I Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Medidas de Comparação de Modelos para Classes Desbalanceadas

Gabriel Gomes Ferreira[1], Jorge Luis Bazán[2]
Universidade de São Paulo

1 Introdução

Uma das principais metodologias quando pensamos em analisar dados é a utilização de métodos de classificação, mais especificamente, de classificadores binários. Ou seja, situações em que se é necessário analisar uma série de atributos para definir se uma observação pertence à uma classe ou à outra. Apesar da existência de diversos metodos para classificadores automáticos, um problema ainda muito comum é a incapacidade deles de lidar com classes desbalancedas.

Além das dificuldade de se construir modelos com boa performance nesse cenário, outro grande agravante é conseguir mensurar a efetividade desses classificadores. Isso ocorre porque medidas que analisam a quantidade de acertos apenas, podem gerar conclusões equivocadas, sendo necessário assim expandir essa análise para a quantidade de falsos positivos, quantidade de falsos negativos e a ponderação eficiente dessas medidas.

Este trabalho tem como objetivo rever as principais medidas de performance para classificadores aplicados em um estudo de simulação para avaliar o desempenho de dois métodos de classificação bastante conhecidos: Regressão Logística (RL) e Floresta Aleatória (FA), no cenário de dados desbalanceados. E com isso, compreender como as medidas se comportam e como podemos comparar mais efetivamente modelos ajustados para classes desbalanceadas.

2 Modelagem

Para a avaliação das medidas em questão, optamos por desenvolver um estudo de simulação a partir de uma base de dados com respostas desbalanceadas. A estratégia para isso foi considerar a distribuição *power logística* [1], que possui um parâmetro extra λ em que podemos definir um nível de assimetria das curvas o qual produz dados desbalanceados em relação ao modelo logístico.

[1] gabriel.gomesf.think@gmail.com
[2] jlbazan@icmc.usp.br

Neste caso os dados são simulados usando o seguinte algoritmo

$$Y_i \sim Bernoulli(p_i), p_i = F_{PL}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{3i} + \beta_5 x_{5i}), i = 1, \ldots, n$$

onde n é o tamanho da amostra simulada, $\beta_0 = 0,5$, $\beta_1 = 1$, $\beta_2 = 1,5$, $\beta_3 = -1$, $\beta_4 = -1,5$, $\beta_5 = 0,5$ são coeficientes de regressão fixos, x_{ji} em que $j = 1, 2, 3, 4, 5$ são covariáveis de tamanho n, as quais foram simuladas de forma que $x_1 \sim U(-4, 4)$, $x_2 \sim U(-3, 3)$, $x_3 \sim U(-2, 2)$, $x_4 \sim U(0, 2)$ e $x_5 \sim N(0, 1)$ e as probabilidades de sucesso p_i foram obtidas considerando a distribuição acumulada da padrão power logística definida por: $\mathrm{F}_{PL}(\beta_0 + \beta_j x_{ji}) = \left(\frac{exp(\beta_0+\beta_1 x_i+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}+\beta_5 x_{5i})}{1+exp(\beta_0+\beta_1 x_i+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}+\beta_5 x_{5i})}\right)^{\lambda}$ onde $\lambda > 0$ é o parâmetro de forma e $j = 1, ..., 5$ [1].

Com a distribuição e as variáveis definidas, estabelecemos uma relação de dependência e simulamos as probabilidades a partir da distribuição em questão. Utilizando um *treshold* de 0,5 categorizamos a nossa variável resposta em duas classes.

A partir dos parâmetros estabelecidos, assumimos um valor de $\lambda = 4$ e realizamos um total de 10000 réplicas de amostras de tamanho 1000. Onde em todas as amostras temos uma relação estabelecida entre as variáveis explicativas e nossa variável resposta.

