8
Conclusões e Trabalhos Futuros

O uso de métodos de comitê como Boosting melhora a acurácia de diversos algoritmos de aprendizado de máquina. Tais métodos criam uma série de classificadores "fracos" que possuem desempenho superior para diferentes tipos de instâncias. Um método simples de votação, baseado na acurácia de cada um dos classificadores, é utilizado para combinar as suas predições.

Desde a publicação do algoritmo AdaBoost, um problema em aberto é a sua extensão para permitir o uso de distribuições iniciais arbitrárias, apesar de diversos autores terem comentado sobre a sua possibilidade.

Boosting At Start é uma nova abordagem de aprendizado de máquina baseada em Boosting que generaliza o algoritmo *AdaBoost* e permite a utilização de qualquer distribuição inicial para os exemplos de treinamento. O BAS propõe uma correta extensão que introduz uma fórmula original para calcular o parâmetro α_t de atualização do algoritmo Boosting. Neste trabalho, demonstramos que tal resultado resolve o problema em aberto e propomos uma heurística para a determinação da melhor distribuição inicial. Também apresentamos maneiras de construir classificadores eficientes por meio de comitês de classificadores BAS.

Além disso, apresentamos um novo esquema de votação, chamado Votação ETL, que melhora a qualidade das predições do algoritmo Comitê BAS. Como o Votação ETL não realiza nenhuma suposição quanto à origem dos votos, ele pode ser utilizado também em qualquer abordagem de comitê, substituindo uma esquema de voto pela maioria.

Realizamos diversos experimentos para evidenciar o bom desempenho das abordagens derivadas do BAS em problemas de classificação de dados textuais e não-textuais. Em todos as tarefas, o BAS mostrou melhores resultados que o AdaBoost. Acreditamos que tal fato decorre da introdução de uma nova forma de conhecimento dos dados representada por meio da distribuição inicial dos exemplos.

Tais evidências indicam que a utilização de tal conhecimento extra na forma de uma distribuição inicial para os exemplos resulta em um melhor aprendizado.

Além disso, observamos os seguintes aspectos positivos a respeito das abordagens BAS apresentadas:

1. Permitem a inclusão de *conhecimento extra* na forma de uma distribuição inicial para os exemplos do conjunto de treinamento. Tal distribuição pode ser uma entrada fornecida pelo usuário ou ser determinada por uma das heurísticas apresentadas

2. A *simplicidade da modelagem* BAS para novas tarefas. Qualquer modelo utilizando o AdaBoost pode ser evoluído em uma estratégia BAS apenas escolhendo uma das formas apresentadas de determinação da distribuição inicial.

3. Permitem a inserção simples de amostras no treinamento, melhorando o agrupamento dos exemplos do algoritmo Comitê BAS, transformando-o em um *esquema semi-supervisionado*.

4. O algoritmo BAS é *simples* e tem a mesma *eficiência* que o algoritmo AdaBoost.

5. O esquema de combinação de classificadores Votação ETL é melhor do que uma simples votação pela maioria pois permite uma *combinação eficaz de votos*. Tal estratégia é genérica e pode ser utilizada em qualquer algoritmo de comitê.

6. O treinamento do algoritmo BAS Comitê pode ser facilmente *paralelizável*, permitindo a utilização de vários processadores para gerar um classificador em um tempo aproximadamente igual ao treinamento do mesmo modelo utilizando o AdaBoost.

7. Todos os classificadores gerados são *independentes* e também podem ser *paralelizáveis*, permitindo um tempo de classificação razoável.

8. O algoritmo BAS pode ser facilmente utilizado como *algoritmo-base* em uma estratégia de comitê. Tal fato é demonstrado por meio do algoritmo Comitê BAS.

Apresentamos diversas heurísticas para determinação da distribuição inicial. Entretanto, a descoberta da melhor distribuição inicial continua sendo um problema aberto.

Mesmo apresentando bons resultados experimentais, as abordagens derivadas do BAS podem ser ainda aprimoradas por meio de algumas extensões.

Outras heurísticas de determinação da distribuição inicial podem ser aplicadas, bem como o processo de combinação dos classificadores gerados pode ser melhorado.

Os dados não rotulados podem ser incorporados ao treinamento do algoritmo BAS. Apresentamos alguns experimentos de aprendizado semi-supervisionado. Entretanto, eles se limitam a utilização das amostras na fase de agrupamento de exemplos do algoritmo Comitê BAS e, caso suportado, no treinamento do algoritmo-base. Uma extensão do BAS de forma a incorporar tais dados não-rotulados, de forma similar ao algoritmo Assemble (Ben02), deve melhorar ainda mais o treinamento de classificadores, sem aumentar o custo para rotular dados.

A utilização do Algoritmo AdaBoost.M1, para as tarefas de classificação multi-classe, pode ter diminuído a melhora relativa das abordagens BAS em relação ao algoritmo-base. Outras extensões multi-classe do AdaBoost, que possuem uma comunicação mais forte com o algoritmo-base, podem ser empregadas no treinamento de tais tarefas de forma a melhorar a qualidade dos classificadores gerados.

A eficiência do algoritmo Comitê BAS pode ser aprimorada com a paralelização da fase de agrupamento de exemplos com um custo bem baixo de desenvolvimento, visto que tais algoritmos são facilmente paralelizáveis.

Mesmo utilizando configurações de parâmetros iguais para todas as tarefas, conseguimos obter, por exemplo, o melhor classificador de sintagmas nominais do Português, utilizando o Corpus SNR-CLIC, cujo desempenho se encontra na faixa de 90%. Para todas as tarefas, entretanto, caso tais parâmetros sejam customizados, os desempenhos podem ser maiores ainda.