Avaliação de Resultados em Classificação Supervisionada

Ana Sousa Ferreira
Faculdade de Psicologia, Universidade de Lisboa, Business Research Unit (BRU-IUL), Lisboa, Portugal,
asferreira@psicologia.ulisboa.pt

Anabela Marques
Escola Superior de Tecnologia do Barreiro, IPS, CIIAS, Barreiro, Portugal,
anabela.marques@estbarreiro.ips.pt

Palavras–chave: Avaliação de resultados; Classificação Supervisionada; Combinação de modelos.

Resumo: Em problemas discretos de classificação supervisionada observa-se, frequentemente, que as observações mal classificadas são diferentes para diferentes modelos. Deste modo, a abordagem pela combinação de modelos tem vindo a ser considerada uma mais valia neste domínio. A avaliação de resultados em classificação baseia-se, habitualmente, na taxa de casos bem classificados. No entanto, alguns autores têm vindo a advertir que esta medida pode não analisar corretamente a qualidade de um modelo. Neste trabalho, pretendemos explorar a avaliação de desempenho de novos modelos combinados, comparando a medida de avaliação mais usual com outros tipos de medidas como a Sensibilidade, Especificidade ou Precisão, Medidas de associação ou concordância ou o Índice de Huberty.

1 Introdução

Em Estatística, fala-se de um problema de classificação supervisionada quando se pretende identificar qual a classe, entre várias definidas *a priori*, a que pertence uma nova observação, baseando-se na

informação fornecida por uma amostra, onde a classe de cada observação é conhecida. Por exemplo, quando se pretende atribuir um diagnóstico a um certo paciente, descrito por um conjunto de características observadas (sexo, pressão arterial, presença ou ausência de alguns sintomas, ...), entre meningite viral ou bacteriana ou quando se precisa de decidir se um dado *email* pertence à classe de *emails* "*spam*"ou "*não spam*". Em qualquer dos exemplos referidos, para identificar a classe a que pertence a nova observação, utiliza-se a informação de uma amostra, denominada habitualmente amostra de treino, tentando perceber se o "perfil" da nova observação, será mais provável de ocorrer na Classe 1 ou na Classe 2.
No caso discreto, os resultados que podem ser observados são denominados por estados. Exemplificando, no caso mais simples de apenas duas variáveis binárias (0 - ausência do sintoma e 1 - presença do sintoma) podem ocorrer os estados seguintes: 00, 01, 10 e 11. Então, os resultados observados numa amostra de treino podem ser apresentados como na Tabela 1:

Tabela 1: Exemplo de estados observados numa amostra de treino

Estados		Classe 1	Classe 2
1	00	4	0
2	01	5	1
3	10	0	4
4	11	1	5
Total		10	10

No caso discreto, o modelo mais natural é o Modelo Multinomial Completo (MMC) onde a probabilidade de ocorrer um certo estado se a observação pertencer a uma determinada classe é estimada pela frequência relativa observada na amostra-treino, em cada classe ([5]). Contudo, quando o número de variáveis consideradas aumenta um pouco, o número de estados possíveis sofre, de imediato, um enorme incremento. Note-se, por exemplo que, no caso mais simples de variáveis binárias, se forem consideradas 10 variá-

veis, teremos $2^p = 2^{10} = 1024$ estados possíveis, exigindo amostras de grandes dimensões para permitir a estimação de todos os parâmetros do modelo.
Deste modo, em classificação supervisionada, no caso discreto, existe frequentemente um problema de dimensionalidade, denominado mesmo na literatura como "a maldição da dimensionalidade":

- Na generalidade dos modelos, o número de parâmetros a ser estimado é demasiado grande;
- Em Ciências Sociais e Humanas, onde o caso discreto tem grande prevalência, não raramente as amostras têm pequena dimensão.