No estudo obtemos que a média da proporção dos dados simulados é igual a 0,331, sendo o que a réplica com maior valor foi de 0,429 e a de menor valor foi 0,243. Ou seja, estamos trabalhando com um desbalanceamento em torno de 30%.

Com os dados simulados, ajustamos em cada réplica a RL usando a funcao glm do pacote base do R e a FA considerando a funcao $randomForest$ do pacote $randomForest$. Com isso, analisamos para cada observação qual é a probabilidade esperada de pertencer a determinada classe e a nossa variável resposta, definida novamente por um *treshold* de 0,5.

3 Estudo

A partir dos modelos ajustados para cada réplica, selecionamos algumas medidas de comparação de modelos para regressão binária. Utilizamos medidas clássicas na comparação de modelos [3], como a acurácia (ACC)(M4), que considera o total de observações classificadas corretamente, a sensibilidade (TPR)(M5) e a especificidade (TNR)(M6), que passa a avaliar os modelos em relação á quantidade de falsos positivos e falsos negativos.

Também utilizamos medidas propostas para entender melhor o desbalanceamento das classes [2], como o *Jaccard index* ou *critical sucess index* (CSI)(M7), *Sokal and Sneath index* (SSI)(M8), *Faith index* (FAITH)(M9), *distance measure pattern difference* (PDIF)(M10) e o *Gilbert skill score* (GS)(M11) [5].

Além dessas medidas, também calculamos o tempo de execução (M1) para o ajuste dos modelos, a estatística de teste qui-quadrado (M2) e a estatística de teste KS (M3) entre as probabilidades (p) reais e as estimadas. Junto à esses resultados, também capturamos a médias da proporções reais para cada réplica junto com as porporções estimadas em cada modelo além do cálculo de medidas clássicas como *recall* (R)(M12), precisão (P)(M13) e o F1 *score* (F1)(M14).

Obtemos que a média das proporções (Prop) observadas é de 0,332 com um desvio padrão de 0,023, enquanto a média das proporções esperadas no modelo ajustado pela regressão logística é de 0,331 e a do modelo de florestas aleatórias é de 0,330. Ou seja, temos uma diferença muito pequena entre os dois ajustes.

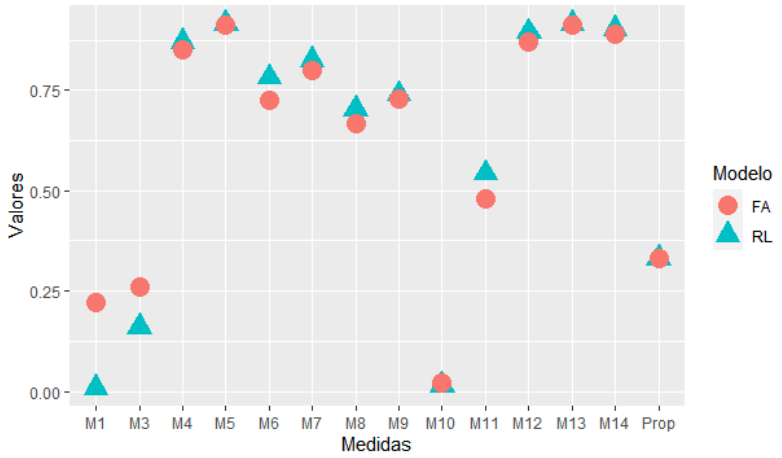

Figura 1: Média das Medidas de desempenho dos modelos usando 10000 replicas.

Podemos ver na Figura 1 que a regressão logística teve uma performance maior em todas as medidas em relação ao modelo de florestas aleatórias.

Na imagem não foi apresentada a medida da estatística Qui Quadrado pois sua ordem de grandeza dificulta a visualização dos demais valores. Mas o resultado obtido foi de 1,391 para o modelo logístico e de 12,382 para o modelo de floresta aleatoria indicando que a regressão logística tem uma performance consideravelmente superior, uma vez que a estatística analisa a distância entre as probabilidades estimadas e as probabilidades reais.

