Comitês de Classificadores

Luiz Felipe de Souza Silva[1]

[1]Departamento de Engenharia da Computação (DCA) – Universidade Federal do Rio Grande do Norte (UFRN)

lf06092004@gmail.com

Resumo. *Este relatório apresenta a implementação de comitês de classificadores para o classificador de raças de gatos e cachorros. O relatório expõe diferentes comitês: Bagging, Boosting, XGBoost, Random Forest e Stacking. A ideia é realizar a implementação e ver como hiperparâmetros podem aumentar a diversidade.*

1. Informações Gerais.

O objetivo da atividade foi realizar a classificação de raças de gatos e cachorros a partir de imagens, utilizando Comitês de classificadores. A ideia é testar diferentes configurações nos comitês e verificar a influência de alguns atributos no teste.

1. Qual o impacto da seleção de atributos no Bagging? Foi bom fazer a distribuição de atributos? Qual modelo foi mais afetado pela distribuição de atributos?

2. Quem forneceu a maior acurácia, Bagging ou Boosting?

3. Qual comitê de classificadores está fornecendo a melhor acurácia? Explique sua resposta.

4. Qual comitê de classificadores está fornecendo a melhor acurácia? Explique sua resposta. Qual a melhor estrutura de comitê, homogênea ou heterogênea?

Os dados de acurácia treinandos com os comitês de classificadores podem ser visto através desse direcionamento para tabela: Tabela_Checkpoint4

Notebooks, extração de características e documentação podem ser visto no repositório do github: https://github.com/luiz-pytech/Classificador_Racas_gato_Cachorro

2. Distribuição de Atributos no Bagging

As bases extraídas com CNNs foram treinadas utilizando o bagging para responder algumas perguntas do tópico 1. O impacto da seleção de atributos em alguns foram casos foram benéficas, em outros casos quase não teve impacto e em outros casos não apresentaram bons resultados. Em casos onde o modelo não é muito sensível, independente da distribuição, o modelo mantém sua acurácia estável ou com pouca variação como o caso do *k*-NN e MLP. No entanto, quando o modelo é muito sensível pela distribuição como a AD, a distribuição apresentou casos benéficos e maléficos. Para à Árvore de Decisão podemos notar que à distribuição apresenta um trabalho melhor de generalização ou perde o poder de generalização. Isso mostra que quando o modelo é mais sensível a seleção de atributo o hiperparâmetro *max_features* fornece um papel crítico no desempenho do comitê.

3. Bagging x Boosting?

Indiscutivelmente o Bagging apresentou acurácias superiores com relação ao Boosting. Isso pode ter ocorrido devido às extrações com o CNN já serem bem classificadas e o Boosting tentar minimizar os mínimos erros e se ajustando demais aos dados de treinamento, diminuindo à acurácia devido ao overfitting. Além disso, redes neurais juntamente com o Bagging demonstraram quase 100% de acurácia, se provando como uma das melhores configurações para classificar imagem e nesse projeto.

4. Random Forest

Random Forest é um técnica de aprendizado de máquina baseada em um conjunto de árvores de decisão usada para classificação e também regressão. Ele cria várias árvores independentes de acordo com um subconjunto de dados e de acordo com um subconjunto de características. Random Forest basicamente é um Bagging com outro nome. Realizando uma análise entre Bagging, Boosting e Random Forest com relação as suas acurácias, é possível notar que Random Forest e Bagging tem acurácias bem parecidas, em alguns casos Random Forest se mostrar superior, e em outros casos o Bagging demonstra superioridade.

A superioridade do Bagging e Random Forest é devido ao fato desses classificadores reduzir a variância e tirar proveito da diversidade gerada pela seleção aleatória de atributos e dados, performando com uma melhor acurácia e melhor generalização.

5. Stacking

O stacking é um modelo de comitê heterogêneo. O stacking combina vários modelos paralelos para realizar uma classificação. Isso permite que o modelo possa combinar pontos fortes de cada modelo de aprendizado e flutuar o erro entre eles. O Stacking foi treinado com 5, 10, 15 e 20 classificadores. A acurácia média total de todos os classificadores ficou acima de 97%, mas o stacking com 5 classificadores se apresentou como o melhor stacking formado. Essa acurácia pode ter sido o processo de superajuste dos dados com mais classificadores, criando características mais correlacionadas no processo de cada nível do projeto. Ainda assim, no stacking foram realizados alguns testes com comitês homogêneos e heterogêneos. Os comitês heterogêneos se mostraram mais robustos na classificação em conjunto, justamente, pela capacidade de juntar os pontos fortes de cada modelo. Os modelos utilizados foram MLP, NB, k-NN e AD.

6. XGBoost

O XGBoost é um modelo de comitê baseado em árvore de decisão que utiliza o gradiente descendente para minimizar o erro a cada interação, além disso é uma otimização do gradiente boosting em termos de velocidade e validação. As bases de dados foram treinadas com o XGBoost afora do Scikit Learn em duas configurações distintas:

- bst1 = XGBClassifier(n_estimators=100, max_depth=5, learning_rate=0.1, eval_metric='mlogloss')

- bst2 = XGBClassifier(n_estimators=300, max_depth=4, learning_rate=0.03, eval_metric='mlogloss')

As acurácias apresentadas foram bem animadoras. Com relação a outros comitês o XGBoost apresentou acurácias melhores que o Bagging usando Árvore de Decisão e o Boosting, além de acurácias bem parecidas com o Random Forest.

7. Conclusão

Comitês bem utilizados são altamente robustos e fortes para classificação. Random Forest, XGBoost, Stacking e Random Forest demonstraram grande capacidade de classificação para esse projeto.