Consequentemente, gera-se facilmente um número elevado de estados não observados, dificultando a estimação de todos os parâmetros. Este problema conduz a que a maior parte dos métodos revelem um fraco desempenho, especialmente quando as classes são pouco separadas e não balanceadas ([6]). Deste modo, em problemas de classificação discretos, a abordagem pela combinação de modelos tem vindo a ser referida como uma mais-valia, resultante de os erros de má classificação observados em diferentes modelos tenderem a ocorrer em objetos diferentes ([2], [7], [11]).
Quando se compara o desempenho destes novos modelos combinados com os modelos originais, usa-se geralmente a Taxa de casos bem classificados ou de casos mal classificados. No entanto, esta medida de avaliação pode não analisar corretamente a qualidade de um modelo, particularmente quando as classes são não balanceadas. Neste trabalho, pretendemos explorar a avaliação de resultados em classificação supervisionada, comparando a taxa de casos bem classificados com outros tipos de medidas ([4], [9]).

2 Combinação de modelos

Geralmente, em face de um problema de classificação complexo, estimam-se diversos modelos e, posteriormente, um único modelo

é selecionado, baseado num determinado critério de validação. Contudo, os modelos descartados contêm frequentemente alguma informação importante sobre o problema de classificação, que se perde pelo facto de se considerar um único modelo ([2]). Por outro lado, verifica-se muitas vezes que as observações mal classificadas são diferentes para diferentes modelos. Este conhecimento tem conduzido a um número crescente de publicações sobre abordagens de combinação de modelos, ainda que referenciadas sob nomes diversos como *Blending*, *Bagging* e *Arcing* entre outros ([8]).

Em problemas de classificação discretos, consideram-se dois modelos de referência: o já referido Modelo Multinomial Completo (MMC) e o Modelo de Independência Condicional de ordem um (MIC) que considera as variáveis independentes dentro de cada classe, reduzindo assim o número de parâmetros a estimar de $2^p - 1$ para p, em cada classe.

Na abordagem de combinação de modelos proposta por Sousa Ferreira ([11]) e continuada por Marques ([7]) consideraram-se combinações lineares de dois modelos de referência no campo discreto. Inicialmente, Sousa Ferreira ([11]) propôs uma combinação linear entre os modelos de referência acima mencionados, MMC e MIC. Esperava-se, naturalmente, que esses dois modelos conduzissem a classificadores diferentes em muitas circunstâncias, dado que o primeiro pressupõe a existência de relações entre as p variáveis binárias e o segundo, considera que dentro de cada classe as p variáveis são independentes. O modelo combinado MMC-MIC resulta da combinação linear entre os dois modelos usando um único coeficiente β, $0 \leq \beta \leq 1$, conduzindo a um modelo intermédio entre MMC e MIC. As várias estratégias adoptadas para estimar β produzem diferentes modelos combinados ([2]). Num segundo momento, verificando que o modelo MMC revela grande dificuldade em estimar todos os parâmetros do modelo quando as amostras têm pequena dimensão, Marques ([7]) desenvolveu uma combinação linear entre o Modelo Gráfico Decomponível (MGD) ([3]) e o modelo MIC, usando também um único coeficiente β, com valores no intervalo [0,1]. O modelo MGD considera as interações mais importantes entre pares de va-

riáveis para estimar a função de probabilidade por classe, utilizando uma estrutura de árvore (grafo), que se baseia na informação mútua. O algoritmo considerado foi o proposto por Chow e Liu ([3]). Também neste caso se esperava que estes dois modelos conduzissem a classificadores diferentes, uma vez que o primeiro pressupõe a existência de interações entre as p variáveis binárias e o segundo, considera que dentro de cada classe as p variáveis são independentes. No caso de múltiplas classes *a priori*, ambos os modelos combinados, MMC-MIC e MGD-MIC, consideram o Modelo de Emparelhamento Hierárquico (MHIERM) que decompõe um problema de múltiplas classes em múltiplos problemas de duas classes ([2], [11]).
A abordagem de combinação de modelos proposta por Sousa Ferreira e continuada por Marques foi avaliada comparativamente com outros algoritmos existentes quer sobre dados reais quer simulados ([7],[8],[11]) revelando uma boa capacidade preditiva em casos de amostras de pequena ou moderada dimensão.
Neste trabalho, pretendemos continuar a explorar os resultados desta abordagem de combinação de modelos, usando outras medidas de avaliação da qualidade dos modelos ([4], [9]).

