

Lista de Exercício 3

Questão 1

Como comparar um pool de classificadores antes e após a poda? Quais alternativas temos para realizar essa tarefa? Como comparar essas alternativas?

Resposta

Podemos comparar utilizando taxas de erro (ou acerto) variando o número de classificadores do *pool*, utilizar testes estatísticos com essas taxas para confirmar hipóteses (ou rejeitar) a cerca da superioridade do desempenho dos resultados do *pool* antes e depois da poda. Assim como podemos utilizar medidas de diversidade para comparar *pool* antes e depois da poda, observando o que acontece com a diversidade quando o número de classificadores varia e também utilizar testes estatísticos com essas resultados para confirmar hipóteses (ou rejeitar) a cerca da superioridade método de poda. Cada alternativa tem sua importância, vai depender do objetivo de se realizar a poda, caso queira manter a diversidade, deve-se utilizar alguma medida de diversidade para confirmar se esse objetivo foi alcançado. Caso o objetivo seja melhorar o desempenho, deve-se utilizar taxas de acerto para confirmar se o objetivo foi alcançado. Caso tanto a melhora do desempenho, quanto a diversidade sejam objetivos, é aconselhado usar as duas alternativas citadas para validar se os objetivos foram alcançados.

Questão 2

Proponha um novo método de poda. Elabore uma arquitetura, descreva um pseudo-código e explique cada etapa. Justifique suas escolhas. Indique em qual categoria o seu algoritmo se encaixa. Não precisa implementar.

Resposta

O método proposto de poda é uma combinação entre os métodos *Kappa Prunning* e Redução de Erro. A motivação para o uso desse método é reduzir a quantidade de classificadores do *pool* sem perder a diversidade e, ao mesmo tempo, sem deteriorar o desempenho do *ensemble* criado. O método se encaixa na categoria de ordenação.

A figura 1 ilustra a arquitetura do método de poda que realiza a combinação do *Kappa Prunning* e do método Redução de Erro. Na primeira etapa, o *pool* é dado como entrada ao método *Kappa Prunning*. Então, T classificadores são

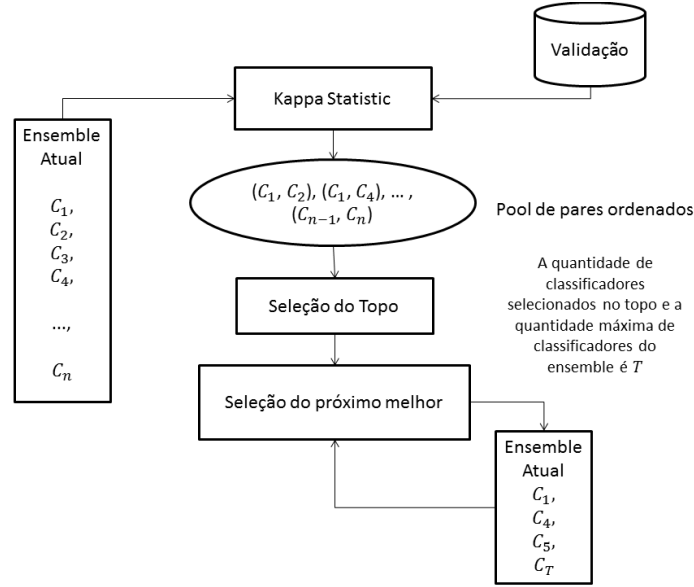


Figura 1: Arquitetura do novo método de poda. Combinação entre os métodos de poda por ordenação *Kappa Prunning* e Redução de Erro.

retirados do topo da lista de ordenação por diversidade e são avaliados pelo método de Redução de Erro. O *ensemble* criado pelo novo método tem um tamanho máximo igual a T . Esse método diminui o tempo computacional do método de Redução de Erro, pois ele não precisa ser executado diretamente no *pool* de classificadores e sim na lista de classificadores de tamanho T .

O pseudo-código do método proposto considera como entradas: um conjunto de validação C , um *pool* de classificadores P e as funções $Ensemble = KappaPrunning(P, C, T)$ e $Ensemble = ReducaoErro(P, C, T)$, onde T é o tamanho do *ensemble* para a primeira função e é o tamanho máximo do *ensemble* para a segunda.

