1
Introdução

1.1
Tema

Aprendizado de Máquina (Mit97) é o campo da Inteligência Artificial responsável pelo desenvolvimento de modelos inferidos automaticamente a partir de dados. Existem diversas aplicações para aprendizado de máquina, tais como Processamento de Linguagem Natural, Mecanismos de Busca para Internet, Detecção de Fraudes, Reconhecimento de Padrões, entre outras. O objetivo principal do aprendizado de máquina é a geração de sistemas computacionais que possam substituir o trabalho humano a um baixo custo. Também é desejável que tais sistemas possam aprender com a experiência, melhorando continuamente o seu desempenho.

Métodos de Comitê (Pol06) é o campo do aprendizado de máquina que constrói um grupo de classificadores, denominados classificadores-base, com o objetivo de ser mais preciso que o melhor dos elementos do grupo. Tais classificadores-base são treinados iterativamente utilizando o mesmo algoritmo-base ou algoritmos-base diferentes. A abordagem mais simples baseada neste algoritmo é o Voto da Maioria Simples, onde diversos classificadores são combinados em uma estratégia de voto. Como resultado final, a resposta que receber o maior número de votos é considerada a resposta do comitê.

Bagging (Bre96), ou *Bootstrap Aggregating*, é um método de comitê que utiliza re-amostragem aleatória uniforme dos exemplos com reposição na construção dos classificadores do comitê. Dado um conjunto de exemplos de treinamento E de tamanho n, o *Bagging* gera b novos conjuntos de treinamento de tamanho $n' \leq n$. Os b modelos são então combinados de forma a definir uma resposta final, que pode ser por meio de uma média, em caso de problemas de regressão, ou votos, em caso de problemas de classificação. Apesar de bem simples, o *Bagging* possui um desempenho muito bom quando combinado com árvores de decisão, diminuindo a variância e a possibilidade de sobre-ajuste (*over-fitting*).

Boosting (Sch90, Fre90) é um meta algoritmo de aprendizado de máquina

que combina um conjunto de classificadores-base "fracos", ou com um baixo desempenho, criando um classificador "forte", ou com um alto desempenho. Os primeiros algoritmos que se aproveitaram de tal paradigma foram o *Aprendizado Boosting* (Sch90) e o *Boosting pela Maioria* (Fre90), entretanto não obtiveram muito sucesso por não serem adaptáveis aos classificadores-base utilizados (Dru93).

AdaBoost (Fre95a), ou *Adaptative Boosting*, é o algoritmo baseado em Boosting mais famoso e utilizado. Seu sucesso deve-se exatamente a sua capacidade de se adaptar aos classificadores-base. No algoritmo AdaBoost, os classificadores são gerados subsequentemente de forma a favorecer exemplos erroneamente classificados pelos classificadores anteriores.

Mineração de Dados (Fay96) é utilizada para determinar padrões latentes de dados. Com o advento da Internet, cada vez mais são geradas grandes quantidades de dados, o que inviabiliza a extração manual de informação. Usualmente, a mineração de dados é aplicada através de dois processos. O primeiro deles é o descobrimento do conhecimento, que permite a extração de características sobre os dados coletados. O segundo é a predição, que permite a construção de um modelo capaz de prever eventos futuros. A mineração de dados consiste em quatro tarefas básicas:

- *Regressão*, que encontra uma função de ajuste para os dados com o menor erro possível;
- *Classificação*, que organiza os dados em classes pré-definidas;
- *Agrupamento*, que organiza os dados em classes que não estão pré-definidas; e
- *Associação de regras*, que encontra relações entre dados.

Mineração de texto (Coh08), inspirada na mineração de dados, é o processo de obtenção de informações provenientes de texto com uma alta qualidade. A mineração de texto envolve, em sua essência, os processos de estruturação do texto, análise sintática e inserção de atributos linguísticos. Algumas tarefas de mineração de texto que podem ser citadas:

- *Categorização*, que classifica documentos em uma ou mais categorias escolhidas;
- *Agrupamento*, que forma grupos de documentos similares;
- *Sumarização*, que cria uma versão menor e sucinta do documento;
- *Extração de Entidades*, que identifica pequenos artefatos do texto que possuem uma característica singular;

– *Análise de Sentimento*, que determina, para cada parágrafo de um texto, o posicionamento do autor em relação a um determinado tema. Normalmente, visa determinar se o autor está expressando uma opinião ou não.

1.2 Objetivos

Nesta tese, apresentamos o algoritmo *Boosting At Start* (BAS), uma nova abordagem de aprendizado de máquina baseada em Boosting. BAS é uma generalização do algoritmo AdaBoost que permite a utilização de uma distribuição inicial arbitrária para os exemplos. Dessa maneira, o BAS permite a inclusão de um conhecimento extra sobre o problema por meio da distribuição inicial dos exemplos.

A despeito do fato de considerarmos o problema de encontrar a melhor distribuição inicial para os exemplos em aberto, neste trabalho, apresentamos uma heurística baseada em computação evolucionária que utiliza algoritmos genéticos de forma a obter uma boa distribuição inicial.

Também neste trabalho é apresentado um algoritmo chamado Comitê BAS, que explora o poder da abordagem BAS de utilizar diferentes distribuições iniciais. Neste caso, é construído um comitê de classificadores BAS. Tal classificador resultante possui um desempenho melhor do que o melhor classificador do comitê.

