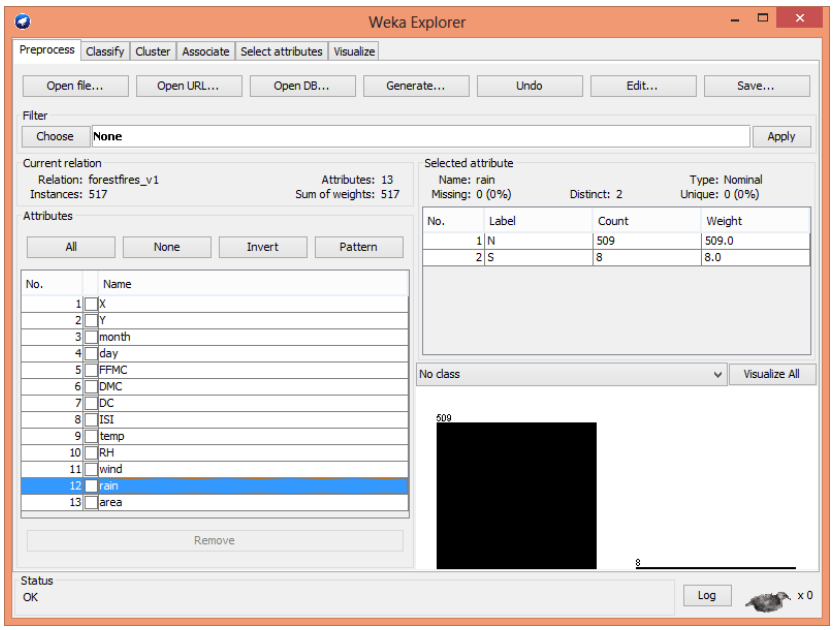

Figura 2.2: Atributo “rain” do *dataset* alterado

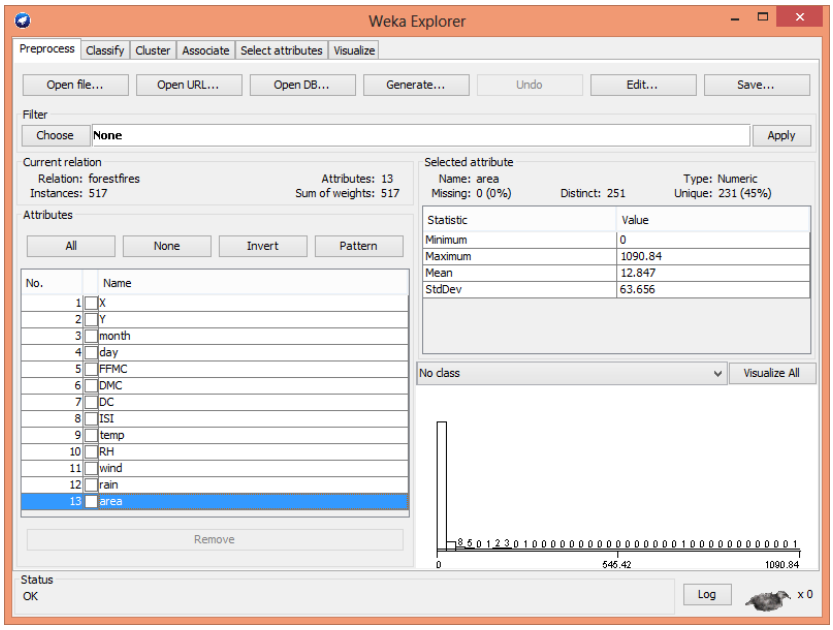

Figura 2.3: Atributo "area" do *dataset* original

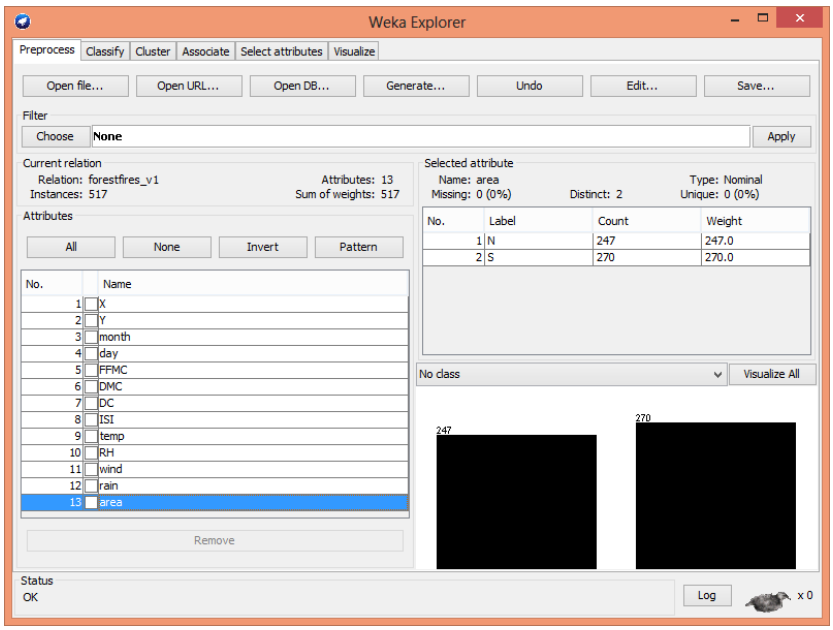

Figura 2.4: Atributo "area" do *dataset* alterado

No atributo "rain", considerou-se que se os valores fossem superiores a 0.0, então seria S (significa que choveu), caso contrário seria N (significa que não choveu). O modo de alteração do atributo "area" seguiu a mesma metodologia que o atributo "rain".

Alterados os atributos "rain" e "area", os tipos de dados encontrados no *dataset* dos incêndios florestais são de dois tipos: numérico e nominal.

- As variáveis numéricas são: X, Y, FFMC, DMC, DC, ISI, temp, RH e wind.

variável	**mínimo**	**máximo**	**média**	**desvio padrão**	**valores distintos**
X	1	9	4.669	2.314	9
Y	2	9	4.3	1.23	7
FFMC	18.7	96.2	90.645	5.52	106
DMC	1.1	291.3	110.872	64.046	215
DC	7.9	860.6	547.94	248.066	219
ISI	0	56.1	9.022	4.559	119
temp	2.2	33.3	18.889	5.807	192
RH	15	100	44.288	16.317	75
wind	0.4	9.4	4.018	1.792	21

Tabela 2.1: Variáveis do tipo numérico do *dataset* Incêndios Florestais

- As variáveis nominais são: month, day, rain e area.

variável	**valores**	**número de casos**	**valores distintos**
month	jan	2	12
	feb	20	
	mar	54	
	apr	9	
	may	2	
	jun	17	
	jul	32	
	aug	184	
	sep	172	
	oct	15	
	nov	1	
	dec	9	
day	mon	74	7
	tue	64	
	wed	54	
	thu	61	
	fri	85	
	sat	84	
	sun	95	
rain	S	8	2
	N	509	
area	S	270	2
	N	247	

Tabela 2.2: Variáveis do tipo nominal do *dataset* Incêndios Florestais

2.2 Processo de extração de conhecimento

2.2.1 Associação

Os objetivos das regras de associação para extração de conhecimento são:

- Encontrar padrões frequentes, associações, correlações ou estruturas ocasionais em conjuntos de dados (*datasets*);
- A descoberta de regras de associação é usada para encontrar elementos que ocorrem conjuntamente em *datasets*;
