

Técnicas de data mining para agrupamento dos vinhos certificados das sub-regiões da região demarcada dos Vinhos Verdes

Data mining techniques for the grouping of certified wines from the sub-regions of the demarcated region of Vinho Verde

Souza-Roza R., LIAAD – INESC.TEC, Portugal, ruisouzaroza@gmail.com

Brazdil P., LIAAD – INESC.TEC/FEP Universidade do Porto, Portugal, pbrazdil@inesc.pt

Reis J.L., Portugal, IPAM LAB/CVRVV/CETRAD-UTAD, jreis@ipam.pt

Cerdeira A., CVRVV, Portugal, acerdeira@vinhoverde.pt

Martins P., CVRVV, Portugal, pmartins@vinhoverde.pt

Felgueiras O., FCUP Universidade do Porto, Portugal, olfelgue@fc.up.pt

Resumo

A conjugação da informação, obtida a partir do tratamento, com técnicas de *data mining*, sobre os dados das análises físico-químicas e organoléticas permitiu obter semelhanças entre os vinhos das nove sub-regiões existentes na *Região Demarcada dos Vinhos Verdes*. Através das técnicas de agrupamento (*clustering*) foram identificados quatro agrupamentos (*clusters*), cada um caracterizado pelo seu centróide. A medida de ganho de informação, em conjunto com a aprendizagem supervisionada baseada em regras, foi usada para encontrar as características diferenciadoras. Este estudo permitiu a interligação das características dos vinhos dessas sub-regiões, o que pode melhorar as tomadas de decisões sobre os perfis desses mesmos vinhos.

Palavras-chave: Verdes; análises físico-químicas; análises organoléticas; *data mining*; *clusters*.

Abstract

The combination of information obtained from data mining technique from physicochemical and organoleptic data analysis allowed similarities between the wines of the nine sub-regions in the Demarcated Region of Vinho Verde. Through clustering techniques, four clusters were identified, each characterized by its centroid. The measure of information gain, together with supervised rule-based learning, was used to find the differentiating characteristics. This study allowed the interconnection of the characteristics of the wines of these sub-regions, which can improve the decision making on the profiles of these same wines.

Keywords: Vinho Verde; physicochemical analysis; organoleptic analysis; *data mining*; *clusters*.

1. INTRODUÇÃO

Nas últimas décadas, tanto o sector vitivinícola como as tecnologias da informação sofreram grandes desenvolvimentos. No que diz respeito ao sector vitivinícola, verificaram-se mudanças

bastante radicais, desde os sistemas de condução da vinha até à modernização da infra-estrutura, controlo de qualidade e gestão do solo, até à adaptação às alterações climáticas (Silva, 2015). As tecnologias da informação também têm contribuído para o progresso. A aplicação da informação para os negócios, que engloba novos campos, tais como *data warehousing* e *data mining*, possibilitou a recolha, armazenamento e processamento de dados quantitativos e qualitativos, proporcionando aos decisores meios úteis para a tomada de decisões (Braga, 2009).

Nesta área já foram realizados vários estudos que combinam as duas áreas. Por exemplo, Cortez et al. (2009) compararam as características físico-químicas das amostras de vinho com características organoléticas de carácter mais subjectivo, como cor, clareza, aroma e sabor. Além disso, Ribeiro et al. (2009a) utilizaram diferentes modelos de classificação e regressão, como Árvores de Decisão, Redes Neurais Artificiais e Regressões Lineares para prever os parâmetros organoléticos a partir dos parâmetros químicos do processo de vinificação. Um estudo semelhante (Ribeiro et al., 2009b) teve como objetivo prever os valores dos atributos subjectivos, incluindo o gosto, a cor e o sabor. Com estas técnicas, é possível prever os valores subjectivos dos parâmetros, variando os valores de alguns parâmetros químicos.

Este estudo, realizado em colaboração com a Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), processou amostras de vinho com o objetivo de responder às seguintes questões:

1. São produzidos vinhos similares em diferentes sub-regiões da Região dos Vinhos Verdes e, em caso afirmativo, onde.
2. Como podemos caracterizar grupos de vinhos semelhantes?

