Universidade Federal do Rio Grande do Norte
Aluno: Thiago Theiry de Oliveira
Disciplina: Aprendizado de Máquina

Checkpoint 2

1 - A sua base de dados vai precisar de algum tipo de processamento? Se sim, descreva o que foi feito para limpar e organizar a base.

R = Inicialmente meus dados são de uma imagem sísmica de boa qualidade, em outras situações, algumas imagens sísmicas possuem múltiplas (um espelho de um sinal que vai se propagando no dado) ou dados faltantes (aparece um branco quando faz o plot). Mas não é o caso, o dado em questão é limpo, já vem no ponto de ser utilizado. Porém para o uso da atividade e verificar o comportamento com a redução de atributos, apliquei ao dado original e usei ele como base tendo 5 atributos sendo aplicado na imagem: ['Amplitude', 'Variância Local', 'Entropia Local', 'Gradiente', 'Média Gaussiana']. O único processamento feito nele anteriormente era apenas em dividi-lo, já que treinar uma rede com imagem é inviável. Mas agora foi feita a aplicação dos atributos no dado original que possui (2326, 23777) para (2326, 23777, 5).

2. A sua base tem outliers? Para responder esta pergunta, aplique um box-plot nos atributos numéricos e análise o box-plot para concluir se existe ou não outliers.

R = O boxplot mostra uma distribuição das médias de intensidade dos patches extraídos dos dados sísmicos. Embora o gráfico tenha indicado a presença de diversos valores identificados como outliers estatísticos (pontos acima do limite superior de 1.5×IQR), uma análise qualitativa dos patches revela que esses valores não são erros ou ruídos, mas sim refletem uma diferença natural e esperada nas características do dado.

Mais especificamente, os patches considerados "outliers" concentram-se predominantemente em regiões de água (ou seja, fora da seção sísmica principal). Estas regiões apresentam padrões visivelmente diferentes dos patches da parte sísmica, são mais homogêneas. Já os patches da região sísmica possuem variações, o que é esperado pela natureza da resposta sísmica.

Portanto, embora o boxplot aponte esses pontos como outliers do ponto de vista estatístico, eles não são outliers do ponto de vista geológico ou do domínio do

problema. São, na verdade, representações legítimas de uma classe distinta dentro dos dados (água vs. subsuperfície sísmica).

3. Aplique um método para redução de instâncias na sua base de dados. Pode ser um visto em sala de aula ou um escolhido pelo aluno. Caso seja um método escolhido pelo aluno, é importante fazer uma descrição dele. Defina esta base como: BaseReduzida1

R = Foi aplicada a amostragem estratificada para equilibrar as classes, de modo a garantir que cada classe estivesse igualmente representada. A base foi reduzida para 3.135 instâncias, mantendo a distribuição balanceada entre as diferentes classes.

4. Responda: Qual foi a redução de instâncias como resultado do método de redução?

R = Por possuir uma base de dados bem desbalanceada:

Classe 'agua': 2813 patches
Classe 'bacia sedimentar': 4231 patches
Classe 'crosta continental': 4016 patches
Classe 'crosta oceanica': 627 patches
Classe 'manto': 1669 patches

A ideia foi buscar a menor base e igualar buscando abaixar o valor das outras bases, com isso. Temos:

5 × 627 = 3135 patches na nova base (BaseReduzida1).

Cada classe estará representada igualmente.

5. Aplique um método para seleção de atributos na sua base de dados. Novamente, pode ser um visto em sala de aula ou um escolhido pelo aluno. Caso seja um método escolhido pelo aluno, é importante fazer uma descrição do mesmo. Defina esta base como: BaseReduzida2

R = A base foi reduzida usando a árvore de decisão.

6. Responda: Qual foi a redução no número de atributos como resultado do método de redução?

R = Foi usado as importâncias da árvore para somar a importância de cada atributo (canal). Assim selecionando apenas os N atributos mais importantes (valor de N ajustável). Com isso, diminuiu a quantidade de canais e a quantidade de atributos.

7. Aplique um método para extração de atributos na sua base de dados. Para esta etapa, utilize o PCA. Defina esta base como: BaseReduzida3

R = O PCA foi utilizado para reduzir a dimensionalidade dos dados. O número de componentes principais foi ajustado para 50, permitindo uma representação mais compacta dos dados, mantendo a maior parte da variação original.

8. Responda a seguinte pergunta: Qual foi a redução no número de atributos como resultado do método de redução?

R = Com o PCA foi reduzido a dimensionalidade dos dados de entrada (que originalmente eram 64×64×5 = 20.480 atributos por patch) para 50 componentes principais, então essa é a nova quantidade de atributos extraídos.

	Número de Instâncias	Número de atributos
Base Original	13.356	5 (20.480 variáveis por amostra)
Base Reduzida 1	3.135	5 (20.480 variáveis por amostra)
Base Reduzida 2	13.356	2 (8.192 variáveis por amostra)
Base Reduzida 3	13.356	50 componentes principais
Observações	A dado original resultou em 13.3256 patches de 64×64 com atributos: Amplitude, Variância Local,Entropia Local, Gradiente, Média Gaussiana. Com 5 atributos temos (64,64,5) resultando no valor apresentado na tabela.	

	Acurácia	
	Modelo 1 KNN	Modelo 2 (Árvore de Decisão)
Base Original	0.7620	0.6946
Base Reduzida 1	0.6178	0.5414
Base Reduzida 2	0.7300	0.7041
Base Reduzida 3	0.7764	0.7106

O que aconteceu com a acurácia quando diminuímos o número de instâncias?

R = Quando houve uma diminuição das instâncias, o dataset ficou com bem menos amostras, apesar de balanceada. O que resultou numa piora dos resultados, dados sísmicos precisam ter um volume extenso para conseguir generalizar bem.

O que aconteceu com a acurácia quando diminuímos o número de atributos?

R = Em contrapartida, a diminuição dos atributos não reflete em uma piora significativa, no modelo 2 de árvore de decisão até melhorou (pouca coisa). Mas os modelos se saíram bem melhor do que a redução de instâncias.

O comportamento foi o mesmo para os dois modelos?

R = Praticamente sim, ambos foram melhorando conforme a abordagem estava mudando, o resultado da basereduzida3 foram os melhores em ambos modelos, e foram melhores que a baseReduzida2 e 1. Apenas uma ressalva para o modelo KNN que os dados originais só foi superado pela baseReduzida3.

De uma forma geral, por que você acha que o comportamento da tabela acima foi o detectado?

R = Acredito que a redução de instância para bases que já possuem dados limitados de fato resultaria em uma piora, e escolhendo os melhores atributos resultaria em uma possível melhora, já que a questão de atributos excessivos podem causar uma dificuldade maior na generalização do modelo. Em geral, o KNN sofreu, assim como a Árvore de Decisão na redução de instâncias, já na redução de atributos, a Árvore de Decisão conseguiu uma melhora mais significativa, especialmente quando a redução resulta em menor complexidade. O KNN se beneficia mais da redução de dimensionalidade (PCA), pois ajuda a manter as distâncias relevantes entre pontos. A Árvore de Decisão também lida bem com a redução de atributos e dimensionalidade, uma vez que suas divisões são feitas de forma hierárquica.