3 Medidas de avaliação

Na literatura de Estatística, a avaliação do desempenho de qualquer modelo de classificação supervisionada baseia-se, genericamente, na diagonal da matriz de confusão que confronta as classes preditas pelo modelo com as classes originais.
Diversas medidas de desempenho de um modelo podem ser definidas a partir dessa matriz, sendo tradicionalmente usadas a Taxa de casos bem classificados ou de casos mal classificados, estimadas por resubstituição, amostra-teste ou validação cruzada. Diferentes autores têm vindo a referir, contudo, que estas estatísticas tradicionais de avaliação de resultados em classificação podem não analisar corretamente a qualidade de um algoritmo ou modelo ([4], [9], [10]).

Num problema de classificação discreto, com duas classes, tem-se a matriz de confusão apresentada na Tabela 2:

Tabela 2: Matriz de Confusão

		Classes preditas	
		1	2
Classes verdadeiras	1	a	b
	2	c	d

onde:
a - nº de casos bem classificados na classe 1
b - nº de casos da classe 1 classificados na classe 2
c - nº de casos da classe 2 classificados na classe 1
d - nº de casos bem classificados na classe 2

Em Medicina, os valores de a, b, c e d são denominados habitualmente por *Verdadeiros Positivos*, *Falsos Negativos*, *Falsos Positivos* e *Verdadeiros Negativos*, respetivamente. Esta terminologia, que se generalizou a muitos outros campos de aplicação, deriva de, por exemplo, se constatar que um exame complementar de diagnóstico indica que um certo sujeito está doente mas, na realidade, o sujeito está saudável. Teremos, então, um caso de *Falso Positivo*. Algumas medidas de avaliação em classificação estão associadas a este tipo de problemas de classificação.
Na Tabela 3 apresentam-se algumas medidas de avaliação baseadas na matriz de confusão. A *Taxa de casos bem classificados* ou *Acuracia* (Ac) é a medida mais comummente usada e mede a eficiência global do modelo. Na verdade, a *Acuracia* pretende responder à questão: "Globalmente, com que frequência o modelo de classificação decide corretamente?"
A *Taxa de casos bem classificados na classe 1* é também denominada por *Sensibilidade* e mede a eficiência na classe 1 e a *Taxa de casos bem classificados na classe 2* é também denominada por *Especificidade* e mede a eficiência na classe 2. Como referido anteriormente, se

Tabela 3: Medidas de avaliação baseadas na matriz de confusão

Medidas	Definição
Taxa de casos bem classificados ou *Acuracia*	$\frac{a+d}{a+b+c+d}$
Taxa de casos bem classificados na classe 1 ou *Sensibilidade*	$\frac{a}{a+b}$
Taxa de casos bem classificados na classe 2 ou *Especificidade*	$\frac{d}{c+d}$
Precisão	$\frac{a}{a+c}$

um exame complementar de diagnóstico indicar que um certo sujeito está doente mas, na realidade, esse sujeito estiver saudável, temos um caso de *Falso Positivo*, pelo contrário, se esse sujeito estiver mesmo doente, temos um caso de *Verdadeiro Positivo*. Do mesmo modo, se o exame complementar de diagnóstico indicar que o sujeito não está doente e, de facto, esse sujeito estiver saudável, estamos perante um caso *Verdadeiro Negativo*. Naturalmente, um bom modelo de classificação deverá ser capaz de identificar quer os casos de *Verdadeiro Positivo* quer os de *Verdadeiro Negativo*.
A *Sensibilidade* é, exatamente, a taxa de casos *Verdadeiro Positivo* e a *Especificidade* a taxa de casos *Verdadeiro Negativo*, respondendo respetivamente às questões "Se um sujeito pertence à Classe 1, qual a frequência com que o modelo de classificação consegue identificar corretamente a classe desse sujeito?", e, "Se um sujeito pertence à Classe 2, qual a frequência com que o modelo de classificação consegue identificar corretamente a classe desse sujeito?". Finalmente, a *Precisão*, também denominada por valor preditivo positivo, mede a exatidão do modelo respondendo a outra questão: "Entre os casos que o modelo classificou como *Positivos*, isto é, pertencentes à Classe

1, quantos efetivamente o são?". Um valor de *Precisão* elevado revela, pois, um modelo que é um bom preditor.