- Definição de regras de relacionamento (implicação ou correlação) entre elementos que ocorrem em comum.

O grupo decidiu procurar 10 regras de associação no *dataset* dos incêndios florestais. Através de testes realizados na ferramenta Weka, concluiu-se que o atributo *rain* estava a enviesar os resultados, e estes tornavam-se inconclusivos. Em 517 casos havia apenas 8 casos onde havia a ocorrência de chuva. Como resultado obtiveram-se várias regras de associação onde a implicação destas levava a concluir que não tinha chovido. Como a ocorrência de chuva surgia em 1.55% dos dados reportados, o grupo decidiu remover este atributo para assim obter regras de associação mais conclusivas.

Para equilibrar os dados, foi aplicado um filtro não supervisionado *discretize*. Este filtro discretiza um intervalo de atributos numéricos.

2.2.2 Segmentação

A Segmentação/*Clustering* de dados é um processo através do qual se particiona um conjunto de dados em segmentos/*clusters* de menor dimensão, que agrupam conjuntos de dados similares.

Para segmentar o *dataset* dos incêndios florestais, foi necessário efetuar o *discretize* dos dados no Weka, e na secção *Cluster* escolheu-se o algoritmo `SimpleKMean`.

Através da segmentação dos dados no Weka, o grupo definiu 5 *clusters* para serem gerados pela ferramenta.

2.2.3 Classificação

Sendo o WEKA uma ferramenta com capacidade de oferecer os mais variados mecanismos de extração de conhecimento, a Classificação foi um dos que captou a atenção do grupo, pois o uso deste método adequa-se ao tipo de dados que se está a trabalhar.

A Classificação consiste na organização e categorização de dados de classes distintas, isto é, o uso deste método tem como objetivo prognosticar valores discretos. Para isto acontecer corretamente, é criado um modelo tendo por base a distribuição dos dados, modelo esse que será utilizado na classificação de novos

dados. Para além disso, com o modelo torna-se possível prever uma nova classe para mapear novos dados.