As respostas às perguntas 1 e 2 são de interesse não apenas para a CVRVV, mas também para um público-alvo que inclui produtores de vinho e todos os envolvidos na comercialização. Este trabalho é de interesse também para investigadores do campo da Inteligência Artificial (IA) e, em particular, nas áreas de aprendizagem automática (*machine learning*) e da extracção de dados (*data mining*). Embora tenham sido reutilizadas várias técnicas de extracção de dados, nomeadamente, de agrupamento (*clustering*), não foram obtidas respostas imediatas às questões colocadas.

Uma questão específica a resolver prende-se com a caracterização dos *clusters* gerados. O simples uso de, por exemplo, o *centróide* não satisfaz, tendo sido necessário identificar as características importantes que distinguem um *cluster* dos outros.

2. ESTUDO DA SIMILARIDADE DE AMOSTRAS DE VINHO DIFERENTES

Antes da descrição da metodologia utilizada, nas secções seguintes é apresentada uma breve descrição do conjunto de dados utilizados neste estudo.

2.1 Conjunto de dados das amostras usadas

O conjunto de dados fornecidos pela CVRVV contém amostras de vinho recolhidas de janeiro de 2004 a dezembro de 2015 com dados das características físico-químicas e sensoriais. Estes dados não contêm qualquer informação que possa identificar a empresa envolvida, nem a identificação do respectivo vinho. O conjunto de dados utilizado inclui 4941 amostras de vinho verde branco e contém 4 características gerais do produto (amostras, elenco, sub-região e denominações especiais), 13 características físico-químicas e 8 características organoléticas. Todos os atributos que tinham mais de 1% dos valores em falta foram eliminados deste estudo, pois processar dados sem informação não faria sentido.

A Tabela 1 mostra as diferentes sub-regiões representadas no conjunto de dados.

Sub-Região	Nº Amostras	%
Amarante	501	10.1
Ave	471	9.5
Baião	320	6.5
Basto	391	7.9
Cávado	431	8.7
Lima	465	9.4
Monção e Melgaço	1808	36.6
Paiva	155	3.1
Sousa	399	8.1
Total	4941	100

Tabela 1 - Lista de sub-regiões e a correspondente distribuição das amostras

Relativamente às 13 características físico-químicas, elas estão listadas na Tabela 2, juntamente com a média, e outras estatísticas descritivas. Todas essas variáveis foram normalizadas para uma escala comum aplicando a normalização gaussiana (Han et al., 2012; Witten et al., 2011).

Esta normalização é necessária, uma vez que evita dar uma importância indevida a atributos com valores relativamente grandes e útil para o agrupamento que será apresentado num ponto seguinte.

As características organoléticas estão listadas na Tabela 3.

Nº	Características	Descrição	Média	Mediana	D.Pad.	Min.	Max.
1	AcidFix	Acidez Fixa	6.18	6.10	0.89	3.50	13.90
2	AcidTot	Acidez Total	6.55	6.50	0.88	3.80	14.20
3	AcidVolat	Acidez Volátil	0.30	0.29	0.10	0.05	1.08
4	AcidCitric	Acidez Citrica	0.31	0.29	0.11	-0.01	2.07
5	Cloretos		0.04	0.03	0.03	0.01	0.50
6	SO2Free	SO2 Livre	30.16	29.00	14.01	0.00	339.00
7	SO2Tot	SO2 Total	115.75	112.00	31.69	9.00	481.00
8	ExtrNRed	Extrato Não-Reduct	18.89	18.70	1.98	14.40	34.70
9	ExtrDryTot	Extract Seco Total	23.69	22.80	5.65	15.40	162.00
10	Dens	Densidade	0.99	0.99	0.00	0.99	1.04
11	pH	Acidez/alcalinidade	3.21	3.22	0.15	2.58	3.82
12	Sulphat	Sulfitos	0.48	0.46	0.15	0.17	1.93
13	AlcoAcquir	Alcool asquirido	11.90	12.00	0.95	8.60	14.90

Tabela 2 - Visão geral das características físico-químicas nas amostras

Nº	Características	Descrição	Média	Mediana	D.PAd	Min.	Max.
1	AromaQual	Aroma qualidade	6.51	7	0.73	2	9
2	TasteQual	Sabor qualidade	6.51	7	0.73	2	9
3	AromaTypic	Aroma tipicidade	6.54	7	0.69	5	9
4	TasteTypic	Sabor tipicidade	6.54	7	0.69	5	9
5	AromaDef	Aroma defeito*	1.99	2	0.10	1	2
6	TasteDef	Sabor defeito*	1.99	2	0.10	1	2
7	Color	Cor	2.24	2	0.49	1	5
8	Clarity	Limpidez	1.05	1	0.30	1	5

* Defeito: os valores desta característica são: 1- com defeito; 2- sem defeito

Tabela 3 - Visão geral das características organoléticas

Todos os valores estão na escala de 0 a 10 e têm interpretação predefinida. Os valores relativos à qualidade do aroma e sua interpretação são mostrados na Tabela 4. Este atributo deve atingir a classificação de pelo menos 6 para satisfazer os requisitos de vinho certificado como Vinho Verde.

