Universidade Federal de Pernambuco – UFPE

Programa de Pós Graduação em Ciência da Computação

Soluções de Mineração de Dados

IN1166

Prof: Leandro Maciel

Aluno: Irvin Soares Bezerra

Relatório Final

Análise de algoritmos de ML para a base Dry Beans

Introdução

Este projeto de mineração foi desenvolvido utilizando as melhores práticas contidas no CRISP-DM (Cross Industry Standard Process for Data Mining), que descreve as etapas necessárias para execução do projeto de mineração de dados.

Este projeto utilizou a base de dados Dry Beans, que descreve características presentes em um grão de feijão.

Para este projeto foram avaliados os seguintes algoritmos para aprendizagem de máquina: K-NN, Árvore de decisão simples, Random Forest, Rede neural MLP, um Comitê de Redes Neurais e um Comitê heterogêneo. Utilizando a linguagem Python, o framework para ciência de dados Jupyter Notebooks hospedado no Google Colab.

1 – Entendimento do problema

Nesta etapa, foi realizado o entendimento do domínio do problema. Existem diferentes tipos de grãos de feijão, com tamanho e forma específica, que indica o tipo e qualidade. A indústria alimentícia necessita de técnicas que agilizem e aumente a confiabilidade, para a identificação das qualidades dos grãos de feijão produzidos. Desta forma ficou entendido que existe uma necessidade de escolha de um modelo de algoritmo de aprendizagem de máquina, para a identificação dos grãos. Esta etapa é de grande importância, visto que durante as próximas etapas, informações extraídas a partir do entendimento do problema são cruciais para tomada de decisão. Foi criado um relatório específico para esta etapa.

2 – Compreensão dos dados.

Quais dados estão presentes na base de dados, por que estes dados e qual a finalidade. Este é o objetivo para esta etapa, compreender cada tipo de dado presente na base. Foram identificados um total de 13611 registros, distribuídos por 16 variáveis numéricas e 01 variável classe. Após analisar estatisticamente as variáveis numéricas, inicialmente não apresentaram valores discrepantes, também não apresentou valores nulos ou faltantes como também não apresentou dados corrompidos. Todas as variáveis de entrada são numéricas, algumas variáveis possuem uma escala maior que outras. A classe resposta está distribuída entre 7 categorias. As variáveis numéricas apresentam dados referentes as características dos grãos, 6 categorias apresentaram uma boa relação com a classe resposta. Foram criados gráficos para a visualização

das informações apresentadas, as informações obtidas nesta etapa foram utilizadas para a tomada de decisão acerca do melhor algoritmo de aprendizado de máquina para esta base de dados. Após entender os dados, conclui-se que será necessário utilizar um algoritmo para classificação dos dados.

3 – Preparação dos dados.

Não foram necessários aplicação de técnicas de complemento de valores ausentes ou remoção de valores da base de dados, visto que a não foram encontrados valores nulos, corrompidos ou faltantes. Foram conferidas as escalas das variáveis numéricas, algumas destas devem ser reescaladas antes de se serem aplicadas a algum modelo. Dividimos a base de dados em X para as variáveis de entrada e y para a classe resposta. Por fim, a base de dados foi separada em treino e teste sendo, 75% para treino e 25% para teste. Treino e teste foram separados em dois grupos, um com os dados reescalados e outro sem reescala.

4 – Modelagem

Foram escolhidos os seguintes algoritmos de aprendizagem de máquina: K-NN, Árvore de decisão, Random Forest, Rede neural MLP, Comitê de Redes Neurais e um Comitê heterogêneo. A base de dados reescalada foi utilizada com os algoritmos KNN e Rede neural MLP, os algoritmos Árvore de decisão e Random Forest não necessitam de reescala de dados. Os comitês utilizados, aproveitam os modelos criados pelos algoritmos de forma solo. Foram definidos valores de hiper parâmetros de acordo com as características encontradas na base de dados.

5 – Avaliação

Todas as medições de acurácia e desvio padrão foram realizadas utilizando a técnica de K-fold cross validation com o valor de 10. Todas as informações acerca da base, foram importantes determinar hiper parâmetros adotados para os algoritmos.

5.1 – KNN

Foram testados diferentes valores para K, começando com o valor de 5 até o valor ótimo para K de 15, utilizando este valor atingiu-se a acurácia de 91.85%, utilizando o cross validation com k=10. Valores maiores para K resultaram em uma acurácia menor. A partir do valor de K igual a 8 houve uma discreta melhoria na acurácia, mas atingindo um desvio padrão menor, sendo este desvio igual a 0.0109. A matriz de correlação apresenta 7 variáveis com uma forte relação com a variável resposta, sendo este o valor de vizinho com maior relevância para predizer a classe resposta.

5.2 – Decision Tree.

Para a árvore de decisão, os dados de treinamento não foram normalizados. Visto que este modelo é robusto para dados em diferentes escalas. Foram testados diferentes valores para a profundidade da árvore. Os testes se iniciaram utilizando uma ramificação igual a 1, sendo testados os seguintes valores: 5, 8, 10 e 13. A partir de 8 ramificações foi possível chegar a uma acurácia de 89.89% O valor ideal foi atingido utilizando 13 ramificações, atingindo a acurácia de 88.80% e o menor valor para o desvio padrão de 0.0081.

