Novas análises e experimentos em modelo local-global para classificação

Rodrigo T. Peres
CAp-UERJ
Rio de Janeiro, Brasil

Carlos E. Pedreira
COPPE-PEE e Faculdade de Medicina, UFRJ
Rio de Janeiro, Brasil

***Resumo* — Apresentam-se novas análises para um procedimento com um enfoque local-global para classificação dicotômica anteriormente proposto. Entre os avanços está uma extensão para o contexto multiclasse. Talvez a principal contribuição seja uma análise sobre a quantidade de protótipos - que reflete o quanto o modelo é local- na fase de particionamento não supervisionado. Quanto mais local o modelo, maior é sua complexidade. Nos experimentos numéricos mostra-se entre outros aspectos relevantes que o modelo se comporta bastante bem em situações de amostras com poucas observações e com desbalanceamento entre classes.**

Keywords—classificação de padrões, classificação local, local-global, regra de Bayes, quantização vetorial.

I. INTRODUÇÃO

A tarefa de um classificador consiste em associar a cada observação um, e somente um rótulo, correspondendo a uma classe dentre um conjunto de n classes possíveis [1]. Utiliza-se para tal um conjunto de atributos medidos de cada uma das observações.

Os classificadores podem ter uma abordagem local ou global. No contexto global, todo o conjunto de observações do conjunto dentro-da-amostra é utilizado conjuntamente para o treinamento do modelo. Um exemplo de um método clássico que toma uso desta abordagem são as redes neurais [3]. Já em abordagens locais, a tarefa de associar uma observação a uma das classes é realizada a partir de subconjuntos da amostra escolhidos de acordo com a localização da observação no espaço. O k-nn [4], que classifica uma observação a partir da classe mais comum dentre seus k vizinhos, é um exemplo de enfoque local. Outra abordagem local muito utilizada para classificação são as Máquinas de Vetores de Suporte (SVM) [5], que constrói o hiperplano separador de classes através dos vetores de suporte, um subconjunto da amostra original.

Em [2], foi proposto um procedimento com um enfoque local-global para classificação dicotômica. Nessa contribuição anterior, o espaço de observações era particionado gerando uma divisão do espaço em células. Cada nova observação era, em seguida, associada a uma destas células. Na sequencia, dois critérios de classificação foram utilizados: se a célula possuía todas as observações associadas a apenas uma classe, a nova observação recebia o mesmo rótulo. Se a célula possuía observações de classes diferentes, era feito um procedimento baseado na regra de Bayes.

Na presente comunicação, apresentamos novas análises para o modelo proposto em [2], incluindo uma extensão para o contexto multiclasse. É feita uma análise sobre a quantidade de protótipos -que reflete o quanto o modelo é local- na fase de particionamento não supervisionado. Quanto mais local for o modelo, maior a sua complexidade, como será discutido na sequencia. Também se analisa situações de amostras com poucas observações e com desbalanceamento entre classes.

II. METODOLOGIA

Nesta seção será apresentado o fundamento metodológico do método desta contribuição, composto por uma síntese da metodologia originalmente proposta em [2] e uma extensão para um ambiente multiclasse, apresentada na presente comunicação.

Seja X uma amostra de tamanho m em um espaço p dimensional composta por observações $\{x_1, x_2, ..., x_m \in R^p, \forall i=1,...,m\}$. Cada observação x_i está associada a um, e somente um rótulo representando uma classe entre n possíveis.

O procedimento proposto em [2] consiste inicialmente em particionar o espaço através de uma técnica não supervisionada, conhecida como LBG [7]. Este algoritmo utiliza protótipos representantes que segmentam o espaço em células, onde cada observação fica associada a um único protótipo. É importante notar que, dentro de uma filosofia não supervisionada, os rótulos das classes não são levados em consideração e evidentemente é possível que haja células com observações de diferentes classes. Chamamos uma célula de 'heterogênea' quando pelo menos duas classes estão presentes. Uma célula na qual todas as observações pertencem a uma única classe é chamada de 'homogênea'.

Para determinar a qual classe uma observação está associada, em primeiro lugar verifica-se a qual protótipo esta observação está mais próxima. Se a célula cujo protótipo representante for homogênea, a classificação é trivial e a observação é associada à mesma classe das outras observações da célula. Se a célula mais próxima for heterogênea, um esquema de classificação supervisionado inspirado na regra de Bayes é realizado. Sem perda de generalidade, vamos focar em uma célula heterogênea arbitrária, C.

Pode-se estimar a probabilidade a priori de cada uma das n classes em C através da frequência de cada classe na célula. Assim:

$$f_j = \frac{\#S_j}{\#C} \quad (1)$$

onde # representa cardinalidade e $S_j = \{x_i \in C \wedge x_i \in \text{classe } j\}$, j = 1,...,n.