Após algum trabalho de pesquisa sobre os diferentes tipos de filtros e sobre as diferentes vantagens ou utilidades que eles ofereciam em relação aos outros, o grupo escolheu o tipo de filtros *Tree*, pois, tendo em conta que os dados são compostos por alguns atributos, nada melhor que construir árvores de decisão. Dentro destas, o algoritmo utilizado foi o *J48* porque as pesquisas revelaram que se tratava de um algoritmo clássico e bastante usado para a representação de árvores de decisão, sendo muito rápido e com uma forma bastante poderosa de exprimir os dados. Além disso, também foi o algoritmo de classificação utilizado numa das aulas práticas, pelo que os elementos do grupo sentiram-se mais à vontade com o mesmo.

Em relação à preparação do conjunto de dados, para efetuar este método não houve grandes dificuldades. Para além do tratamento de dados inicial que já foi explicado anteriormente, a alteração foi remover o atributo *rain* pois existia uma grande discrepância de valores para dias que choveu ou não, e portanto não era útil fazer uma previsão com este atributo. E também, ao contrário dos métodos anteriores, foi criado um ficheiro de *training* com apenas as primeiras 200 instâncias do *dataset* original.

Posto isto, o atributo principal a analisar foi a *area*, de forma a puder posteriormente conseguir prever se para determinadas condições, teríamos área incendiada ou não. Mais à frente serão apresentados os resultados e conclusões.

2.3 Resultados

Neste capítulo são apresentados os resultados obtidos.

2.3.1 Associação

Através dos resultados obtidos, foi possível registar as seguintes regras de associação com os seguintes graus de confiança:

1. Se o mês é setembro e os valores do índice ISI se encontrarem no intervalo [5.61 - 11.22] (137 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (137 casos). O grau de confiança desta regra é de 100%.

2. Se os valores do índice DC se encontrarem no intervalo [690.06 - 775.33] e os valores do índice ISI se encontrarem no intervalo [5.61 - 11.22] (118 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (118 casos). O grau de confiança desta regra é de 100%.

3. Se os valores do índice DMC se encontrarem no intervalo [117.18 - 146.2] (113 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (113 casos). O grau de confiança desta regra é de 100%.

4. Se os valores do índice ISI se encontrarem no intervalo [5.61 - 11.22] e não houve incêndio (150 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (149 casos). O grau de confiança desta regra é de 99%.

5. Se a coordenada no eixo dos yy's se situar no intervalo [3.4 - 4.1] e os valores do índice ISI se encontrarem no intervalo [5.61 - 11.22] (120 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (119 casos). O grau de confiança desta regra é de 99%.

6. Se os valores do índice DMC se encontrarem no intervalo [88.16 - 117.18] (114 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (112 casos). O grau de confiança desta regra é de 98%.

7. Se os valores do índice ISI se encontrarem no intervalo [5.61 - 11.22] (312 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (106 casos). O grau de confiança desta regra é de 98%.

8. Se a temperatura se encontrar no intervalo [17.75 - 20.86] (127 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (124 casos). O grau de confiança desta regra é de 98%.

9. Se os valores do índice ISI se encontrarem no intervalo [5.61 - 11.22] e houve incêndio (162 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (157 casos). O grau de confiança desta regra é de 97%.

10. Se os valores do índice DC se encontrarem no intervalo [690.06 - 775.33] (151 casos), então os valores do índice FFMC encontram-se no intervalo [88.45 - inf] (146 casos). O grau de confiança desta regra é de 97%.