10	9	8	7	6	5	4	3	2	1	0
Excelente	Muito bom	Bom	Suficiente	Mediocre	Mau					

Tabela 4 - Valores para a qualidade do aroma

Os valores relativos à Tipicidade do Aroma e Tipicidade do Sabor e a sua interpretação são mostrados na Tabela 5. Este atributo deve atingir a classificação de pelo menos 5 para satisfazer os requisitos de vinho certificado como Vinho Verde.

10	9	8	7	6	5	4	3	2	1	0
Típico					Atípico					

Tabela 5 - Valores da tipicidade do aroma e da tipicidade do sabor

2.2 Agrupamento (*clustering*) das amostras de vinho

No primeiro estudo, recorreu-se ao agrupamento das amostras de vinho. Isso foi motivado pela observação de que amostras de vinho similares deveriam ser agrupadas no mesmo *cluster*. Todas as características físico-químicas e organoléticas foram utilizadas neste processo, mas o atributo sub-região foi deixada de fora, pois afectaria o processo de agrupamento. O objetivo foi analisar os *clusters* e as sub-regiões associadas.

Existem três tipos de técnicas de agrupamento, tais como, hierárquicos, particionais, baseados em densidade, entre outros (Gama et al., 2015). Neste trabalho foi usado o algoritmo de agrupamento *k-means* que pertence aos métodos particionais. Este algoritmo foi escolhido porque é muito comum, embora tenha alguns defeitos, tais como: alta sensibilidade aos *outliers* e baixa capacidade

para ultrapassar o mínimo local. No entanto, observamos que este *software* gerou sempre 4 *clusters* com os dados após diferentes inicializações.

Como o *k-means* requer o número de *clusters* como entrada, foi usada outra técnica que o permite determinar. O método utilizado foi EM, Maximização da Expectativa (Knime, 2015). Este procedimento começa com o número mínimo de *clusters* (isto é, 1) e mantém-se aumentando enquanto a medida de probabilidade aumenta. Este procedimento retornou 4 como o melhor número de *clusters* para o nosso conjunto de dados. O *k-means* foi executado com esta entrada. Assim, quando os *k-means* foram executados com o número de agregados = 4 para as 4941 amostras, obteve-se o seguinte resultado (ver Tabela 6).

<i>Cluster</i>	Nº Amostras	%
1	2076	42.02
2	641	12.97
3	1259	25.48
4	965	19.53
Totais	4941	100

Tabela 6 - *Clusters* gerados e correspondente número de amostras de vinho

2.2 Caracterização dos *clusters* individuais

Os resultados do processo de *clustering* não são só por si muito significativos, pois não fornecem uma informação muito útil para os especialistas no domínio se fizermos uma análise mais aprofundada. Poderiam ser inspeccionados os itens individuais em cada *cluster*, mas inspeccionar centenas ou milhares de itens excede os limites humanos. Portanto, é necessário caracterizar cada conjunto de forma mais concisa.

A primeira ideia útil é o conceito de *centróide* (Han et al., 2012). O *centróide* de um *cluster* tem a forma de um elemento, em que o valor de cada atributo é representado pela média (*mean*), se o atributo inclui valores numéricos ou moda (*mode*), se o atributo for categórico. Assim, pode-se dizer que o *centróide* de um *cluster* representa esse *cluster*. O simples uso do *centróide* não é inteiramente satisfatório, pois não é possível distinguir características importantes das não importantes. Portanto, foram utilizados outros meios que complementaram essa informação.

Supondo que se deseja caracterizar o agrupamento C_i e compará-lo com C_{other} em que $C_{other} = C - C_i$, representa os itens em todos os outros *clusters*. A caracterização de C_i envolve as seguintes etapas:

1. Calcular o *ganho de informação* de cada atributo supondo que se deseja classificar itens em C_i ou C_{other} ;
2. Calcular a diferença entre os valores de um dado atributo dos *centróides* de C_i e C_{other} ;

3. Gerar regras de classificação que permitam classificar os casos em C_i ou C_{other} ;

Nas secções seguintes são apresentados os detalhes de cada um dos pontos acima mencionados.