Outra metodologia utilizada para avaliar o desempenho das diferentes medidas consideradas foi a análise fatorial [4], uma técnica de redução de dimensionalidade, que foi utilizada para resumirmos a informação de todas as medidas com o obetivo de encontrarmos algum padrão na forma que as medidas se agrupam. Podemos ver na Figura 2 a apresentação dos dois fatores calculados para cada um dos modelos, em que vemos por meio das setas como as variáveis são divididas em cada fator (foi utilizado um ponte de corte de 0,7 para a criação da imagem).

Podemos observar três interessantes aspectos da forma que as medidas foram agrupadas: a) O primeiro fator em ambos os modelos, de forma generalista agrupa as medidas que tentam balancear de certa forma o total de acertos diante do total de possibilidades. Esta formado pelas medidas CSI, SSI, ACC, PDIF, GS e FAITH. b) O segundo fator em ambos casos agrupou as medidas Recall, Sensibilidade e Especificidade que são medidas que fazem ponderações mais rígidas em relação à acertos e erros dentro de determinada classe, como a quantidade de falsos positivos e falsos negativos. c) A estatística KS e a estatística Qui-Quadrado não são agrupadas em nenhum dos fatores calculados.

4 Conclusão

Conforme podemos perceber, o RL teve performance superior em todas as medidas calculadas, inclusive no tempo de execução, em que levou em média 1% do tempo necessário para a execução do modelo de FA (0,009 versus 0,9). Dando assim, indícios de que este modelo é mais indicado

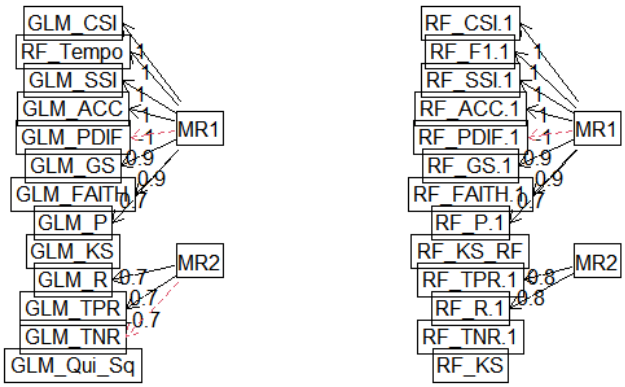

Figura 2: Fatores calculado (RL à esquerda).

para o nível de desbalanceamento proposto no projeto. Porém, isso não significa que o modelo tem um ajuste melhor. Num exemplo similar reportado em de la Cruz [2] foi observado que a RL perde para modelos de regressao binária com funcoes de ligação assimétricas.

Outro ponto importante, é que podemos concluir que quando analisamos comparativamente modelos de classificação binária para classes desbalanceadas, devemos procurar entender a sua eficácia de três formas: avaliando as probabilidades estimadas, entendendo a relação total de acertos da variável resposta e a relação específica de acertos da variável resposta em cada uma das classes conforme a análise dos fatores apresentada na seção anterior.

Referências

[1] Bazán, J., Torres-Avilés, F., Suzuki, A., Louzada, F. *Power and reversal power links for binary regressions: An application for motor insurance policyholders*. Applied Stochastic Models inBusiness and Industry 33, 22–34, 2017.

[2] de la Cruz Huayanay, Alex, et al. *Performance of asymmetric links and correction methods for imbalanced data in binary regression.* Journal of Statistical Computation and Simulation 89.9: 1694-1714, 2019.

[3] Fawcett, Tom. *An introduction to ROC analysis.* Pattern recognition letters 27.8: 861-874, 2016.

[4] Kline, Paul. *An easy guide to factor analysis. Routledge, 2014.*

[5] Schaefer, Joseph T. The critical success index as an indicator of warning skill. Weather and forecasting 5.4: 570-575, 1990.