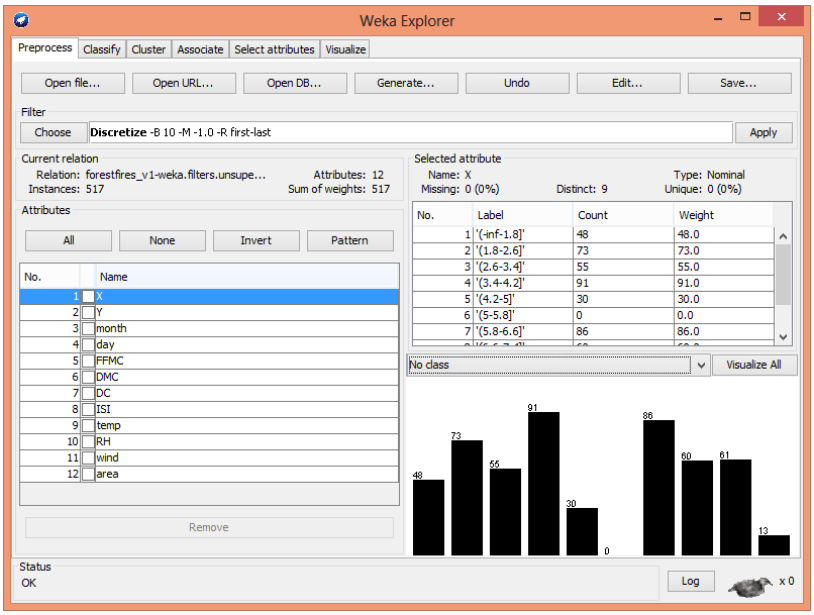

Figura 2.5: *Discretize* através da ferramenta Weka

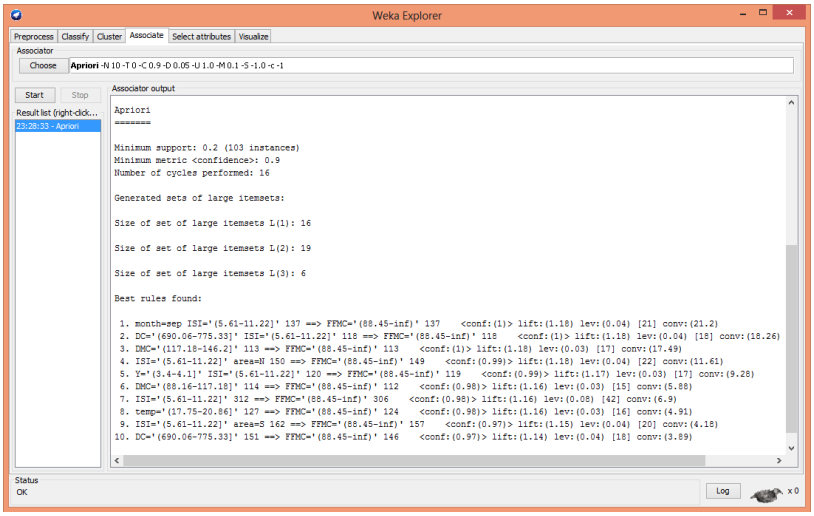

Figura 2.6: Associação através da ferramenta Weka

A informação obtida através do programa em Java é a seguinte:

```
- Numero de Instâncias: 517
- Atributos: 12

Apriori
=======

Minimum support: 0.2 (103 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 19

Size of set of large itemsets L(3): 6

Best rules found:

 1. month=sep ISI='(5.61-11.22]' 137 ==> FFMC='(88.45-inf)' 137
    <conf:(1)> lift:(1.18) lev:(0.04) [21] conv:(21.2)
 2. DC='(690.06-775.33]' ISI='(5.61-11.22]' 118 ==> FFMC='(88.45-inf)' 118
    <conf:(1)> lift:(1.18) lev:(0.04) [18] conv:(18.26)
 3. DMC='(117.18-146.2]' 113 ==> FFMC='(88.45-inf)' 113
    <conf:(1)> lift:(1.18) lev:(0.03) [17] conv:(17.49)
 4. ISI='(5.61-11.22]' area=N 150 ==> FFMC='(88.45-inf)' 149
    <conf:(0.99)> lift:(1.18) lev:(0.04) [22] conv:(11.61)
 5. Y='(3.4-4.1]' ISI='(5.61-11.22]' 120 ==> FFMC='(88.45-inf)' 119
    <conf:(0.99)> lift:(1.17) lev:(0.03) [17] conv:(9.28)
 6. DMC='(88.16-117.18]' 114 ==> FFMC='(88.45-inf)' 112
    <conf:(0.98)> lift:(1.16) lev:(0.03) [15] conv:(5.88)
 7. ISI='(5.61-11.22]' 312 ==> FFMC='(88.45-inf)' 306
    <conf:(0.98)> lift:(1.16) lev:(0.08) [42] conv:(6.9)
 8. temp='(17.75-20.86]' 127 ==> FFMC='(88.45-inf)' 124
    <conf:(0.98)> lift:(1.16) lev:(0.03) [16] conv:(4.91)
 9. ISI='(5.61-11.22]' area=S 162 ==> FFMC='(88.45-inf)' 157
    <conf:(0.97)> lift:(1.15) lev:(0.04) [20] conv:(4.18)
10. DC='(690.06-775.33]' 151 ==> FFMC='(88.45-inf)' 146
    <conf:(0.97)> lift:(1.14) lev:(0.04) [18] conv:(3.89)
```

2.3.2 Segmentação

O grupo procedeu ao treino do *dataset* e os resultados apresentados de seguida. Através do *output* foi possível concluir que nos 517 casos registados:

- **Cluster 0:** em 165 (32%) casos verifica-se a ocorrência de um incêndio no mês de setembro, sem chuva, à segunda-feira, com a velocidade do vento a variar entre os 1.3-2.2 kms/h, com um clima seco (humidade relativa entre os 32-40.5%), uma temperatura amena a variar entre os 17.75-20.86 °C, um índice ISI a variar entre os 5.61-11.22 valores, um índice DC a variar entre os 690.06-775.33 valores, um índice DMC a variar entre os 117.18-146.2 valores, um índice FFMC a variar entre os 88.45-291.3 valores, e a localização da área ardida a situar-se nas coordenadas (5.8-6.6) no eixo dos xx's e (4.8-5.5) no eixo dos yy's.