As medidas de avaliação usadas, em geral, não fornecem um equilíbrio entre os casos falsos positivos (c) e os falsos negativos (b). As medidas de avaliação combinadas, apresentadas na Tabela 4, tentam obter uma melhor paridade entre eles.

Tabela 4: Medidas de avaliação combinadas

Medidas	Definição
Taxa de casos bem classificados balanceada	$\frac{\text{Sensibilidade}+\text{Especificidade}}{2}$
Média Geométrica entre Sensibilidade e Especificidade	$\sqrt{\text{Sensibilidade} \times \text{Especificidade}}$
Medida F	$\frac{2\times\text{Sensibilidade}\times\text{Precisão}}{\text{Sensibilidade}+\text{Precisão}}$

A *Taxa de casos bem classificados balanceada* ou *Acuracia balanceada* é a média aritmética entre a *Sensibilidade* e a *Especificidade* e, comparada com a *Acuracia global*, tenderá a ser menor quando o modelo não consegue classificar igualmente bem as duas classes. A *Média Geométrica* entre as duas medidas mede o equilíbrio entre a classificação nas duas classes. Um valor de *Média Geométrica* baixo indica um desempenho fraco na classe dita positiva (geralmente, considerada como classe de maior interesse). A *Medida F* combina as medidas *Sensibilidade* e *Precisão*, mesmo quando as classes de dados são verdadeiramente desiquilibradas. As medidas de avaliação já apresentadas anteriormente, sendo genericamente taxas, simples ou combinadas, variam naturalmente no intervalo [0,1].

Um outro tipo de medidas de avaliação, que indicam a associação ou o acordo entre classes verdadeiras e preditas têm vindo a ser referidas por alguns autores. Por outro lado, parece ser relevante avaliar a melhoria efetiva que um modelo introduz relativamente à regra da maioria. Estas medidas de avaliação menos tradicionais em classificação supervisionada são apresentadas na Tabela 5.

O *Coeficiente* ϕ é uma conhecida medida de associação entre duas

Tabela 5: Outro tipo de medidas de avaliação

Medidas	Definição
Coeficiente ϕ	$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$
Estatística K de Cohen	$\frac{A_c - P_{acaso}}{1-P_{acaso}}$, onde $P_{acaso} = (\frac{a+b}{N} \times \frac{a+c}{N}) + (\frac{c+d}{N} \times \frac{b+d}{N})$ e $N = a+b+c+d$
Índice de Huberty	$\frac{P_{cc}-P_m}{1-P_m}$, onde P_{cc} - % casos corretamente classificados e P_m - % casos corretamente classificados de acordo com a regra da maioria

variáveis binárias, podendo tomar valores no intervalo [−1,1]. O sinal positivo deste coeficiente indica um maior número de casos em que o modelo de classificação decidiu corretamente e, o sinal negativo, pelo contrário, revela que existem mais casos de decisão incorreta. Por outro lado, a *Estatística K de Cohen* pode ser definida como a proporção de acordo entre duas classificações após ser retirada a proporção de acordo devida ao acaso, podendo também tomar valores no intervalo [−1,1]. Por último, o *Índice de Huberty* avalia o desempenho de um modelo como o grau de correção da classificação realizada, comparando com a percentagem de casos bem classificados

pela regra da maioria, sendo definido como a razão entre a melhoria efectiva e a melhoria possível na classificação. Este índice é a única medida de avaliação apresentada que pode tomar valores fora do intervalo $[-1,1]$.