A verossimilhança pode ser estimada pelo inverso da distância entre x_i e a média da classe j dentro de C, ou seja,

$$\hat{L}_j = \frac{1}{d(x_i, m_{c,j})} \quad (2)$$

onde $m_{c,j}$ é a média das observações pertencentes a classe j dentro de C e d é a distância Euclidiana; j=1,...,n.

Usando (1) e (2) pode-se estimar a chance de uma observação x_i pertencer a uma das n classes em C da seguinte forma:

$$y_j = f_j \hat{L}_j \quad (3)$$

Se não existirem observações em C para alguma classe i, $1 \leq i \leq n$, então y_i será igual a zero. Após calcular (3) para todas as classes em C, x_i será associada à classe ξ tal que $\xi = \max(y_j)$, j = 1,..., n.

III. RESULTADOS E DISCUSSÃO

Nesta seção os resultados dos experimentos numéricos serão apresentados intermeados pela discussão associada a estes experimentos.

Como acontece em qualquer aplicação de aprendizado por máquina, em modelos de classificação '*data-driven*' é fundamental que o erro fora-da-amostra acompanhe o erro dentro-da-amostra, indicando a potencialidade de generalização [8]. É esperado que, em modelos mais complexos que envolvam mais parâmetros, haja mais dificuldade de manter o erro fora-da-amostra próximo ao erro dentro-da-amostra. É importante notar que, no caso da técnica proposta nesta comunicação, a complexidade está relacionada ao número de protótipos utilizados. O aumento do número de protótipos implica na utilização de mais parâmetros livres e consequentemente em mais complexidade. Em paralelo, quanto mais protótipos forem utilizados, mais local será o modelo, já que se espera que as células (definidas na seção anterior) contenham cada vez menos observações associadas.

Três experimentos são apresentados sendo o comportamento do erro fora-da-amostra com relação ao erro dentro-da-amostra focos importantes.

A. *Experimento 1*

Neste experimento se utiliza um banco de dados sintético, gerado especialmente para a análise em questão. Foram geradas 1000 observações provenientes de duas distribuições Normais, cada uma destas emulando uma classe com 500 observações. Na figura 1, podem-se observar os dados (classe 1 em azul e classe 2 em vermelho). Na tabela 1 apresenta-se o desempenho do método para diferentes números de protótipos e a quantidade de células homogêneas e heterogêneas (veja definições na seção II).

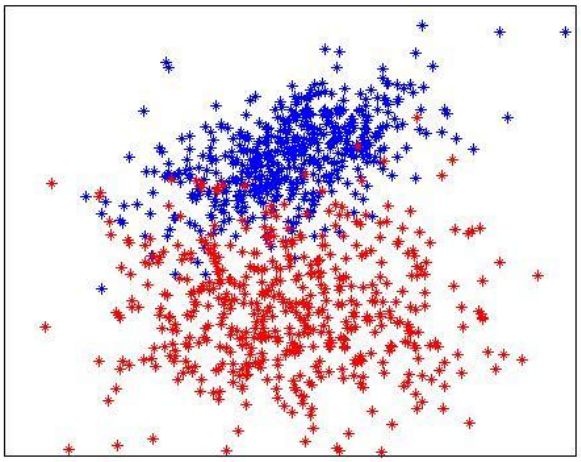

Fig. 1: Exemplo sintético

Observe-se que, como esperado, o desempenho dentro da amostra melhora à medida que o número de protótipos aumenta (Tabela 1), ou seja, o nível de acerto dentro-da-amostra melhora à medida que o algoritmo se torna mais local. É importante mais uma vez ressaltar que, ao mesmo tempo em que se torna mais local, o modelo se torna mais complexo com o aumento de parâmetros envolvidos.

Como era esperado, o desempenho fora-da-amostra não acompanha totalmente a melhora de desempenho dentro-da-amostra à medida que o modelo vai se tornando mais complexo. Ou seja, a diferença entre o erro dentro e fora da amostra se distancia conforme se aumenta a complexidade do modelo (veja tabela 1). Para um aumento de 4 para 512 protótipos verifica-se uma melhora de desempenho de quase 8% dentro-da-amostra e de apenas 1,1% fora-da-amostra.

Em relação à frequência de células homogêneas e heterogêneas, quando o modelo utiliza menos parâmetros, há uma predominância de heterogeneidade. Isto é facilmente explicado uma vez que, com poucas células, a maioria destas possuirá observações de classes diferentes. À medida que se aumenta a complexidade, com a adição de mais protótipos, a predominância passa a ser de células homogêneas. Neste caso, muitas células possuem observações de apenas uma das classes, enquanto que as células próximas à fronteira de decisão possuem observações de classes distintas. Observe-se ainda que uma célula homogênea evidentemente propicia uma classificação muito simples fora da amostra, já que toda observação capturada por um protótipo de uma célula homogênea, recebe automaticamente o rótulo da única classe presente na célula. Para as observações capturadas por protótipos de células heterogêneas, aplica-se a regra de classificação descrita na seção anterior. De acordo com o

algoritmo original, o número total de células pode ser diferente do número de protótipos. Isto se dá porque há um processo de limpeza e protótipos que tenham uma célula vazia (não capturam nenhuma observação dentro da amostra), são retirados antes da classificação.