Cálculo do ganho de informação de cada atributo

O facto de o centróide envolver todos os atributos pode não ser elucidativo porque não chama a atenção para os atributos importantes. Estes são os que discriminam bem os itens do *cluster* C_i em relação a todos os outros itens (isto é, C_{other}). Poderão ser adoptadas várias medidas para esse efeito. Neste caso optamos pelo ganho de informação (*InfoGain*) (Han et al., 2012). Os atributos são classificados por ordem decrescente desta medida. Para simplificar a análise, mostramos apenas os atributos com $InfoGain > 0,1$. Podemos ver que, por exemplo, no *Cluster* C_1 (ver Tabela 7) o atributo mais informativo é *TasteQual*, com o valor de $InfoGain = 0.649$.

Diferença entre os valores de atributo dos centróides de C_i e C_{other}

Procuramos fornecer informações adicionais que permitam caracterizar cada *cluster*. Para além de calcular o centróide de C_i este processo pode ser repetido também para C_{other} para comparar os valores.

Analise os valores do grupo C_1 (ver Tabela 7). Notamos que o centróide deste *cluster* (coluna 4) para *TasteQual* = 7.133, que é um valor muito maior do que os valores de C_{other} (ou seja, 5.964 na coluna 5). Assim, este atributo não é apenas informativo, como vimos antes *InfoGain* é relativamente elevado, indicando que este *cluster* inclui vinhos de alta qualidade.

Quanto às características físico-químicas, apresentamos os valores normalizados e não normalizados originais. Por exemplo, consideremos *AcidFix* na Tabela 7. O valor mostrado na coluna 5 é - 0.425 (5.8) indica que o valor normalizado é - 0.425 e o não normalizado é 5.8.

Gerar regras de classificação que permitam classificar casos em C_i ou C_{other}

Existe uma outra maneira de caracterizar cada *cluster* C_i . Podemos empregar regras discriminativas geradas por um sistema de árvore de decisão e transformar o resultado em regras (Han et al., 2012). O sistema é treinado com o objetivo de distinguir os casos de C_i que ficam fora de C_{other} . Por outras palavras, todos os casos de *cluster* C_i são rotulados como pertencentes à classe positiva, enquanto todos os outros são considerados como pertencendo à classe negativa.

Todas as regras que classificam casos na classe positiva (correspondente a C_i) são analisadas e as que cobrem um número mínimo de casos são consideradas para uso posterior. Se a regra inclui um atributo Atr_i na sua condição, então ele é anexado a esse atributo na tabela listando todos os atributos para o *cluster* C_i .

Para ilustrar, consideremos os atributos da Tabela 7 para o *cluster* C_1 . A regra $AcidFi \leq 0.75$ foi gerada pelo sistema entre várias outras regras. É anexado ao atributo *AcidFix* nesta tabela (na coluna 7). Observe-se que o valor 0.75 está na escala normalizada. Esta regra diz-nos que se a condição $AcidFix \leq 0.75$ é satisfeita para uma determinada amostra de vinho, e pode ser incluída no *cluster* C_1 . Esta regra abrange apenas alguns dos itens. Outras regras abrangem outros itens. Note-se também que o valor do centróide (-0.425) é muito menor que o centróide dos outros grupos (-0.202) e também muito inferior ao limite de 0.75 que aparece na regra. Portanto, se a regra foi usada para classificar o item centróide, ele dispararia e sugeriria C_1 , como deveria ser.

3. DESCRIÇÃO E CARACTERIZAÇÃO DOS QUATRO CLUSTERS GERADOS

Esta secção inclui quatro subsecções, cada uma dedicada a um dos *clusters* gerados. Cada subsecção inclui uma tabela onde cada coluna inclui:

1. Nome do atributo,
2. Tipo de características (sensoriais, isto é, organolépticas, ou físico-químicas),
3. Valor *InfoGain* fornecendo informações sobre como o atributo é informativo,
4. Valor centróide deste atributo no agrupamento C_i ,
5. Valor centróide deste atributo nos remanescentes agrupamentos C_{other} ,
6. Diferença entre os valores das colunas anteriores,
7. A condição antecedente de uma regra gerada com a maior cobertura.