4 Resultados Numéricos

Neste estudo, analisaram-se dados simulados, com duas classes e quatro variáveis binárias e consideraram-se três importantes fatores que influenciam o desempenho dos modelos: dados balanceados ou não balanceados, separabilidade das classes (baixa ou elevada) e dimensão das amostras (pequena ou grande). Considerando os oito cenários referidos, especificam-se seguidamente os valores considerados para cada fator: *i*. Equilíbrio - Classes balanceadas quando $n_1 = n_2$ e não balanceadas quando $n_1 = \frac{1}{9} \times n_2$; *ii*. Separabilidade - sendo medida pelo Coeficiente de Afinidade ([1]) definido no intervalo [0,1]. Este coeficiente mede a afinidade ou semelhança entre as classes pelo que, quanto mais pequeno for o seu valor, mais separadas são as classes consideradas e, por isso, a tarefa do modelo de classificação fica simplificada. Deste modo, considerou-se separabilidade baixa quando o coeficiente de afinidade toma valores superiores a 0,7 e elevada quando este coeficiente toma valores inferiores a 0,3; *iii*. Dimensão das amostras - pequena quando n=60 e grande quando n=400.
Considerando os dois graus de separabilidade das classes (Baixa ou Elevada), os dados em análise foram simulados segundo a Distribuição Multinomial com os parâmetros, isto é, as probabilidades de ocorrência das quatro variáveis preditoras binárias, apresentados na Tabela 6.

No estudo apresentado, para cada um dos oito cenários considerados, geraram-se 10 réplicas. Baseados nos 80 conjuntos de dados gerados, pretendemos averiguar a vantagem comparativa do modelo combinado MGD-MIC, usando diversas medidas de avaliação do de-

Tabela 6: Parâmetros da Distribuição Multinomial usados na simulação dos dados, de acordo com o grau de separabilidade (Baixa ou Elevada) entre as duas classes consideradas

	C_1	C_2
S. Baixa	(0,5;0,5;0,5;0,5;0,5;0,5;0,5;0,5)	(0,5;0,5;0,5;0,5;0,5;0,5;0,5;0,5)
S. Elevada	(0,1;0,9;0,7;0,3;0,2;0,8;0,6;0,4)	(0,9;0,1;0,3;0,7;0,8;0,2;0,4;0,6)

sempenho. As medidas de avaliação de desempenho dos modelos foram todas estimadas por *2-fold cross validation*. O desempenho dos modelos, simples ou combinados, são apresentados nas Tabelas 7, 8, 9 e 10, onde se mostram os resultados médios intra-cenários (e respetivo desvio-padrão), destacando-se a negrito o melhor resultado obtido em cada cenário.

Na Tabela 7, podemos notar que, no caso de maior complexidade, quase todas as medidas elegem o modelo combinado como o melhor modelo, embora a sua capacidade preditiva seja apenas ligeiramente superior à dos modelos originais. Note-se, ainda, que neste caso de classes balanceadas, $\phi = Kappa = I.Huberty$. Quando a separabilidade é elevada, o modelo combinado revela um desempenho muito semelhante ao do modelo MIC, ambos demonstrando uma excelente capacidade preditiva. As outras medidas de avaliação mostram também resultados elevados, podendo pois dizer-se que os modelos MIC e MGD-MIC obtêm uma melhoria efectiva na classificação.

Na Tabela 8, quando a separabilidade é baixa, observa-se a seleção do modelo MGD ou MGD-MIC como o melhor modelo, e também neste caso, de classes balanceadas $\phi = Kappa = I.Huberty$, revelando embora uma melhoria efectiva muito baixa. Quando as classes são bem separadas, o modelo MIC obtém resultados muito semelhantes aos de MGD-MIC mas ainda superiores para algumas medidas.

Na Tabela 9, quando se apresentam os resultados para classes não ba-

Tabela 7: Avaliação do desempenho do modelo combinado no caso de classes balanceadas e amostras de pequena dimensão (resultados médios e desvios padrão intra-cenários)