- **Cluster 1:** em 131 (25%) casos verifica-se a ocorrência de um incêndio no mês de agosto, sem chuva, ao sábado, com a velocidade do vento a variar entre os 3.1-4 kms/h, com um clima seco (humidade relativa entre os 23.5-32%), uma temperatura amena a variar entre os 14.64-17.75 °C, um índice ISI a variar entre os 5.61-11.22 valores, um índice DC a variar entre os 604.79-690.06 valores, um índice DMC a variar entre os 30.12-59.14 valores, um índice FFMC a variar entre os 88.45-291.3 valores, e a localização da área ardida a situar-se nas coordenadas (7.4-8.2) no eixo dos xx's e (3.4-4.1) no eixo dos yy's.

- **Cluster 2:** em 68 (13%) casos verifica-se a não ocorrência de qualquer incêndio no mês de fevereiro, à segunda-feira, sem chuva, com a velocidade do vento a variar entre os 2.2-3.1 kms/h, com um clima ameno (humidade relativa entre os 40.5-49%), uma temperatura a variar entre os 2.2-5.31 °C, um índice ISI a variar entre os 0.0-5.61 valores, um índice DC a variar entre os 7.9-93.17 valores, um índice DMC a variar entre os 1.1-93.17 valores, um índice FFMC a variar entre os 80.7-88.45 valores, e a área em questão a situar-se nas coordenadas (3.4-4.2) no eixo dos xx's e (3.4-4.1) no eixo dos yy's.

- **Cluster 3:** em 91 (18%) casos verifica-se a não ocorrência de qualquer incêndio no mês de agosto, à sexta-feira, sem chuva, com a velocidade do vento a variar entre os 4-4.9 kms/h, com um clima seco (humidade relativa entre os 32-40.5%), uma temperatura a variar entre os 17.75-20.86 °, um índice ISI a variar entre os 5.61-11.22 valores, um índice DC a variar entre os 604.79-690.06 valores, um índice DMC a variar entre os 88.16-117.18 valores, um índice FFMC a variar entre os 88.45-96.20 valores, e a área em questão a situar-se nas coordenadas (3.4-4.2) no eixo dos xx's e (3.4-4.1) no eixo dos yy's.

- **Cluster 4:** em 62 (12%) casos verifica-se a não ocorrência de qualquer incêndio no mês de agosto, à sexta-feira, sem chuva, com a velocidade do vento a variar entre os 3.1-4 kms/h, com um clima ameno (humidade

relativa entre os 40.5-49%), uma temperatura a variar entre os 20.86-23.97 °, um índice ISI a variar entre os 5.61-11.22 valores, um índice DC a variar entre os 690.06-775.33 valores, um índice DMC a variar entre os 146.2-175.22 valores, um índice FFMC a variar entre os 88.45-96.20 valores, e a área em questão a situar-se nas coordenadas (1-1.8) no eixo dos xx's e (3.4-4.1) no eixo dos yy's.