Os elementos nesta tabela são ordenados pelo *InfoGain*.

Cluster C_1 - Vinhos de qualidade superior e elevado teor de álcool

Este *cluster* inclui 2076 elementos, representando 42,02% dos casos. A Tabela 7 mostra as características mais relevantes deste *cluster*.

Atributo	Tipo	InfoGain	Centróide	Centróide	Centróide diferença	Regra
<i>TasteQual</i>	Sens.	0.649	7.133	5.964	1.169	
<i>AromaQual</i>	Sens.	0.644	7.133	5.967	1.166	
<i>TasteTypic</i>	Sens.	0.637	7.137	6.007	1.130	
<i>AromaTypic</i>	Sens.	0.634	7.136	6.009	1.127	
<i>AcidFix</i>	Phys-Chem	0.148	-0.425 (5.8)	0.308	-0.733	$AcidFix \leq 0.75$ (6.85)
<i>AcidTot</i>	Phys-Chem	0.131	-0.407 (6.2)	0.295	-0.702	
<i>AlcoAcquir</i>	Phys-Chem	0.123	0.447 (12.3)	-0.324	0.771	$AlcoAcquir > -2.06$ (9.9)
<i>Dens</i>	Phys-Chem	0.102	-0.342 (0.99)	0.248	-0.590	$Dens \leq 1.11$ (0.98)

Tabela 7– Caracterização do *Cluster* C_1

Quanto às características sensoriais deste agrupamento, tanto o sabor como o aroma são caracterizados por boas classificações acima de 7.1, isto é, mais de 1.1 valores acima dos valores dos outros agrupamentos.

No nível das características físico-químicas, observa-se que este grupo é caracterizado por uma acidez relativamente baixa (tanto *AcidFix* quanto *AcidTot*), teor de álcool relativamente alto (*AlcoAcquir*) para Vinho Verde (12,3 °) e densidade relativamente baixa (*Dens*) indicando um nível mais baixo de sólidos.

Análise da distribuição de amostras no agrupamento C_1 em diferentes regiões sub-vitivinícolas

Embora a informação sobre a sub-região do vinho não tenha sido utilizada no *clustering*, a informação existe e está associada a cada amostra de vinho. Portanto, podemos recuperar essas informações e anexá-las às amostras em cada *cluster*.

A Tabela 8 mostra a distribuição das amostras por sub-região. Apenas são mostradas as sub-regiões onde a proporção mudou em mais de 2% quando comparada com a Tabela 1. Observamos que este *cluster* é composto principalmente por vinhos da sub-região de Monção e Melgaço (52%), sendo a proporção maior do que quando comparada com a não agrupada (36,6%).

Sub-Região	Nº Amostras	Amostras %	Diferença todos os dados
Amarante	167	8.04	-2.06
Ave	123	5.92	-3.58
Baiao	84	4.05	-2.45
Monção e Melgaço	1078	51.93	15.33

Tabela 8 – Distribuição das amostras de vinho por sub-região para o *cluster* C_1

Cluster C_2 - Vinhos com alguma acidez

Este *cluster* inclui 641 elementos, representando 12,97% dos casos. A Tabela 9 mostra as características mais relevantes deste *cluster*.

Atributo	Tipo	InfoGai	Centróide	Centrói	Centrói	Regra
Dens	Phys-	0.460	1.322	-0.197	1.519	1.265 (0.995) < Dens ≤
SO2Tot	Phys-	0.307	1.213	-0.181	1.394	-0.324 (105) < SO2Tot ≤ 2.706
ExtrDryT	Phys-	0.306	1.035	-0.154	1.189	ExtrDryTot ≤ 2.071 (35.4)
AlcoAcqu	Phys-	0.281	-1.119	0.167	-1.286	AlcoAcquir > -3.069 (9)
ExtrNRed	Phys-	0.185	0.908	-0.135	1.043	
Cloret	Phys-	0.140	0.626	-0.093	0.719	-0.279 (0.029) < Cloret ≤ 5.226
TasteTypi	Sens.	0.100	6.153	6.594	-0.441	
TasteQua	Sens.	0.099	6.119	6.569	-0.450	

Tabela 9 – Caracterização do *cluster* C_2

A densidade relativamente alta contrasta com baixo teor de álcool (*AlcoAcquir*). É também caracterizada por um nível mais elevado de dióxido sulfuroso (*SO2Tot*) que afeta negativamente o

sabor. Todas as características sensoriais estão em torno de 6.1, bem abaixo dos níveis do grupo C₁ (acima de 7.1).