	$n_1 = n_2 = 30$					
	Separabilidade Baixa			Separabilidade Elevada		
Medidas	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
Tx. Bem Class.	0,60	0,61	**0,62**	**0,94**	0,89	**0,94**
	(0,03)	(0,09)	(0,07)	(0,04)	(0,04)	(0,04)
Sensibilidade	0,60	0,62	**0,66**	0,94	0,90	**0,96**
	(0,07)	(0,07)	(0,10)	(0,04)	(0,07)	(0,07)
Especificidade	**0,60**	0,59	0,59	**0,94**	0,88	0,92
	(0,08)	(0,11)	(0,09)	(0,05)	(0,05)	(0,04)
Precisão	0,60	0,62	**0,63**	**0,94**	0,89	0,93
	(0,03)	(0,09)	(0,07)	(0,04)	(0,04)	(0,04)
Média Geométrica	0,59	0,60	**0,61**	**0,94**	0,89	**0,94**
	(0,03)	(0,09)	(0,08)	(0,04)	(0,04)	(0,04)
Medida F	0,60	0,61	**0,64**	**0,94**	0,89	**0,94**
	(0,04)	(0,08)	(0,08)	(0,04)	(0,04)	(0,05)
Tx. Bem Clas. Bal.	0,60	0,61	**0,62**	**0,94**	0,89	**0,94**
	(0,03)	(0,09)	(0,07)	(0,04)	(0,04)	(0,04)
Coeficiente ϕ	0,20	0,21	**0,25**	0,88	0,79	**0,89**
	(0,05)	(0,18)	(0,15)	(0,08)	(0,07)	(0,09)
Estatística Kappa	0,20	0,21	**0,25**	0,87	0,78	**0,88**
	(0,05)	(0,17)	(0,15)	(0,08)	(0,07)	(0,09)
Índice de Huberty	0,20	0,21	**0,25**	0,87	0,78	**0,88**
	(0,05)	(0,17)	(0,15)	(0,08)	(0,07)	(0,09)

lanceadas e pouco separadas, sobressai, para a maioria das medidas, o modelo combinado. Neste caso, não balanceado, $\phi \neq Kappa \neq I.Huberty$, e os valores obtidos pelo índice de Huberty revelam um pior desempenho do modelo do que se observaria pela aplicação da regra da maioria. Quando as classes são bem separadas, só a *Sensibilidade* não elege o modelo combinado como o melhor modelo.

A análise das classes não balanceadas e amostras de grande dimen-

Tabela 8: Avaliação do desempenho do modelo combinado no caso de classes balanceadas e amostras de grande dimensão (resultados médios e desvios padrão intra-cenários)

	$n_1 = n_2 = 200$					
	Separabilidade Baixa			Separabilidade Elevada		
Medidas	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
Tx. Bem Class.	0,52	**0,54**	**0,54**	**0,93**	0,90	0,92
	(0,02)	(0,02)	(0,02)	(0,02)	(0,02)	(0,02)
Sensibilidade	0,51	0,51	**0,56**	**0,91**	0,88	**0,91**
	(0,03)	(0,05)	(0,04)	(0,03)	(0,05)	(0,03)
Especificidade	0,54	**0,56**	0,51	**0,94**	0,92	0,92
	(0,02)	(0,04)	(0,05)	(0,02)	(0,02)	(0,01)
Precisão	0,52	**0,53**	**0,53**	**0,94**	0,91	0,92
	(0,02)	(0,02)	(0,02)	(0,02)	(0,02)	(0,01)
Média Geométrica	0,52	**0,53**	**0,53**	**0,93**	0,90	0,92
	(0,02)	(0,02)	(0,02)	(0,02)	(0,03)	(0,02)
Medida F	0,52	0,52	**0,54**	**0,92**	0,90	**0,92**
	(0,02)	(0,03)	(0,03)	(0,02)	(0,03)	(0,02)
Tx. Bem Clas. Bal.	0,52	**0,54**	**0,54**	**0,93**	0,90	0,92
	(0,02)	(0,02)	(0,02)	(0,02)	(0,02)	(0,02)
Coeficiente ϕ	0,05	**0,07**	**0,07**	**0,85**	0,80	0,84
	(0,03)	(0,04)	(0,04)	(0,04)	(0,05)	(0,04)
Estatística Kappa	0,05	**0,07**	**0,07**	**0,85**	0,80	0,84
	(0,03)	(0,04)	(0,04)	(0,04)	(0,05)	(0,04)
Índice de Huberty	0,05	**0,07**	**0,07**	**0,85**	0,80	0,84
	(0,03)	(0,04)	(0,04)	(0,04)	(0,05)	(0,04)

são (ver Tabela 10), mostra que, quando pouco separadas, o modelo MGD é eleito por todas as medidas como o melhor modelo, embora com resultados quase sempre iguais aos do modelo combinado. O *Índice de Huberty* volta a revelar um pior desempenho do que se obteria pela regra da maioria. Quando a separabilidade é elevada, o modelo combinado é eleito como o melhor modelo por todas as medidas.