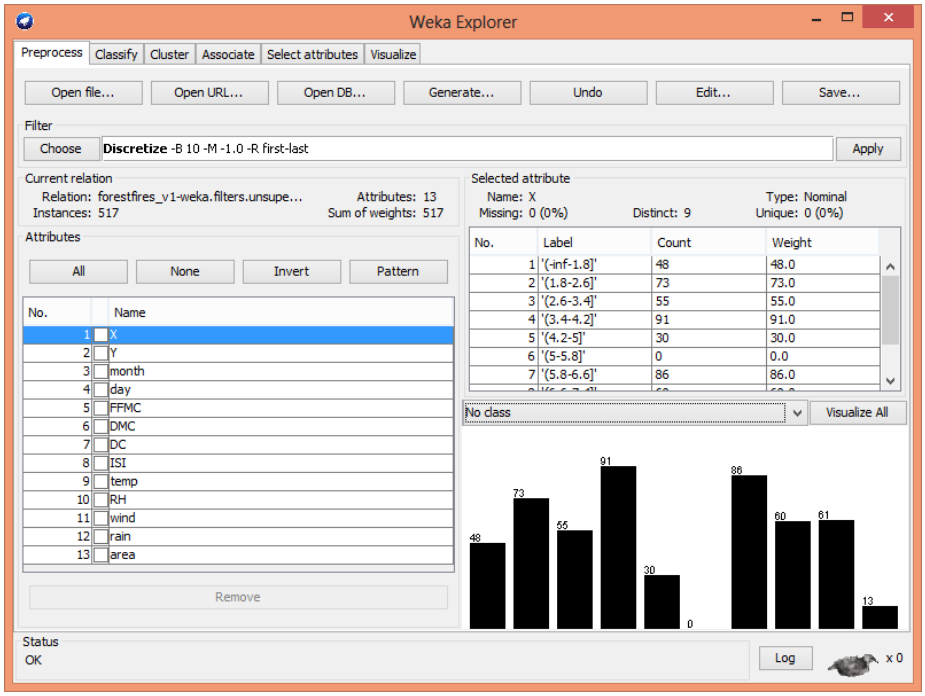

Figura 2.7: *Discretize* através da ferramenta Weka

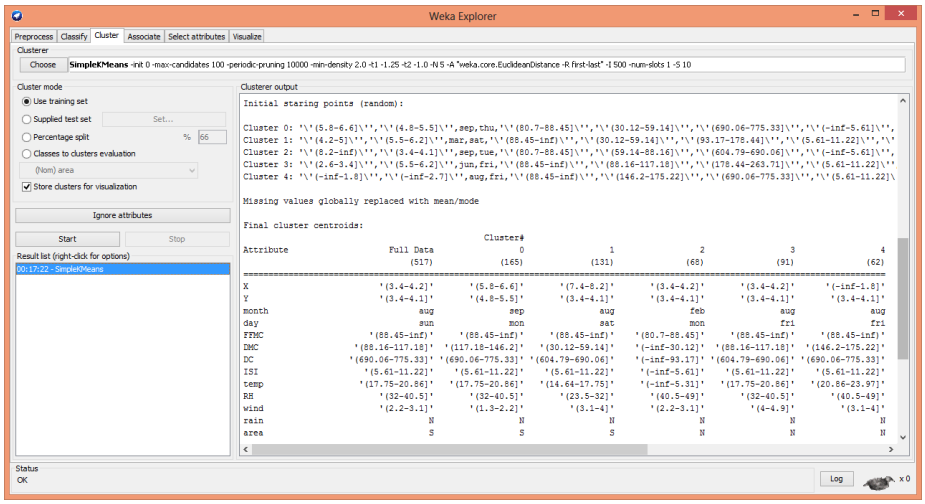

Figura 2.8: Segmentação através da ferramenta Weka

A informação obtida através do programa em Java é a seguinte:

```
- Numero de Instâncias: 517
- Atributos: 13

Algoritmo otimizado:
Capabilities: [Nominal attributes, Binary attributes, Unary attributes,
Empty nominal attributes, Numeric attributes, Missing values, No class]
Dependencies: []
min # Instance: 1


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 3139.0

Initial staring points (random):

Cluster 0: '\'(5.8-6.6]\'','\'(4.8-5.5]\'',sep,thu,'\'(80.7-88.45]\'',
           '\'(30.12-59.14]\'','\'(690.06-775.33]\'','\'(-inf-5.61]\'',
           '\'(11.53-14.64]\'','\'(49-57.5]\'','\'(1.3-2.2]\'',N,S
Cluster 1: '\'(4.2-5]\'','\'(5.5-6.2]\'',mar,sat,'\'(88.45-inf)\'',
           '\'(30.12-59.14]\'','\'(93.17-178.44]\'','\'(5.61-11.22]\'',
           '\'(14.64-17.75]\'','\'(57.5-66]\'','\'(3.1-4]\'',N,S
Cluster 2: '\'(8.2-inf)\'','\'(3.4-4.1]\'',sep,tue,'\'(80.7-88.45]\'',
           '\'(59.14-88.16]\'','\'(604.79-690.06]\'','\'(-inf-5.61]\'',
```

```
           '\'(23.97-27.08]\'','\'(32-40.5]\'','\'(2.2-3.1]\'',N,S
Cluster 3: '\'(2.6-3.4]\'','\'(5.5-6.2]\'',jun,fri,'\'(88.45-inf)\'',
           '\'(88.16-117.18]\'','\'(178.44-263.71]\'','\'(5.61-11.22]\'',
           '\'(17.75-20.86]\'','\'(32-40.5]\'','\'(4-4.9]\'',N,N
Cluster 4: '\'(-inf-1.8]\'','\'(-inf-2.7]\'',aug,fri,'\'(88.45-inf)\'',
           '\'(146.2-175.22]\'','\'(690.06-775.33]\'','\'(5.61-11.22]\'',
           '\'(23.97-27.08]\'','\'(40.5-49]\'','\'(3.1-4]\'',N,N

Missing values globally replaced with mean/mode

Final cluster centroids:
                          Cluster#
Attribute        Full Data                 0                 1
                     (517)             (165)             (131)
=============================================================
X               '(3.4-4.2]'       '(5.8-6.6]'       '(7.4-8.2]'
Y               '(3.4-4.1]'       '(4.8-5.5]'       '(3.4-4.1]'
month                   aug               sep               aug
day                     sun               mon               sat
FFMC          '(88.45-inf)'     '(88.45-inf)'     '(88.45-inf)'
DMC        '(88.16-117.18]'  '(117.18-146.2]'   '(30.12-59.14]'
DC        '(690.06-775.33]' '(690.06-775.33]' '(604.79-690.06]'
ISI           '(5.61-11.22]'    '(5.61-11.22]'    '(5.61-11.22]'
temp        '(17.75-20.86]'   '(17.75-20.86]'   '(14.64-17.75]'
RH              '(32-40.5]'       '(32-40.5]'       '(23.5-32]'
wind            '(2.2-3.1]'       '(1.3-2.2]'         '(3.1-4]'
rain                      N                 N                 N
area                      S                 S                 S

                    Cluster#
            2                 3                 4
         (68)              (91)              (62)
=============================================
   '(3.4-4.2]'       '(3.4-4.2]'      '(-inf-1.8]'
   '(3.4-4.1]'       '(3.4-4.1]'       '(3.4-4.1]'
           feb               aug               aug
           mon               fri               fri
'(80.7-88.45]'     '(88.45-inf)'     '(88.45-inf)'
'(-inf-30.12]'  '(88.16-117.18]'  '(146.2-175.22]'
'(-inf-93.17]' '(604.79-690.06]' '(690.06-775.33]'
 '(-inf-5.61]'    '(5.61-11.22]'    '(5.61-11.22]'
 '(-inf-5.31]'   '(17.75-20.86]'   '(20.86-23.97]'
   '(40.5-49]'       '(32-40.5]'       '(40.5-49]'
   '(2.2-3.1]'         '(4-4.9]'         '(3.1-4]'
             N                 N                 N
             N                 N                 N
```