5.3 – Random Forest.

Para a floresta de árvores, foi adotado o valor de 13 ramificações para cada árvore da floresta, visto que foi o valor de ótimo encontrado durante a experimentação de hiper parâmetros da árvore de decisão. Para aumentar a acurácia da floresta, foi experimentado diferentes valores para as quantidades de árvores, começando com o valor de 20, e incrementado até o valor de 100. A partir de 70 árvores foi obtido uma acurácia de 91.10%, para outros incrementos, houve uma discreta melhoria para a acurácia. O valor de ótimo foi obtido com 100 árvores, apresentando uma acurácia de 91.36% e um desvio padrão de 0.0074.

5.4 – Rede Neural MLP

Para a rede neural de múltiplas camadas perceptron, foram alterados os seguintes hiperparâmetros, densidade de neurônios e a quantidade de epochs que calculam a saída e os pesos do conjunto de treinamento. Foram experimentados valores de 20 até 90 para a densidade, e de 100 até 500 para os epochs. Houve uma grande variação nos resultados de acurácia, sendo de 35.38% para densidade 20 e 100 epochs. O valor de ótimo foi encontrado com a densidade 90 e 500 epochs, apresentando uma acurácia de 70.90% e um desvio padrão de 0.0218. Como teste, foi possível obter um acurácia maior, utilizando grandes valores de densidade e de epochs, mas sendo necessário um elevado custo computacional.

5.5 – Ensemble de Redes Neurais

Após encontrar os valores de ótimo para a rede neural MLP, foram criados 5, 8 e 10 modelos para a avaliação do Ensemble. Foi treinado um meta-classificador utilizando como modelo os modelos obtidos a partir da execução solo do algoritmo MLP. Apesar de variar a quantidade de redes para encontrar o melhor valor, todas as execuções apresentaram os mesmos resultados, sendo 92.30% de acurácia e 0.0057 de desvio padrão. Dessa forma foi adotado o valor de 5 modelos como ótimo para a criação do Ensemble, visto que mais modelos elevam o custo computacional. O ensemble apresentou um desempenho superior a execução solo do algoritmo que foi de 70.90% e um desvio padrão de 0.0218.

5.6 – Ensemble heterogêneo

Após encontrar os valores de ótimo para os quatros modelo avaliados, foi possível criar um ensemble com diferentes modelos. A execução do ensemble apresentou a seguinte acurácia 92.01% e um desvio padrão de 0.0075. Obtendo um melhor desempenho que todos os modelos solo.

6 – Conclusões

A execução dos algoritmos apresento os seguintes resultados:

#	Acurácia K-fold	Desvio padrão	Tempo de execução	Algoritmo
5	92.30%	0.0057	5s	EN_MLP
6	92.01%	0.0075	24s	EN_HET
1	91.85%	0.0109	1s	KNN
3	91.36%	0.0074	16s	RF
2	88.80%	0.0081	1s	DT
4	70.90%	0.0218	1m19s	MLP

A melhor acurácia, testada através de amostra estratificada e utilizando o valor de k = 10, foi obtida pelo comitê de redes neurais perceptron, obtendo uma acurácia de 92.30%, também apresentou o menor desvio padrão de 0.0057 e um tempo de execução de apenas 5s.

O comitê de modelos heterogêneos vem em segundo colocado, utilizando a mesma estratégia de validação, apresentou uma acurácia de 92001%, um desvio padrão de 0.0075, valores ligeiramente inferiores aos obtidos no comitê MLP. Mas apresentando um tempo de execução de 16s, sendo mais demorado que o MLP.

Os algoritmos KNN e RF apresentaram boa acurácias sendo 91.85% e 91.36% respectivamente, e desvio padrão de 0.0109 para o KNN e 0.0074 para o RF. O KNN teve um tempo de execução muito rápido de apenas 1s, já o RF levou 16s para realizar todo o processamento.

Os algoritmos DT e MLP apresentaram as menores acurácias entres os algoritmos solo, 88.80% e 70.90% respectivamente. O DT obteve um tempo de execução de apenas 1s, o MLP foi o algoritmo solo mais demorado entre os avaliados, levando 1m19s para realizar todo o processamento, além de apresentar o maior desvio padrão de 0.0218. Já o DT apresentou um desvio padrão de 0.0081.

Após analisar os desempenhos dos algoritmos e comitês é possível tirar algumas conclusões, a base de dados, conforme identificado durante a etapa de compreensão requer um aprendizado voltado a classificação dos dados.

A melhor acurácia e desvio padrão foi obtido aplicando os comitês de redes neurais e heterogêneo, mas para a aplicação de ambos, será necessário aplicar outros modelos, que impactam diretamente em tempo de codificação, tempo de execução e custo computacional, o MLP é um dos modelos mais demorados, e combinados consomem ainda mais tempo.

Os algoritmos MLP e DT apresentaram resultados que implicam em rejeição para utilização, além de que são modelos voltados a lidar com outros propósitos de aprendizado de máquina, e base de dados com uma maior quantidade de dados.

Desta forma para solução do problema proposto para a classificação dos grãos de feijão, será indicado a utilização dos algoritmos KNN ou RF, ambos consomem pouco tempo de codificação, são modelos rápidos de serem executados e apresentam uma boa acurácia.