Trabalho prático de MDLE - Fase 3

Instituto Superior de Engenharia de Lisboa
Mineração de Dados em Larga Escala
1 de junho de 2023
Grupo 08:
Gonçalo Fonseca - A50185
Pedro Diogo - A47573

I. Introdução

No contexto do "*Influenza Outbreak Twitter Data*", a terceira fase do projeto da unidade curricular Mineração de Dados em Larga Escala visa abordar o problema de mineração de dados abordados nas fases anteriores, tendo como objetivo principal aplicar corretamente técnicas de validação de modelos e avaliar, de forma crítica, os resultados obtidos. Neste documento são enunciadas algumas métricas que achámos mais pertinentes para o problema e são também apresentados os resultados e conclusões do trabalho desenvolvido.

II. Técnicas de Validação dos Modelos

Nesta secção são descritas as métricas que achámos que seriam mais pertinentes para fazer a validação dos modelos obtidos durante a fase de treino. No contexto deste problema que trata a identificação de pacientes diagnosticados com gripe (ocorrência de um surto), é importante tentar maximizar o número de verdadeiros positivos visto que é preferível diagnosticar alguém como positivo do que não. Com base nisso foram escolhidas as seguintes métricas:

- **Sensitivity** e **Specificity**: Estas métricas têm a propriedade de não dependerem da prevalência da classe - a sensibilidade (***sensitivity***) é a precisão entre os verdadeiros positivos, enquanto a especificidade (***specificity***) é a precisão entre os verdadeiros negativos. Uma vez que os verdadeiros positivos e os verdadeiros negativos são tratados de forma completamente separada por estas métricas, as suas proporções relativas não são importantes [1]. A sensibilidade foi escolhida pois quanto maior for, menor a probabilidade de o modelo falhar em identificar casos positivos, e a especificidade pois um valor alto indica que o modelo está a reduzir os falsos positivos, ou seja, a evitar classificar erradamente casos negativos como positivos. Existirá sempre um compromisso entre os verdadeiros positivos e verdadeiros negativos, porém achamos que o primeiro tem maior relevância.

- **Taxa de Falsos Negativos**: A taxa de Falsos Negativos é uma métrica de desempenho que mede a probabilidade do modelo prever um valor negativo quando o valor real é positivo [3]. Se o objetivo for minimizar os casos em que um resultado positivo é erradamente classificado como negativo, a taxa de falsos negativos é relevante, sendo tal aplicado no nosso caso. Foi escolhida esta métrica pois uma baixa taxa de falsos negativos indica que o modelo está a realizar uma deteção robusta e a evitar falhas críticas na identificação de casos positivos.

- **Taxa de Falsos Positivos**: Um Falso Positivo é quando os dados são incorretamente identificados como não sendo anómalos, ou seja, quando os dados são classificados como "normais", quando na realidade são anómalos [2]. Normalmente são estes os casos que se pretendem reduzir na área da saúde, por exemplo. Foi escolhida esta métrica pois uma baixa taxa de falsos positivos indica que o modelo está a conseguir fazer uma deteção precisa de casos positivos.

- **AUC (Area Under ROC Curve)**: Na generalidade, esta métrica mostra que quanto mais elevada for a AUC, presume-se que melhor será o desempenho do modelo na distinção entre as classes positivas e negativas [4]. No entanto, esta métrica poderá não ser a melhor devido à invariância de escala (por exemplo, por vezes precisamos de resultados de probabilidade bem calibrados e a **AUC** não nos diz nada sobre isso) e também a invariância do limiar de classificação (por exemplo, nos casos em que existem grandes disparidades na taxa dos falsos negativos em relação aos falsos positivos, pode ser fundamental minimizar um tipo de erro de classificação) [5]. Esta métrica foi escolhida pois, apesar de não indicar em concreto se um modelo é melhor que outro, permite ter uma ideia do seu desempenho geral em distinguir entre casos positivos e negativos.