2.3.3 Classificação

A definição de uma árvore de decisão ocorre através da divisão de um conjunto de dados em subconjuntos de forma recursiva. Uma árvore é formada por nós, ramos e folhas. Os nós representam regiões onde são realizados testes lógicos para a separação dos dados. O primeiro nó é chamado o nó raiz e é o nó principal da árvore de decisão. Os nós que estão localizados abaixo do nó raiz são os nó filhos e esses nós estão conectados por ramos.

Posto isto, nas imagens a seguir é possível verificar a árvore de decisão gerada e o *output* do algoritmo *J48*.

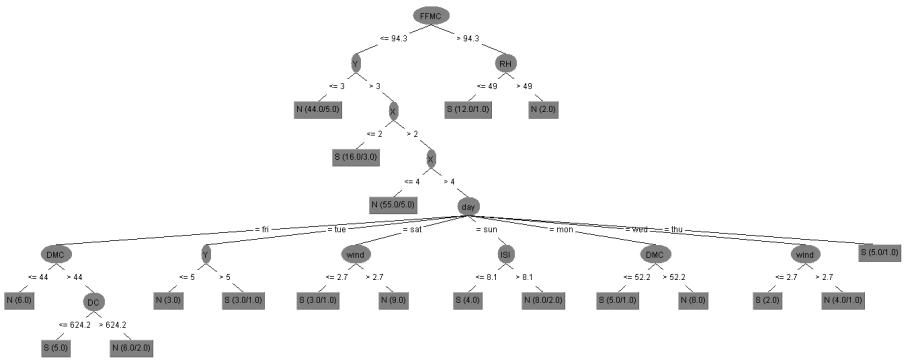

Figura 2.9: Árvore de Decisão

A informação obtida através do programa em Java é a seguinte:

```
- Numero de Instâncias: 200
- Atributos: 12

Parametros do algoritmo otimizado:
-C 0.25 -M 2

Algoritmo otimizado:
J48 pruned tree
------------------

FFMC <= 94.3
|   Y <= 3: N (44.0/5.0)
|   Y > 3
|   |   X <= 2: S (16.0/3.0)
|   |   X > 2
|   |   |   X <= 4: N (55.0/5.0)
|   |   |   X > 4
```

```
|   |   |   |   day = fri
|   |   |   |   |   DMC <= 44: N (6.0)
|   |   |   |   |   DMC > 44
|   |   |   |   |   |   DC <= 624.2: S (5.0)
|   |   |   |   |   |   DC > 624.2: N (6.0/2.0)
|   |   |   |   day = tue
|   |   |   |   |   Y <= 5: N (3.0)
|   |   |   |   |   Y > 5: S (3.0/1.0)
|   |   |   |   day = sat
|   |   |   |   |   wind <= 2.7: S (3.0/1.0)
|   |   |   |   |   wind > 2.7: N (9.0)
|   |   |   |   day = sun
|   |   |   |   |   ISI <= 8.1: S (4.0)
|   |   |   |   |   ISI > 8.1: N (8.0/2.0)
|   |   |   |   day = mon
|   |   |   |   |   DMC <= 52.2: S (5.0/1.0)
|   |   |   |   |   DMC > 52.2: N (8.0)
|   |   |   |   day = wed
|   |   |   |   |   wind <= 2.7: S (2.0)
|   |   |   |   |   wind > 2.7: N (4.0/1.0)
|   |   |   |   day = thu: S (5.0/1.0)
FFMC > 94.3
|   RH <= 49: S (12.0/1.0)
|   RH > 49: N (2.0)

Number of Leaves  :   19

Size of the tree :  32

Resultados nos Dados de Teste
Correctly Classified Instances         177               88.5    %
Incorrectly Classified Instances        23               11.5    %
Kappa statistic                          0.7226
Mean absolute error                      0.1885
Root mean squared error                  0.307
Relative absolute error                 43.9847 %
Root relative squared error             66.3758 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      90.25   %
Total Number of Instances              200
```

