

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ  
CURSO DE TECNOLOGIA EM SISTEMAS PARA INTERNET  
CÂMPUS GUARAPUAVA

Fabio Kenji Oshiro Takatuzi

# **Algoritmo Incremental para Aprendizagem de Árvores de Decisão Adaptativas**

PROPOSTA DE TRABALHO DE CONCLUSÃO DE CURSO DO CURSO SUPERIOR EM  
TECNOLOGIA EM SISTEMAS PARA INTERNET

GUARAPUAVA  
1º Semestre de 2016

Fabio Kenji Oshiro Takatuzi

# **Algoritmo Incremental para Aprendizagem de Árvores de Decisão Adaptativas**

Proposta de Trabalho de Conclusão de Curso de graduação, apresentado à disciplina de Trabalho de Conclusão de Curso 1, do Curso Superior de Tecnologia em Sistemas para Internet – TSI – da Universidade Tecnológica Federal do Paraná – UTFPR – Câmpus Guarapuava, como requisito parcial para obtenção do título de Tecnólogo em Sistemas para Internet.

Orientador: Profa. Ms. Renata Luiza Stange Carneiro Gomes

GUARAPUAVA  
1º Semestre de 2016

## 1. SUMÁRIO DA PROPOSTA DE TRABALHO

### 1.1. Título

Algoritmo Incremental para Aprendizagem de Árvores de Decisão Adaptativas.

### 1.2. Modalidade do Trabalho

Pesquisa

Desenvolvimento de sistemas

### 1.3. Área do Trabalho

Inteligência Artificial, Aprendizagem de Máquina, Árvores de Decisão e Tecnologia Adaptativa.

### 1.4. Resumo

A estrutura das árvores de decisão é hoje um dos métodos mais utilizados para a resolução de problemas de classificação, tomada de decisão e aprendizado de máquina. Através dela, é possível que um conjunto de dados sejam classificados com base em seus atributos, gerando assim um dado de saída que age como resposta aos valores de entrada. Entretanto, existem muitas situações em que os dados de entrada se encontram em constante mudança, como nos problemas de aprendizagem incremental, em que o conjunto de treinamento trabalha de maneira dinâmica. Para problemas como esse, a adaptatividade apresenta uma solução bastante consistente, pois permite que uma estrutura se automodifique em resposta a estímulos externos, incorporando a ela novas informações. Nesse contexto, o trabalho tem como objetivo propor um algoritmo para indução de árvores de decisão utilizando tecnologias adaptativas. O algoritmo deve ser capaz de alterar dinamicamente sua estrutura hierárquica, em resposta a estímulos do ambiente, durante o processo de classificação. Dessa forma espera-se que o algoritmo apresente uma alternativa para os métodos de aprendizagem de máquina tradicionais e também acarrete soluções expressivas para problemas de aprendizado incremental.

## 2. PROPOSTA DE TRABALHO

### 2.1. Introdução

A resolução de um problema computacional bem especificado pode ser dada através de um algoritmo. Um algoritmo, por sua vez, é uma sequência de instruções computacionais que produzem, a partir de um valor de entrada, ou um conjunto deles, um valor ou conjunto de valores como saída (CORMEN et al., 2009).

No entanto, determinadas tarefas ou problemas existentes não são passíveis de serem realizadas ou resolvidos utilizando algoritmos tradicionais, como por exemplo o correto reconhecimento de caracteres manuscritos, ou até mesmo qual a melhor sequência de jogadas para se ganhar uma partida de xadrez. Esse tipo de situação necessita que o programa “aprenda” quais são as melhores soluções para os problemas ou qual a melhor maneira de se executar uma tarefa específica. Nesse contexto, o aprendizado de máquina é uma área da Inteligência Artificial que estuda a capacidade dos sistemas computacionais de aprender e modificar seu comportamento através da experiência adquirida a fim de melhorar seu desempenho em execuções posteriores (ALPAYDIN, 2010).

No campo da aprendizagem de máquina existem diversos métodos de aprendizado, tais como aprendizado Bayesiano, redes neurais artificiais, modelo oculto de Markov e indução de árvores de decisão. Nesse contexto, a estrutura das árvores de decisão é um dos métodos mais práticos e utilizados para a **inferência indutiva** (MITCHELL, 1997). Através dessa estrutura, a representação de funções discretas sobre múltiplas variáveis possuindo características hierárquicas, facilitam a inspeção, compreensão e a sua utilização por seres humanos (PISTORI, 2003).

Os algoritmos de aprendizagem de árvores de decisão utilizam a estratégia dividir para conquistar, de maneira que um problema mais complexo é decomposto em subproblemas mais simples e, dessa maneira, recursivamente, a mesma estratégia é aplicada para cada subproblema estruturando assim a árvore. Os atributos de maior importância vão sendo colocados na árvore de decisão desde a raiz até os nós folha, de acordo com alguma medida estatística (ex. ganho de informação). Muitos algoritmos que têm sido desenvolvidos para a indução de árvores de decisão baseiam-se neste

paradigma *top-down*, que constroem a árvore escolhendo o atributo raiz das subárvores. Alguns dos exemplos mais conhecidos são o ID3 (Quinlan 1986) e o seu sucessor C4.5 (Quinlan 1993) (MITCHELL, 1997).