$Ensemble = CombKappaReduErr(P, C, T)$:

1. $L_0 = KappaPrunning(P, C, T)$;
2. $Ensemble = ReducaoErro(L_0, C, T)$;
3. *retorne* $Ensemble$;

Questão 3

Implemente e avalie o método de poda EPIC. Utilize as seguintes diretrizes:

- Classificador base: Árvore de decisão;
- Banco de dados: Algum banco com mais de 2 classes do UCI Machine Learning Repository;

- Tamanho do pool de classificadores: 10, 20, 30, 40, ..., 100;
- Tamanho do ensemble: 10% do tamanho do pool;
- Estratégia de combinação: votação.

Resposta

O método de poda EPIC foi implementado em Matlab e sua avaliação foi feita utilizando o método de validação cruzada KFold com $K = 10$, considerando 8 partições de treino, 1 partição de validação para a poda e 1 partição para teste. A base de dados escolhida foi do UCI Machine Learning Repository a Ecoli .

A figura 2 ilustra os resultados dos testes utilizando a menor taxa de erro obtida do KFold para o *pool* e o *ensemble*. A Tabela 1 mostra a média e desvio padrão das taxas de erro obtidas do KFold para o *pool* e o *ensemble*.

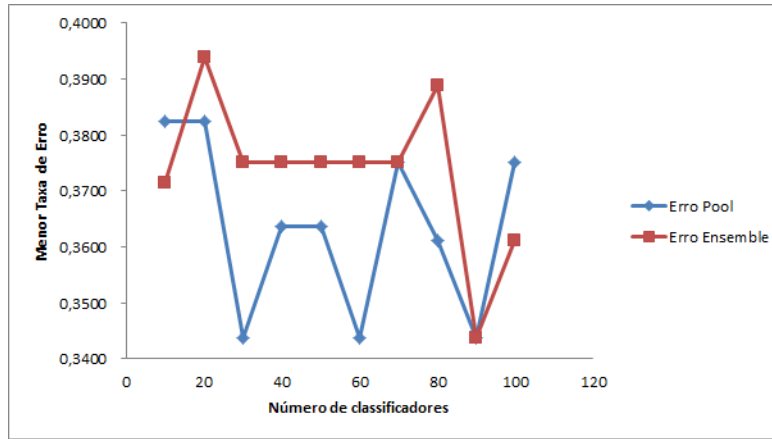


Figura 2: Resultados do *ensemble* e do *pool* utilizando a menor taxa de erro do KFold com $K = 10$ e variando o número de classificadores do *pool*.

Observando a figura 2, vemos que o menor erro do *pool* é menor que o do *ensemble* na maioria dos casos. Os casos em que o *ensemble* foi melhor que o *pool* foram aqueles onde o *pool* possuía 10 e 100 classificadores. Quando o número de classificadores aumenta até 80 classificadores, a menor taxa de erro se deteriora em relação à quantidade de 10 classificadores, Permanecendo constante na maioria destes casos. Nos casos de 90 e 100 classificadores, a menor taxa melhora em relação ao caso de 10 classificadores, alcançando o melhor resultado e igualando-se ao *pool*. Já ao se observar os valores na Tabela 1, vemos que na maior parte dos casos a média da taxa de erro do KFold e seu desvio padrão para o *ensemble* são menores que a média e desvio do *pool*, indicando, neste estudo, uma superioridade de desempenho quando utilizado o método EPIC para poda do *pool*.

Tabela 1: Resultados da média e desvio padrão das taxas de erro obtidas pelo EPIC e pelo *pool* após o Kfold com $K = 10$.

Classificadores	EPIC		Pool	
	Média	Desvio Padrão	Média	Desvio Padrão
10	0,4502	0,0420	0,4550	0,0397
20	0,4800	0,0515	0,4557	0,0682
30	0,4391	0,0433	0,4374	0,0784
40	0,4491	0,0675	0,4552	0,0699
50	0,4293	0,0811	0,4402	0,0717
60	0,4434	0,0473	0,4493	0,0726
70	0,4493	0,0470	0,4434	0,0654
80	0,4409	0,0397	0,4498	0,0668
90	0,4140	0,0511	0,4432	0,0725
100	0,4323	0,0595	0,4461	0,0679