B. Experimento II – Waveform

Neste experimento, utilizaram-se dados disponíveis no repositório da UCI (http://archive.ics.uci.edu/ml/). Usamos aqui o 'banco waveform' Trata-se de um problema com 3 classes, com 40 atributos, todos estes com ruído adicionado (gaussiano com média zero e variância um). Os últimos 19 atributos são completamente ruidosos. Das 5000 observações, foram selecionadas 3000 para o conjunto dentro-da-amostra e 2000 observações fora-da-amostra.

Com o objetivo de verificar o comportamento do desempenho do modelo para tamanhos distintos do conjunto dentro-da-amostra, criou-se subconjuntos (das 3000 observações disponíveis) com tamanhos 200, 500 e 1000. O desempenho para estes 4 conjuntos está apresentado na Tabela 2. No subconjunto com 200 observações há no máximo 128 protótipos e com 500, no máximo 256.

Note na tabela 2 que, com poucos protótipos, por exemplo, 4, o comportamento é excessivamente global levando a um desempenho ruim do erro dentro-da-amostra mesmo com 3000 observações. Por outro lado, com 256 protótipos foram necessárias 3000 observações para manter a diferença entre os desempenhos dentro e fora da amostra próximos a 4%. Confirma-se assim o esperado teoricamente, à medida que se aumenta a complexidade do modelo é necessário aumentar também o tamanho da amostra, caso contrário o modelo perde a capacidade de generalização. Mantidos os 256 protótipos, a diferença entre os desempenhos dentro e fora da amostra foi de 14% para 1000 observações e de mais de 18% para 500. Para 4 protótipos, são necessárias apenas 200 observações para manter a diferença dos desempenho dentro e fora da amostra em torno de 5%. Entretanto, neste caso, o desempenho dentro da amostra é relativamente baixo. Ou seja, é preciso, como esperado, encontrar um balanço entre a complexidade do modelo que viabiliza um melhor desempenho dentro-da-amostra, com uma quantidade de observações suficientemente grande, para esta complexidade, o que possibilita que o desempenho dentro-da-amostra se mantenha próximo ao desempenho fora-da-amostra.

C. Experimento III – Isolet

Neste experimento utilizamos o banco 'Isolet', disponível no repositório da UCI (http://archive.ics.uci.edu/ml/). O objetivo deste experimento é analisar o comportamento do modelo em um banco desbalanceado com relação às classes. Redefinimos as 26 classes originalmente disponíveis as agrupando em duas: classe 1-letra A; classe 2-demais letras. Com isso, das 6238 observações dentro da amostra -seguindo a divisão dentro e fora da amostra proposta no banco original-, 240 ficaram pertencendo à classe 1 e 5998 à classe 2. Para as 1559 observações fora da amostra, 60 ficaram sendo da classe 1 e 1499 da classe 2. Outro fator importante de dificuldade neste banco é a grande quantidade de atributos, 617.

Os resultados aparecem na Tabela 3. Devido às classes estarem desbalanceadas, apresentamos os resultados discriminados por classe. O desempenho para a classe 2, que é a majoritária, é sempre muito bom dentro e fora da amostra. A dificuldade maior, conforme esperado, é em relação à classe com poucas observações. Note que, em uma abordagem pouco local, com 4 protótipos, uma única classe é escolhida, o que resulta num acerto total da classe 2 e em um erro completo da classe 1. Para 8 e 16 protótipos, o desempenho é praticamente o mesmo, com um acerto muito pequeno dentro-da-amostra para a classe 1 e errando todas as observações fora-da-amostra. Com 32 protótipos, já há um ganho real de desempenho para a classe 1, mas ainda muito distante do obtido para a classe majoritária. Finalmente, com 64 protótipos houve um equilíbrio entre o desempenho por classe. Isto mostra que, apesar do método quantizar o espaço vetorialmente sem levar em consideração os rótulos das classes, pode-se obter sucesso para identificação das observações menos frequentes, como esperado, com o aumento da complexidade do modelo.

IV. CONCLUSÕES

Neste artigo apresentamos a extensão de um método de classificação local global de duas para n classes. Além disso, realizamos uma análise focada na complexidade do método proposto e na capacidade do algoritmo aproximar o erro fora da amostra pelo erro dentro da amostra.

