Laboratório 3 - MDLE - G01

Pedro Carvalho
Mestrado em Informática e Computadores
Instituto Superior de Engenharia de Lisboa
Lisboa, Portugal
47113

Nuno Bartolomeu
Mestrado em Informática e Computadores
Instituto Superior de Engenharia de Lisboa
Lisboa, Portugal
47233

***Abstract*—Neste documento encontra-se presente o relatório do laboratório 3 da displina Mineração de Dados em Larga Escala do 2º semestre do ano 2023/2024. O documento é composto pelas perguntas presentes no enunciado, seguido das respostas às mesmas.**

I. Introdução

Dado o dataset Influenza com 545 atributos, 2 Classes e 2190 instâncias, tem-se como problema/tarefa classificar usando um dataset desequilibrado

II. Visualzar o dataset

A. *Use the sparklyr sdf schema function to check the schema of the df variable.*

Usou-se o a função sdf_schema do sparklyr para ler o esquema do Spark DataFrame, observou-se o nome e o tipo de cada coluna no dataframe.

B. *Check the content of the SPARK data frame df, using the head function.*

Com recurso à função head foi possível observou-se as 6 primeiras instâncias e as 21 primeiras colunas do dataframe assim como os respetivos valores.

C. *Use the stopifnot function to guarantee that the number of columns and rows in df is correct. To achieve this goal, apply the nrow and ncol functions (or the equivalent in Sparklyr), and compare the values with the ones in Table 1.*

Na tabela 1 podemos extrair a informação de que o dataset tem 545 atributos e 2190 instâncias. A função stopifnot permite verificar se neste caso o nº de colunas e linhas é o correto, o número de linhas é o número de instâncias indicado na tabela que é 2190, o número de colunas será o número de atributos mais a classe daí o valor suposto ser 546(nº de atributos + 1). Ao executar o comando este gera um não gera nenhum erro indicando que os valores estão corretos.

III. Seleção de atributos

A. *Use the magrittr's pipe operator, %>% and the select function to reduce df features to the features in the indexes 1, 2, 5, 6, 9, 10, 11, 14, 16, 17, 19, 21, 24, 25, 26, 31, 32, 33, 34, 35, 41, 44, 49, 50, 54. Store the resulting SPARK Dataframe in the df.sel variable. Notice that the first feature is the dependent variable, named CLASS.*

O pipe operator %>% permite direcionar o dataset df para o primeiro argumento da função select que irá selecionar as colunas correspondentes ao índice presente no vetor idx.

B. *Use the head function to overview the resulting dataset.*

Usando a função head é possível visualizar os 25 atributos previamente selecionados.

IV. Utilização de técnicas genéricas de amostragem

A. *Apply the sparklyr sdf random split function to produce two datasets: one for training (2/3) and other for testing (1/3). Use the seed value 123, for this and all random functions from this point forward.*

Usou-se a função set.seed para usar o valor 123. A função random_split permite repartir o dataframe em múltiplos grupos, neste caso 2/3 para treino e 1/3 para teste.

B. *Use the R table function to determine the number of instances for each class in both datasets. Explain why this function cannot be used directly on df.train and df.test.*

O número de instâncias dos dados para treino foram para CLASS==0 de 1356 e para a CLASS==1 foram 73 e o número de instâncias dos dados para teste foram para a CLASS==0 de 719 e para a CLASS==1 é 42. A função table não pode ser diretamente usado nas variáveis df.train e df.test pois a função table espera um objeto R local e essas variáveis são objetos distribuídos do Spark DataFrame que residem em um cluster do Spark.

C. *Use the ml random forest function to generate a classification model, with the formula: "CLASS ~.".*

A função ml_random_forest recebe o dataset df.train e a fórmula "CLASS ~." que indica que o modelo irá prever 'Class' com base todas as outras variáveis no dataset '~.'

D. ***Using the helper function mdle.printConfusionMatrix and ml predict, check the performance of the model. Consider it as the baseline model.***

Na figura 1 podemos observar que a taxa de falsos positivos é muito elevada, o que acaba por levar a um mau diagonóstico no caso de doença. Apresenta um grande nível de precisão, que indica a capacidade do modelo prever com perícia as classificações. O valor de Kappa é baixo concluindo que não existe uma grande concordância entre as classificações do modelo e as reais. O valor positivo previsto é baixo comparado com o valor negativo previsto, ou seja, o modelo é mais capaz de detetar casos negativos em vez de positivos.

```
Confusion Matrix and Statistics: Random Forest Baseline

     0   1
  0 709  10
  1  34   8
False Positive Rate : 0.810
Accuracy            : 0.942
Kappa               : 0.242
Pos Pred Value      : 0.190
Neg Pred Value      : 0.986
```