Análise da distribuição de amostras no C₂ em diferentes regiões vitivinícolas

Apenas algumas das sub-regiões são mostradas na Tabela 10. São aqueles onde a proporção mudou em mais de 2% quando comparada com a Tabela 1. Observamos que este *cluster* inclui mais amostras de quatro sub-regiões - Amarante, Ave, Cávado e Sousa. A sub-região de Monção e Melgaço está sub-representada quando comparada com os dados não agrupados.

Sub-Região	Nº amostras	Amostras %	Diferença Todos os dados
Amarante	99	15.44	5.34
Ave	98	15.29	5.79
Cavado	93	14.51	5.81
Monção e Melgaço	119	18.56	-18.04
Sousa	72	11.23	3.13

Tabela 10 – Distribuição das amostras de vinho por sub-região para o *cluster* C₂

Cluster C₃ – Vinhos com acidez elevada

Este *cluster* inclui 1259 elementos, representando 25,48% dos casos. A Tabela 11 apresenta as características mais relevantes deste *cluster*.

Atributo	Tipo	InfoGain	Centróide	Centróide	Centróide	Regra
TasteQual	Sens.	0.463	5.886	6.725	-0.839	TasteQua \leq 6.5
AromaQual	Sens.	0.455	5.888	6.726	-0.838	
TasteTypic	Sens.	0.422	5.933	6.744	-0.811	
AromaTypic	Sens.	0.414	5.936	6.743	-0.807	
AcidFix	Phys-Chem	0.175	-0.509 (5.7)	0.175	-0.684	AcidFix \leq 0.529 (6.7)
AcidTot	Phys-Chem	0.164	-0.523 (6.1)	0.179	-0.702	

Tabela 11 – Caracterização do *cluster* C₃

Este agrupamento é caracterizado por valores bastante baixos de características sensoriais (cerca de 5.9, isto é, no nível de suficiente). Isto compara desfavoravelmente com o grupo C₁ (acima de 7.1) e C₂ (em torno de 6.1 que é considerado bom). Isto é provavelmente devido a altos níveis de acidez (tanto *AcidFix* e *AcidTot*).

Análise da distribuição de amostras no C₃ em diferentes sub-regiões vitivinícolas

Este *cluster* segue a tendência geral de dados não agrupados. Apenas duas sub-regiões se afastam ligeiramente do padrão geral (ver Tabela 12):

Sub-Região	Nº amostras	Amostras %	Diferença todos os dados
Ave	156	12.39	2.89
Monção e Melgaço	425	33.76	-2.84

Tabela 12 – Distribuição das amostras de vinho por sub-região para o cluster C₃**Cluster C₄ - Vinhos com baixa Acidez e baixo Extracto Não Redutor**

Este *cluster* inclui 965 elementos, representando 19,5% dos casos. A Tabela 13 mostra as características mais relevantes deste *cluster*.

Atributo	Tipo	InfoGain	Centróide	Centróide C _{other}	Centróide diferença	Regra
AcidFix	Phys-Chem	0.642	1.313 (7.347)	-0.319	1.632	AcidFix > 0.305(6.45)
AcidTot	Phys-Chem	0.619	1.304 (7.697)	-0.316	1.62	
pH	Phys-Chem	0.242	-0.852 (3.091)	-0.126	0.645	pH ≤ 1.037 (3.37)
ExtrNRed	Phys-Chem	0.104	0.519 (19.91)	-0.057	0.292	ExtrNRed ≤ 4.721 (28.1)
Dens	Phys-Chem	0.093	0.235 (0.992)	0.207	-1.059	Dens ≤ 2.653 (0.998)

Tabela 13 – Caracterização do cluster C₄

Este agrupamento inclui amostras com acidez muito baixa (*AcidFix*), bem acima do valor de referência, níveis normais de pH, baixo extracto não redutor (*ExtrNRed*), tornando-o "mais leve" na boca. Além disso, este *cluster* tem densidade relativamente baixa (*Dens*). Os vinhos são de boa qualidade.