Depois de escolhidas as métricas que mais se adequam para o nosso problema, desenvolveu-se o código em R baseado no código usado nas fases anteriores tendo sido feitas algumas alterações de forma a tirar partido da infraestrutura computacional que o Spark disponibiliza.

III. Análise dos Resultados

Para obtenção dos resultados, decidiu-se usar novamente o conjunto de dados da região 1 (**train_data_1.csv** e **test_data_1.csv**), visto que este foi usado nas fases anteriores, mantendo a consistência dos processos. A Tabela I contém os resultados obtidos para este *dataset* após o treino e teste de vários modelos com combinações de redução de atributos, amostragem e classificadores.

Model	Sensitivity	Specificity	FP Rate	FN Rate	AUC
SVD — Baseline — Random Forest	**1.000**	0.000	1.000	**0.000**	0.522
SVD — Baseline — SVC	0.944	0.095	0.905	0.056	0.447
SVD — Undersample — Random Forest	0.741	**0.381**	**0.619**	0.259	0.549
SVD — Undersample — SVC	0.914	0.143	0.857	0.086	0.518
SVD — Oversample — Random Forest	0.916	0.048	0.952	0.084	0.485
SVD — Oversample — SVC	0.925	0.143	0.857	0.075	0.528
SVD — BL-SMOTE — Random Forest	0.966	0.000	1.000	0.034	**0.557**
SVD — BL-SMOTE — SVC	0.948	0.095	0.905	0.052	0.540

Tabela I – Métricas dos modelos com o dataset 1

No geral, os resultados mostram que nenhum dos modelos conseguiu obter uma alta sensibilidade e especificidade em simultâneo. A deteção de casos positivos é uma prioridade, mas é importante considerar a redução de falsos positivos consideravelmente, tendo em conta o problema. Sendo mais específico, os modelos tiveram na generalidade um valor alto de sensibilidade (0.741 a 1.000), tendo no entanto um baixo valor de especificidade (0.000 a 0.381), podendo indicar imprecisões na classificação dos casos negativos, sendo que também nenhum modelo conseguiu eliminar completamente os falsos negativos.

Contudo, o que se pode observar é que o modelo que se destaca é o **SVD — Baseline — Random Forest** visto que, apesar de ter a especificidade a 0.0000 (taxa alta de falsos positivos), foi necessário dar prioridade à sensibilidade (casos positivos) com resultados de 1.000 e ainda diminuir o número de falsos negativos (com resultados de 0.000) pois é preferível prever que existe um surto e não existir, do que o contrário.

Para se testar que os modelos estão a funcionar corretamente para este problema, experimentou-se realizar os mesmos processos para os conjuntos de dados das regiões 41 e 11, tendo sido obtidos os resultados presentes nas Tabelas III e IV presentes nos Anexos, respetivamente. Em ambos os *datasets*, o modelo que obteve melhor sensibilidade foi mais uma vez o **SVD — Baseline — Random Forest**. Com estes resultados consistentes entre *datasets* podemos assumir que para este problema este modelo é o mais adequado.

A. Hiperparâmetros

Nesta subsecção pretendemos averiguar se, ao manipular os hiperparâmetros do modelo **SVD — Baseline — Random Forest** conseguimos otimizar o mesmo. Para tal foi desenvolvido código que tira partido da infraestrutura computacional do Spark para realizar validação cruzada com 10 *folds*. Os parâmetros que foram variados foram o número de árvores e a profundidade máxima das mesmas tendo sido obtidos 36 modelos diferentes.

Com base no modelo que obteve o melhor valor para o avaliador passado (AUC), foi testado o *dataset* 1 tendo sido obtidas as métricas presentes na Tabela II:

Tabela II – Modelo com hiperparâmetros

Sensitivity	Specificity	FP Rate	FN Rate	AUC
0.998	0.000	1.000	0.002	0.597

Neste caso houve uma degradação da sensibilidade e um aumento da AUC quando comparado com o modelo gerado com os parâmetros por defeito da função do classificador *random forest* do Spark.