Fig. 1. Confusion Matrix : Random Forest Baseline

V. Utilização de técnicas de amostragem de correção desequilibrada

A. ***Using the training set from 4.a), apply an undersampling technique to balance the number of cases of each class. Use the function sdf sample. To calculate the fraction use the functions nrow and collect (or sdf nrow alone) on the df.pos.train and df.neg.train variables, and combine them with sdf bind rows. What is the number of instances for each class in the training set after the undersampling?***

Realizou-se um undersampling aleatório simples. Usou-se um filtro para obter as instâncias positivas e negativas, ou seja, aquelas que tiverem a CLASS=1 e CLASS=0, respetivamente. Após obtido o número de instâncias guarda-se numa variável o número mínimo de instâncias. De seguida ocorreu o undersampling das classes positivas e negativas, fez-se uma combinação dos DataFrames e o número de instâncias para a CLASS 0 é 76 e da CLASS 1 é 65.

B. ***Repeat points 4.c) and 4.d), and compare the results with the previous models. Are they better? Are they worst?***

Comparando com o modelo da figura 1(modelo1), o modelo que usa undersampling(modelo2) apresenta uma taxa de falsos positivos menor comparada com do modelo1, concluindo que o modelo1 classifica mais instâncias de classe negativa como positiva.
A precisão do modelo1 é maior do que do modelo2 mas esta não é uma métrica confiável pois em datasets desbalanceados pode ser influenciada pela predominância da classe maioritária. O coeficiente kappa mede a concordância entre as classificações observadas e esperadas, levando em conta a possibilidade de concordância aleatória. O modelo com undersampling apresenta um kappa ligeiramente maior em comparação com o modelo sem undersampling o que indica uma concordância um pouco melhor entre as classificações observadas e esperadas.
O modelo2 tem um valor positivo previsto mais alto (0.786) do que o modelo1 (0.190), o que significa que é mais preciso na previsão da classe positiva.
Ambos os modelos têm valores negativos previstos elevados, mas o modelo1 apresenta um valor ligeiramente mais alto (0.986) em comparação com o modelo2(0.864).
Concluindo, embora o modelo sem undersampling tenha uma precisão geralmente mais alta, o modelo com undersampling apresenta um desempenho melhor em termos de taxa de falsos positivos, valor positivo previsto e coeficiente kappa. Isso sugere que o uso da técnica de undersampling ajudou a melhorar a capacidade do modelo de lidar com o desbalanceamento de classes e a realizar previsões mais precisas para a classe minoritária.

```
Confusion Matrix and Statistics: Random Forest Undersampling

     0   1
  0 621  98
  1   9  33
False Positive Rate : 0.214
Accuracy            : 0.859
Kappa               : 0.325
Pos Pred Value      : 0.786
Neg Pred Value      : 0.864
```

Fig. 2. Confusion Matrix : Random Forest Undersampling

C. ***Using the training set from 4.a), apply an oversampling technique to balance the number of cases of each class. What is the number of instances for each class in the training set after the oversampling?***

Realizou-se a técnica de oversampling aproveitando o algoritmo usado para realizar undersampling só que desta vez em vez de diminuir o número de instâncias da classe maioritária, aumenta-se o número de instâncias da classe minoritária. O número de instâncias para a classe 0 é 1313 e da classe 1 é 1315.

D. ***Repeat points 4.c) and 4.d), and compare the results with the previous models.***

Accuracy: O modelo de Random Forest Baseline tem a maior precisão (0.942), seguido pelo modelo de Random Forest Oversampling (0.930) e, por último, o modelo de Random Forest Undersampling (0.859).

Kappa: O coeficiente kappa mede a concordância entre as previsões do modelo e as observações reais, levando em conta a possibilidade de as previsões ocorrerem por acaso. O modelo de Random Forest Oversampling (0.328) tem o melhor desempenho nessa métrica, seguido pelo modelo

de Random Forest Undersampling (0.325) e, por último, o modelo de Random Forest Baseline (0.242).

Pos Pred Value (Positive Predictive Value): Esta métrica mostra a proporção de verdadeiros positivos em relação a todas as instâncias previstas como positivas pelo modelo. O modelo de Random Forest Undersampling tem o maior valor (0.786), seguido pelo modelo de Random Forest Oversampling (0.690) e, por último, o modelo de Random Forest Baseline (0.190).

Neg Pred Value (Negative Predictive Value): Esta métrica mostra a proporção de verdadeiros negativos em relação a todas as instâncias previstas como negativas pelo modelo. O modelo de Random Forest Baseline tem o maior valor (0.986), seguido pelo modelo de Random Forest Oversampling (0.887) e, por último, o modelo de Random Forest Undersampling (0.864).