O rácio de classificações corretas sobre o total de classificações é de 88.5%, que é considerado bastante bom. Este algoritmo desempenha bastante bem neste conjunto de dados. Teria sido útil, ter testado a opção de fornecer um conjunto de dados para teste (de forma a obter resultados ainda mais conclusivos), porém não possuíamos outra fonte para obter dados da mesma natureza.

Capítulo 3

Notas Finais

3.1 Conclusões

Com este trabalho prático ficaram retidas as principais funcionalidades da ferramenta WEKA no que diz respeito ao pré-processamento, filtragem, classificação, segmentação e associação dos dados de um *dataset*. Conheceram-se os métodos disponibilizados pela API WEKA em Java, e melhorou-se a aplicação destes métodos. Consultou-se e analisou-se um conjunto de dados, e encontraram-se resultados.

No decorrer da realização da segunda parte do trabalho de grupo, foram detetadas as seguintes falhas: erros de compilação do código Java associados à incorreta incorporação dos métodos disponibilizados pelo WEKA; dificuldades na interpretação das funcionalidades da ferramenta WEKA; dificuldades na interpretação dos resultados obtidos pelo WEKA.

Para solucionar estas falhas, o grupo tomou as seguintes atitudes: para corrigir os erros associados à compilação do código, foi necessário recorrer à equipa docente da unidade curricular; para superar as dificuldades na interpretação das funcionalidades do WEKA, recorreu-se ao material fornecido pela equipa docente, e efetuou algumas pesquisas em páginas Web que se encontram mencionadas na Bibliografia; para obter uma melhor interpretação dos resultados obtidos, o grupo recorreu ao material fornecido pela equipa docente, e realizou algumas pesquisas em páginas Web que se encontram mencionadas na Bibliografia.

Através do conjunto de dados escolhido, não foi necessário efetuar um grande esforço para a limpeza de dados, visto que não havia muitos atributos no *dataset*. Apenas fizeram-se algumas alterações, para ser possível obter resultados mais conclusivos.

O facto de o grupo desenvolver uma aplicação em Java, ajudou a analisar os resultados, focando apenas a atenção no conjunto de *outputs* produzidos. Ou seja, abstraiu-se um pouco da ferramenta WEKA, de maneira a ser possível isolar mais a análise dos resultados.

Bibliografia

[1] - Descrição da ferramenta WEKA:
http://pt.wikipedia.org/wiki/Weka

[2] Scuse D.; Reutemann P. (2007) "WEKA Experimenter Tutorial for Version 3-5-5"The University of WAIKATO

[3] - API WEKA:
http://weka.wikispaces.com/Use+Weka+in+your+Java+code

[4] - Use Weka in your Java code:
http://weka.wikispaces.com/Use+Weka+in+your+Java+code

[5] Damasceno, M., "Introdução a Mineração de Dados Utilizando o Weka" Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte

[6] Carlos, R., "Trabalho prático de mineração de dados" Algoritmos de aprendizado para avaliação de carros

[7] - Association Rule Mining with WEKA:
http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/associate.html

[8] - *Forest Fires Data Set*:
https://archive.ics.uci.edu/ml/datasets/Forest+Fires

[9] - Página do Professor Paulo Cortez da Universidade do Minho sobre *Forest Fires Dataset*
http://www3.dsi.uminho.pt/pcortez/forestfires/

[10] - Referência do artigo de Cortez e Morais, 2007:
http://www3.dsi.uminho.pt/pcortez/fires.pdf

[11] - Descrição dos índices de clima de incêndios:
http://www.malagaweather.com/fwi-txt.htm

[12] - *Using Weka 3 for clustering*:
http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-Ex3.html

[13] - *Data mining with WEKA, Part 2: Classification and clustering*:
`http://www.ibm.com/developerworks/library/os-weka2/`

[14] - *Use Weka in your Java code*:
`http://weka.wikispaces.com/Use+Weka+in+your+Java+code`