Tabela 9: Avaliação do desempenho do modelo combinado no caso de classes não balanceadas e amostras de pequena dimensão (resultados médios e desvios padrão intra-cenários)

	$n_1 = 6; n_2 = 54$					
	Separabilidade Baixa			Separabilidade Elevada		
Medidas	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
Tx. Bem Class.	0,67	0,62	**0,76**	0,90	0,80	**0,92**
	(0,09)	(0,09)	(0,07)	(0,03)	(0,12)	(0,02)
Sensibilidade	0,67	**0,70**	0,63	0,85	**0,90**	0,85
	(0,19)	(0,15)	(0,11)	(0,17)	(0,12)	(0,12)
Especificidade	0,67	0,62	**0,78**	0,91	0,79	**0,93**
	(0,10)	(0,10)	(0,09)	(0,04)	(0,14)	(0,03)
Precisão	0,20	0,17	**0,24**	0,54	0,40	**0,61**
	(0,07)	(0,05)	(0,05)	(0,09)	(0,13)	(0,10)
Média Geométrica	**0,65**	0,59	0,63	0,85	0,83	**0,88**
	(0,10)	(0,13)	(0,12)	(0,14)	(0,08)	(0,06)
Medida F	0,30	0,29	**0,35**	0,66	0,52	**0,69**
	(0,09)	(0,06)	(0,07)	(0,08)	(0,12)	(0,06)
Tx. Bem Clas. Bal.	0,67	0,66	**0,71**	0,88	0,85	**0,89**
	(0,10)	(0,08)	(0,06)	(0,09)	(0,07)	(0,05)
Coeficiente ϕ	0,22	0,19	**0,28**	0,64	0,51	**0,67**
	(0,13)	(0,10)	(0,08)	(0,09)	(0,13)	(0,07)
Estatística Kappa	0,17	0,13	**0,24**	0,58	0,44	**0,65**
	(0,10)	(0,08)	(0,08)	(0,12)	(0,15)	(0,07)
Índice de Huberty	-2,33	-2,77	**-1,37**	0,02	-0,98	**0,22**
	(0,85)	(0,88)	(0,74)	(0,34)	(1,17)	(0,19)

Em qualquer das tabelas de resultados pode notar-se que, o desvio padrão relativo a todas as medidas apresentadas, é sempre extremamente baixo, próximo de zero, exceto no caso do *Índice de Huberty* em classes não balanceadas.

Tabela 10: Avaliação do desempenho do modelo combinado no caso de classes não balanceadas e amostras de grande dimensão (resultados médios e desvios padrão intra-cenários)

	$n_1 = 40; n_2 = 360$					
	Separabilidade Baixa			Separabilidade Elevada		
Medidas	MIC	MGD	MGD-MIC	MIC	MGD	MGD-MIC
Tx. Bem Class.	0,54 (0,04)	**0,56** (0,03)	**0,56** (0,04)	0,90 (0,03)	0,89 (0,05)	**0,91** (0,03)
Sensibilidade	0,54 (0,06)	**0,58** (0,07)	0,57 (0,07)	0,90 (0,06)	0,89 (0,04)	**0,91** (0,04)
Especificidade	0,54 (0,05)	**0,55** (0,03)	**0,55** (0,05)	0,90 (0,03)	0,89 (0,05)	**0,91** (0,03)
Precisão	0,12 (0,02)	**0,13** (0,02	**0,13** (0,02)	0,52 (0,08)	0,52 (0,12)	**0,54** (0,10)
Média Geométrica	0,53 (0,04)	**0,56** (0,05)	**0,56** (0,04)	0,90 (0,04)	0,89 (0,03)	**0,91** (0,03)
Medida F	0,19 (0,03)	**0,21** (0,11)	**0,21** (0,03)	0,66 (0,08)	0,64 (0,10)	**0,67** (0,08)
Tx. Bem Clas. Bal.	0,54 (0,04)	**0,57** (0,06)	0,56 (0,04)	0,90 (0,04)	0,89 (0,03)	**0,91** (0,03)
Coeficiente ϕ	0,05 (0,05)	**0,08** (0,07)	**0,08** (0,05)	0,64 (0,08)	0,62 (0,12)	**0,65** (0,08)
Estatística Kappa	0,03 (0,03)	**0,05** (0,08)	**0,05** (0,03)	0,61 (0,09)	0,59 (0,12)	**0,62** (0,08)
Índice de Huberty	-3,59 (0,45)	**-3,44** (0,72)	**-3,44** (0,45)	0,04 (0,32)	-0,09 (0,46)	**0,08** (0,31)