False Positive Rate: Esta métrica mostra a proporção de falsos positivos em relação a todas as instâncias negativas reais. O modelo de Random Forest Baseline tem o maior valor (0.810), seguido pelo modelo de Random Forest Oversampling (0.310) e, por último, o modelo de Random Forest Undersampling (0.214).

Portanto, dependendo do contexto e das necessidades específicas do problema, pode-se escolher o modelo que melhor se adapta aos critérios de avaliação, considerando estas métricas. Por exemplo, se é mais importante minimizar os falsos positivos, o modelo de Random Forest Undersampling pode ser preferível, enquanto se a ênfase estiver na precisão geral, o modelo de Random Forest Baseline pode ser mais adequado.

```
Confusion Matrix and Statistics: Random Forest Oversampling

     0   1
  0 638  81
  1  13  29
False Positive Rate : 0.310
Accuracy            : 0.876
Kappa               : 0.328
Pos Pred Value      : 0.690
Neg Pred Value      : 0.887
```

Fig. 3. Confusion Matrix : Random Forest Oversampling

E. Apply Borderline-SMOTE Sampling to balance the number of cases of each class, using the BLSMOTE function from the smotefamily package. The first parameter is a data set without the class. During the oversampling process, use only R data.frame variables. Indicate what are the values that you used for K, C, and method parameters.

Como o desequilíbrio entre classes é significativo o valor default de K (5) achou-se o suficiente para evitar a geração excessiva de amostras sintéticas. Assim como com K, um valor menor de C pode ser mais adequado para identificar regiões onde as instâncias de minoria estão próximas das instâncias de maioria. Dado o desequilíbrio de classe e a necessidade de uma abordagem mais conservadora na geração de amostras sintéticas, o método "type1" pode ser mais apropriado. Este método é menos agressivo na geração de amostras e pode ser mais adequado para conjuntos de dados altamente desequilibrados como o seu.

F. Repeat points 4.c) and 4.d), and compare the results with the previous models

Comparando os modelos sabemos que:
O modelo que apresenta uma taxa de falsos positivos é o Baseline seguido pelo BLSMOTE, Oversampling e Undersampling, sendo o melhor valor do Undersampling pois a taxa indica que ocorre menos vezes, comparado com os outros modelos, casos de falsos positivos. O modelo Baseline apresenta mais uma vez uma maior precisão no seguido de perto pelo BLSMOTE, Oversampling, Undersampling. O modelo BLSMOTE apresenta um maior coeficiente Kappa, depois o Oversampling, Undersampling e Baseline, indica que existe uma maior concordância no modelo BLSMOTE. O modelo de UnderSampling é o melhor a prever os casos positivos, seguido do Oversampling, BLSMOTE e Baseline. Por último o modelo Baseline é o melhor a prever os casos negativos isto devido ao maior número de instâncias da classe 0, logo o modelo irá estar mais preparado para prever esses casos, o 2º melhor modelo é o BLSMOTE, seguido pelo Oversampling e por último o Undersampling.

```
Confusion Matrix and Statistics: Random Forest BLSMOTE

     0   1
  0 696  23
  1  26  16
False Positive Rate : 0.619
Accuracy            : 0.936
Kappa               : 0.361
Pos Pred Value      : 0.381
Neg Pred Value      : 0.968
```

Fig. 4. Confusion Matrix : Random Forest BLSMOTE

VI. Comparação

A. Based on the results achieved, what sampling technique do you think is probably better for this dataset, considering the problem in Table 1? Present a table, where the best results are highlighted in bold. Explain.

Na seguinte tabela encontra-se disponível os valores uma tabela para as diferentes técnicas de amostragem de correção. Para a taxa de falsos positivos(FPR) o melhor valor é aquele que é o menor entre os 4 valores pois isto indica a taxa de ocorrência de falsos positivos, daí ter que ser o menor valor possível, enquanto nas outras métricas quanto maior o valor melhor.
Tendo em conta que o problema/tarefa é classificação usando um conjunto de dados de desequilíbrado e relacionado a uma

doença, as métricas mais relevantes seriam os falsos positivos, ou seja prever corretamente os casos positivos(presença de doença) e o valor preditivo positivo(PPV) que indica a proporção de casos positivos corretamente identificados entre todos os casos identificados como positivos. Observando a tabela podemos concluir que o melhor modelo para este caso é o Undersampling.

TABLE I
COMPARAÇÃO DE MÉTRICAS DOS MODELOS

Modelo	FPR	Acc	Kappa	PPV	NPV
Baseline	0.810	**0.942**	0.242	0.190	**0.986**
Undersampling	**0.214**	0.859	0.325	**0.786**	0.864
Oversampling	0.310	0.876	0.328	0.690	0.887
BLSMOTE	0.619	0.936	**0.361**	0.381	0.968

REFERENCES

[1] Nuno Datia, 'Dataset reduction Instance manipulation', moodle ISEL, Abril 2024