1.3 Metodologia

De forma a comprovar a eficiência das abordagens propostas, uma série de experimentos é conduzida em problemas de classificação de dados e de texto. Tais experimentos podem ser divididos em dois tipos. O primeiro é o de problemas de classificação de dados extraídos do repositório de dados UCI (Uci98). O segundo é formado por problemas de Processamento de Linguagem Natural (PLN) com tarefas de Mineração de Texto para diversos idiomas. Nosso objetivo é comprovar o desempenho das abordagens propostas. Para tal, elas são comparadas com o algoritmo AdaBoost original, que utiliza uma distribuição inicial uniforme para os exemplos, e com outros algoritmos de estado-da-arte especificamente desenvolvidos para tais tarefas.

É interessante ressaltar aqui, os experimentos de PLN realizados. Na Tabela 1.1, mostramos um extrato da informação dos corpora utilizados.

Tabela 1.1: Características dos corpora de texto utilizados.

Tarefa	Corpus	Idioma	Sentenças	Palavras
Anotação Morfossintática	Mac-Morpho	Português	53.374	1,221.465
	Tycho Brahe	Português	40.932	1,035.592
	Brown	Inglês	57.340	1,161.192
	Tiger	Alemão	50.474	888.578
Anotação de Sintagmas	SNR-CLIC	Português	4.392	104.144
	Rams. & Marc.	Inglês	10.948	259.104
	CoNLL-2000	Inglês	10.948	259.104
	SPSAL-2007	Hindi	1.134	25.000

Para todas as tarefas, uma modelagem genérica é aplicada utilizando algoritmos-base, simples ou encontrados na literatura correspondente a cada problema examinado.

Na Tabela 1.2, sumarizamos os resultados para cada conjunto de dados utilizado. Os melhores resultados podem ser observados em negrito para cada problema. Podemos perceber que os resultados são comparáveis com o estado da arte para as tarefas examinadas.

Tabela 1.2: Desempenho comparativo do algoritmo BAS nas tarefas de PLN.

Tarefa	**Corpus**	**Estado da Arte**		**BAS**	
				BAS100	BAS400
Anotação Morfossintática	Mac-Morpho	Comitê-ETL	96.94	96.96	**96.98**
	Tycho Brahe	Comitê-ETL	**96.72**	96.69	96.66
	Brown	Comitê-ETL	96.83	96.88	**96.89**
	Tiger	Comitê-ETL	96.68	96.69	**96.76**
Anotação de Sintagmas	SNR-CLIC	Comitê-ETL	89.58	89.72	**89.93**
	Rams. & Marc.	SVM	**94.22**	93.31	93.37
	CoNLL-2000	SVM	**94.12**	93.28	93.33
	SPSAL-2007	HMM+CRF	**80.97**	80.54	80.58

Os conceitos principais do BAS e de seus algoritmos derivados são apresentados em uma série de artigos publicados por este autor. Em Milidiú & Duarte (Mil09a), a abordagem Boosting BAS é apresentada, bem como o algoritmo Comitê BAS e sua aplicação em diversos problemas de classificação de dados. A versão semi-supervisionada do algoritmo Comitê BAS é destaque em Milidiú & Duarte (Mil09b), com aplicação em problemas de classificação binária de dados comumente utilizados em experimentos envolvendo estratégias de Boosting e Aprendizado Semi-Supervisionado. Em Milidiú & Duarte (Mil09c), o algoritmo Comitê BAS é estendido por meio da utilização

de um novo esquema de votação final denominado Votação ETL. Finalmente, em Duarte & Milidiú (Dua07), é apresentada uma versão estendida do algoritmo BAS que aceita, além de uma distribuição inicial, uma função de custo de erro para os exemplos.

Além de tais trabalhos diretamente ligados ao algoritmo BAS, alguns outros experimentos no campo do aprendizado de máquina foram reportados em outros trabalhos deste autor. Em Milidiú et al. (Mil08b), apresentamos modelos derivados a partir do algoritmo *Aprendizado de Transformação Guiado pela Entropia* (ETL) para três tarefas de linguagem natural do Português: anotação morfossintática, anotação de sintagmas e reconhecimento de entidades nomeadas. Em Milidiú et al. (Mil08a), propomos um esquema independente do idioma para a tarefa de anotação de sintagmas. Em Milidiú et al. (Mil07a, Mil07b), introduzimos o problema de construção automática de gabaritos para o algoritmo Aprendizado Baseado em Transformações (*Transformation-Based Learning* - TBL), utilizando uma abordagem evolucionária baseada em algoritmos genéticos. Em Milidiú et al. (Mil06b, Mil07c), apresentamos experimentos para a tarefa de reconhecimento de entidades nomeadas do Português utilizando três algoritmos: Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM), Cadeias de Markov Escondidas (*Hidden Markov Models* - HMM) e TBL. Em Freitas et al. (Fre06), mostramos a aplicação dos algoritmos TBL e HMM na tarefa de identificação de apostos do Português. Por último, em Milidiú et al. (Mil06), realizamos experimentos de aprendizado semi-supervisionado utilizando uma combinação de modelos gerados a partir de HMM e TBL.

1.4 Estrutura da Tese

Essa tese está organizada da seguinte forma. No capítulo 2, descrevemos o algoritmo BAS bem como a sua prova de corretude. No capítulo 3, mostramos uma heurística para determinação da melhor distribuição inicial por meio de algoritmos genéticos. No capítulo 4, o algoritmo BAS Comitê é apresentado bem como suas variantes no esquema de votação final e na forma de aprendizado. No capítulo 5, apresentamos uma descrição dos algoritmos-base utilizados pelos algoritmos de comitê relatados nos experimentos. No capítulo 6, apresentamos os resultados das abordagens propostas para problemas de classificação de dados. No capítulo 7, reportamos os resultados com experimentos envolvendo problemas de classificação textual. Finalmente, no capítulo 8, apresentamos nossas considerações finais e sugestões de trabalhos futuros.