IV. Conclusão

Nesta fase do trabalho foram aplicadas técnicas de classificação de modelos com o objetivo de resolver o problema de deteção de surtos. Após análise dos resultados obtidos, verificou-se que o modelo **SVD — Baseline — Random Forest** apresentou a melhor capacidade de deteção de casos positivos, ao mesmo tempo que minimizou a ocorrência de falsos negativos. Embora a especificidade tenha sido menor, priorizou-se a sensibilidade para identificar corretamente os casos positivos.

É importante destacar que, devido a limitações de tempo e recursos, nem todas as combinações de modelos puderam ser exploradas. No entanto, os resultados obtidos até ao momento demonstraram a eficácia do modelo selecionado para abordar o problema em questão.

Em conclusão, embora haja espaço para melhorias e a exploração de outras abordagens, foi possível alcançar um modelo que apresentou resultados próximos do desejado, aumentando a deteção de casos positivos e reduzindo significativamente os falsos negativos. Isso evidencia a relevância e a aplicabilidade das técnicas apresentadas no contexto da deteção de surtos.

REFERÊNCIAS

[1] HOAGIE, Nuclear - Answer to ≪Why we use precision/recall in binary classification but sensitivity(=recall)/specificity in medicine?≫Cross Validated, 13 abr. 2022. [Consult. 30 mai. 2023]. Disponível em https://stats.stackexchange.com/a/571457.

[2] What is False Positive Rate - [Em linha] [Consult. 30 mai. 2023]. Disponível em https://deepchecks.com/glossary/false-positive-rate/.

[3] ROELPI - Performance Metrics: False Negative RateRoel Peters, 14 dez. 2020. [Consult. 30 mai. 2023]. Disponível em https://www.roelpeters.be/glossary/false-negative-rate-machine-learning/.

[4] DEY, Victor - Understanding the AUC-ROC Curve in Machine Learning Classification [Em linha], atual. 5 set. 2021. [Consult. 30 mai. 2023]. Disponível em https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/.

[5] Classification: ROC Curve and AUC — Machine Learning - [Em linha] [Consult. 30 mai. 2023]. Disponível em https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

V. ANEXOS

Model	Sensitivity	Specificity	FP Rate	FN Rate	AUC
SVD — Baseline — Random Forest	1.000	0.000	1.000	0.000	0.874
SVD — Baseline — SVC	0.920	0.222	0.778	0.080	0.549
SVD — Undersample — Random Forest	0.595	0.689	0.311	0.405	0.688
SVD — Undersample — SVC	0.686	0.378	0.622	0.314	0.474
SVD — Oversample — Random Forest	0.998	0.000	1.000	0.002	0.626
SVD — Oversample — SVC	0.893	0.222	0.778	0.107	0.490
SVD — BL-SMOTE — Random Forest	0.998	0.111	0.889	0.002	0.855
SVD — BL-SMOTE — SVC	0.911	0.267	0.733	0.089	0.542

Tabela III – Métricas dos modelos com o dataset 41

Model	Sensitivity	Specificity	FP Rate	FN Rate	AUC
SVD — Baseline — Random Forest	1.000	0.007	0.993	0.000	0.716
SVD — Baseline — SVC	0.923	0.206	0.794	0.077	0.594
SVD — Undersample — Random Forest	0.702	0.574	0.426	0.298	0.706
SVD — Undersample — SVC	0.779	0.346	0.654	0.221	0.495
SVD — Oversample — Random Forest	0.966	0.125	0.875	0.034	0.684
SVD — Oversample — SVC	0.877	0.221	0.779	0.123	0.527
SVD — BL-SMOTE — Random Forest	0.911	0.147	0.853	0.089	0.651
SVD — BL-SMOTE — SVC	0.865	0.316	0.684	0.135	0.544

Tabela IV – Métricas dos modelos com o dataset 11