Três experimentos foram realizados buscando avançar empiricamente na resposta a três perguntas importantes: (P1) qual a relação entre o erro dentro da amostra e a complexidade do modelo proposto? Com o experimento sintético observamos como esperado que, ao aumentar a complexidade refletida no número de protótipos, o erro dentro da amostra tende a diminuir consistentemente; (P2) como ajustar o modelo para que o erro fora da amostra se mantenha próximo ao erro dentro da amostra mantendo bom desempenho? No experimento 2, verificamos que o aumento da complexidade sugere a necessidade do aumento no número de observações para que seja possível aproximar os dois erros, permitindo o aplicação de um modelo mais complexo. O algoritmo para ter bom desempenho necessita ser local, mas há um ponto em que aumentar a complexidade, fixado o número de observações, pode sugerir fazer com que o erro fora da amostra deixe de seguir o erro dentro da amostra; (P3) como é o desempenho do algoritmo em bancos desbalanceados? No experimento 3 o método obteve um bom desempenho em um banco altamente desbalanceado com um razoável aumento de complexidade.

Trabalhos futuros indicam uma análise em um número maior de bancos de dados. Procedimentos para identificar o número ideal de protótipos (problema comum a todos os algoritmos desta família) podem ser conjecturados. Para bancos desbalanceados, talvez quantizar vetorialmente por classe possa ser uma alternativa.

AGRADECIMENTOS

Rodrigo T. Peres foi parcialmente apoiado pela FAPERJ (programa de apoio ao pós-doutorado E-26/102.481/2010). Carlos E. Pedreira foi parcialmente apoiado pelo CNPq (produtividade em Pesquisa: 305081/2011-0 e Universal 472499/2012) e pela FAPERJ (cientista do Nosso Estado: E26/102.946/2011- CNE).

REFERÊNCIAS

[1] R. Duda, P. Hart, G. Stork, Pattern Recognition, 2nd edition, Wiley, 2001.

[2] R.T.Peres, C.E.Pedreira, "A new local-global approach for classification", Neural Networks, vol. 23, issue 7, pp. 887-891, 2010.

[3] C. M. Bishop, Pattern Recognition and Machine Learning, Informations Science and Statistics, 2006.

[4] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, 2001.

[5] J. Shawe-Taylor, N. Cristianini, Kernel methods for pattern analysis, Cambridge University Press, 2004.

[6] R.T. Peres, C.E. Pedreira, "The generalized risk zone: Observations selection for classification", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, issue 7, pp. 1331_1337, 2009.

[7] Y. Linde, A. Buzo, R.M. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, COM-28(1), pp.84_95, 1980.

[8] Y.S. Abu-Mostafa, M. Magdon-Ismail, H-T Lin, Learning from Data, AMLbook.com, 2012.

TABELA 1: Desempenho fora e dentro (entre parênteses) da amostra e quantidade de células homogêneas e heterogêneas por número de protótipos – experimento 1.

	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**
Acerto (%)	91,1% (90,40%)	92,95% (92,80%)	93,90% (93,00%)	93,75% (94,00%)	93,25% (95,90%)	91,90% (96,90%)	92,00% (97,90%)	92,20% (98,20%)
Células Homogêneas	1	2	7	19	42	96	190	287
Células Heterogêneas	3	6	9	13	22	29	39	36

TABELA 2: Desempenho por número de protótipos e quantidade de observações – experimento 2.

	Método Proposto							
Waveform	**4**	**8**	**16**	**32**	**64**	**128**	**256**	**512**
200 obs	66,70% (72,50%)	68,40% (77,00%)	75,95% (79,50%)	78,60% (87,00%)	75,50% (93,50%)	76,65% (98,00%)	-	-
500 obs	72,35% (75,60%)	77,45% (77,80%)	79,05% (85,40%)	80,05% (87,00%)	80,30% (90,60%)	81,10% (96,20%)	78,70% (97,40%)	-
1000 obs	73,15% (72,10%)	78,40% (79,10%)	82,25% (83,10%)	82,60% (84,20%)	80,90% (84,50%)	80,50% (88,10%)	79,50% (93,50%)	77,00% (97,40%)
3000 obs	74,65% (73,53%)	80,70% (79,37%)	84,50% (82,73%)	82,20% (80,03%)	84,25% (82,33%)	83,15% (81,97%)	82,70% (86,60%)	81,80% (91,03%)

TABELA 3: Desempenho por número de protótipos e classe – experimento 3.

	Método Proposto				
Isolet	**4**	**8**	**16**	**32**	**64**
Classe 1	0% (0%)	0% (0,83%)	0% (0,83%)	68,33% (65,42%)	91,67% (90,42%)
Classe 2	100% (100%)	100% (100%)	100% (100%)	100% (99,80%)	99,80% (99,50%)