5 Conclusões

Pensando no objetivo de avaliar o desempenho do modelo combinado comparativamente aos modelos originais, a medida de avaliação usada não parece influenciar a decisão, revelando o modelo combinado particular interesse em situações com nível de complexidade elevado, nomeadamente com amostras de pequena dimensão. No caso balanceado, o modelo eleito como o melhor é o mesmo quer

com a medida tradicional quer com outra medida como a *Taxa de Bem Classificados Balanceada*. No caso não balanceado, com grande desequilíbrio entre a dimensão das classes, o modelo combinado mostra também o seu interesse, mesmo quando a separabilidade é elevada. O modelo MIC revela um bom desempenho quando o nível de complexidade não é demasiado elevado e o modelo MGD só consegue revelar-se superior aos outros dois modelos quando as amostras não têm pequena dimensão e o nível de complexidade não é demasiado elevado. Como esperado, relativamente à comparação entre as medidas de avaliação, as medidas mais usuais e as combinadas mostram resultados muito semelhantes quando as classes têm dimensões pouco desiquilibradas. Quando se regista um forte desiquilíbrio entre a dimensão das classes, as medidas combinadas revelam, então, o seu interesse. Por outro lado, o *Coeficiente* ϕ, a *Estatística Kappa* e o *Índice de Huberty* fornecem claramente uma informação de carácter diferente sobre o classificador, cuja interpretação precisa de ser mais explorada, provavelmente em aplicações com dados reais. As medidas *Sensibilidade* e *Especificidade* só revelam particular interesse quando, num certo campo de aplicação como, por exemplo, em Medicina, um dos erros de classificação é considerado particularmente importante.

A avaliação dos resultados em Classificação Supervisionada continuará a ser explorada recorrendo quer a dados simulados (considerando um maior número de réplicas em cada cenário e um maior número de cenários) quer a dados reais, particularmente no caso de classes não balanceadas, procurando compreender melhor o interesse de outras medidas de avaliação menos tradicionais, como por exemplo, o *Índice de Huberty*.

Referências

[1] Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis. *Methods of Operations Research*, 53, 507–512.

[2] Brito, I., Celeux, G., Sousa Ferreira, A. (2006). Combining methods in supervised classification: A comparative study on discrete and continuous problems. *REVSTAT - Statistical Journal*, 4, 201–225.

[3] Celeux, G., Nakache, J. P. (1994). *Analyse Discriminante sur Variables Qualitatives*. Celeux,G., Nakache, J.P.(eds.), Polytechnica.

[4] Ferreira, A. S., Cardoso, M. G. (2013). Evaluating Discriminant Analysis Results. In: Lita da Silva J., Caeiro F., Natário I., Braumann C. (eds.): *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and and Other Statistical Applications. Studies in Theoretical and Applied Statistics*,155–162, Springer, Berlin, Heidelberg.

[5] Goldstein, M., Dillon, W.R. (1978). *Discrete Discriminant Analysis*. Wiley and Sons.

[6] Ho, T.K., Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 289–300.

[7] Marques, A. (2014). *Análise Discriminante sobre Variáveis Qualitativas*. Tese de Doutoramento, ISCTE - Instituto Universitário de Lisboa.

[8] Marques, A., Sousa Ferreira, A., Cardoso, M. (2017). Performance of Combined Models in Discrete Binary Classification. *Methodology* 13(1), 23–37.

[9] Paik, H. (1998). The effect of prior probability on skill in two-group discriminant analysis. *Quality and Quantity*, 32(2), 201–211.

[10] Santafe, G., Inza, I., Lozano, J.A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4), 467–508.

[11] Sousa Ferreira, A. (2000). *Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas*. Tese de Doutoramento, Universidade Nova de Lisboa.