Análise da distribuição de amostras no C₄ em diferentes sub-regiões vitivinícolas

Os detalhes da distribuição de amostras no C₄ em diferentes sub-regiões vitivinícolas encontram-se na Tabela 14.

Sub-Região	Nº amostras	Amostras %	Diferença Todos os dados
Baiao	98	7.63	3.66
Basto	99	7.78	2.36
Lima	117	9.93	2.72
Monção e Melgaço	186	33.76	-17.33
Paiva	57	3.97	2.81
Sousa	104	6.99	2.68

Tabela 14 – Distribuição das amostras de vinho por sub-região para o cluster C₄**4. DISCUSSÃO E CONCLUSÕES**

Neste trabalho foi discutido o problema de como identificar e caracterizar diferentes amostras de vinhos da região dos Vinhos Verdes com base nas suas características físico-químicas e organoléticas. Procurámos, por exemplo, responder à questão de saber se vinhos similares são produzidos em diferentes sub-regiões e, em caso afirmativo, onde. As respostas a esta pergunta

interessam à Comissão da Viticultura da Região dos Vinhos Verdes (CVRVV), mas também aos produtores e aos comerciantes de vinho.

No decorrer deste trabalho, utilizaram-se várias técnicas de extracção de dados, incluindo, por exemplo, *clustering*. Quando este método foi utilizado os Vinhos Verdes foram agrupados em quatro grupos. Um dos *clusters*, por exemplo, inclui os vinhos de qualidade superior que normalmente são encontrados na sub-região de Monção e Melgaço.

Neste trabalho foram caracterizados os *clusters* gerados. O uso simples, por exemplo, do centróide não satisfaz, sendo preciso identificar as características importantes que distinguem um *cluster* dos outros. Sugerimos três métodos diferentes que ajudam a atingir este objetivo:

- (1) Concentrar a atenção nos atributos mais informativos (com o valor de *InfoGain* > 0.1);
- (2) Para todos os atributos informativos, calcular a diferença entre o centróide do *cluster* C_i e o centróide de todos os outros dados C_{other} (exemplo: *TasteQual* em C_1 é 1.169 mais elevado na escala normalizada do que nos outros dados);
- (3) Complementar as informações sobre atributos informativos com condições recuperadas das regras de classificação (exemplo: em C_3 o *TasteQual* ≤ 6.5 do que no resto dos dados).

Este tipo de caracterização não é comum, embora pareça bastante útil ao transmitir os resultados aos especialistas neste domínio. Acreditamos que este trabalho é do interesse não só para os especialistas que trabalham na vitivinicultura, mas também noutros domínios de conhecimento onde questões semelhantes podem surgir.

REFERÊNCIAS

- Braga, R. (2009), Viticultura de Precisão. AJAP – Associação dos Jovens Agricultores de Portugal (editores).
- Cortez P., Cerdeira, A., Almeida, F., Matos, T. e Reis, J. (2009), “Modeling wine preferences by data mining from physico chemical properties”. Decision Support Systems, Elsevier, vol. 47, pp. 547-553.
- Gama, J., Carvalho, A.P.L., Faceli, K., Lorena, A.C. e Oliveira, M. (2015), Extração de conhecimento de dados – Data Mining. Edições Sílabo.
- Han, J., Kamber, M. e Pei, J. (2012), Data Mining. Concepts and Techniques. Morgan Kaufmann. Elsevier.
- Knime, (2015), Konstanz Information Miner Software, version 3.2.1, KNIME GmbH, Konstanz, Germany.
- Ribeiro J., Neves, J.M., Machado, J. e Novais, P.J. (2009a), “Wine vinification prediction using data mining tools”. ECC'09 Proceedings of the 3rd international conference on European computing conference. Computing and Computational Intelligence. WSEAS. pp. 78-85. <http://hdl.handle.net/1822/18957>.
- Ribeiro, J., Neves, J., Sanchez, J., Novais, P. e Machado, J. (2009b), “Vinification Mining – A Case Study on Wine Production”. Universidade do Minho. <http://repositorium.sdum.uminho.pt/handle/1822/18924>.
- Silva, J.R.M. (2015), “Novas tecnologias na gestão da vinha”. Diapositivos. Conferência Internacional da Vinha e do Vinho. Reguengos de Monsaraz.
- Witten, I. H., Frank, E. e Hall, M. A. (2011), Data Mining – Practical Machine Learning, Tools and Techniques. Morgan Kaufmann. Elsevier.