De forma geral, a maioria dos algoritmos de aprendizagem de máquina não são incrementais, ou seja, não possuem a capacidade de incluir novos dados e informações no conjunto de treinamento ao longo de sua execução. Um exemplo de algoritmo de indução árvores incremental é o CART incremental (CRAWFORD, 1989), o ID5R (UTGOFF, 14 1989) e o ITI (UTGOFF; BERKMAN; CLOUSE, 1997).

Grande parte dos algoritmos pesquisados atualmente buscam métodos mais eficientes para o ganho de informação e para a determinação do momento em que uma árvore deixe de se expandir (PISTORI; NETO, 2002), uma vez que em alguns casos de operações de paridade ou multiplexação, a estrutura da árvore cresceria rapidamente, no entanto, sem melhorar sua capacidade de predição e classificação (CATAE; ROCHA, 2011). Entretanto, a busca por algoritmos baseados em aprendizado incremental é igualmente importante pois existem diversas aplicações em que novas informações vão surgindo ou se alterando ao longo do tempo, formando assim um cenário em que a utilização de algoritmos incrementais implicaria em resultados bastante expressivos (PISTORI; NETO, 2002).

Considere um sistema de monitoramento climático, o qual realiza leituras ambientais (velocidade do vento, temperatura média do ar, etc.) a todo momento. Suponha então, que o mesmo, por meio das leituras realizadas, precise identificar e aprender quais conjuntos de atributos e valores seriam mais apropriados para que um avião realizasse uma viagem de um lugar para outro. Devido ao ambiente no qual o software trabalha, os dados coletados estão constantemente se alterando, implicando assim em uma frequente mudança no conhecimento a ser aprendido e na repetição do processo de aprendizagem. Dessa forma, a utilização de métodos não incrementais se torna inapropriada, uma vez que o conjunto instâncias de treinamento do algoritmo não são fixas (UTGOFF, 1989) e a repetição no processo de aprendizagem implica em um custo computacional elevado (YOSHIDA, 2007). Neste caso, surge a necessidade de implementar processos dinâmicos baseados na aprendizagem incremental, de maneira que a base de conhecimento possa ser atualizada sem que o processo de aprendizagem necessite ser repetido.

Um conceito novo que tem sido adotado na solução de problemas com características dinâmicas é a adaptatividade. De acordo com Neto (2011) o conceito de adaptatividade é definido pela capacidade que os sistemas, nele baseados, possuem de promover espontaneamente alterações em seu próprio comportamento, de acordo com a necessidade, em função de sua situação e comportamento correntes, e dos valores de suas entradas. Para um problema de aprendizagem incremental, a utilização de técnicas adaptativas pode representar uma solução bastante expressiva (NETO, 2000; STANGE, 2011).

A Tecnologia Adaptativa consiste na utilização da adaptatividade de maneira prática e econômica, correspondendo também ao conjunto de métodos, técnicas e ferramentas que visam a resolução de problemas concretos utilizando modelos baseados em dispositivos adaptativos (STANGE, 2011; NETO, 2011). Um dispositivo adaptativo por sua vez é composto por um dispositivo subjacente não adaptativo e uma camada adaptativa, que confere a este dispositivo a capacidade de se automodificar (NETO, 2011).

Um exemplo de dispositivo adaptativo são as árvores de decisão adaptativas proposta por (PISTORI, 2003; PISTORI, H.; NETO, 2003), que permite que a estrutura hierárquica de uma árvore de decisão comum possa ser dinamicamente alterada durante o processo de decisão, quando a árvore é percorrida da raiz para as folhas.

Neste trabalho, será proposto um algoritmo incremental de indução de árvores de decisão baseado em técnicas adaptativas, que visa facilitar a representação e especificação de dispositivos que alteram continuamente sua estrutura e conjunto de treinamento, tornando-se assim, uma possível alternativa para a construção de árvores de decisão adaptativas existentes (PISTORI; NETO, 2002).

## **2.2. Objetivos**

### **2.2.1. Objetivo Geral**

O objetivo geral deste trabalho é avaliar a utilização da tecnologia adaptativa no processo de aprendizado de máquina aplicado às árvores de decisão.

### 2.2.2. Objetivos Específicos

- Obter um algoritmo de indução de árvores de decisão adaptativa com desempenho comparável aos algoritmos de árvore de decisão.
- Implementar e testar o algoritmo de criação de árvores de decisão adaptativas.
- Comparar o algoritmo proposto com os demais algoritmos de indução de árvores de decisão incrementais e não incrementais.

### 2.3. Estado da arte

Pistori e Neto (2002) propõe um algoritmo de indução de árvores de decisão chamado *Adaptree* utilizando conceitos da teoria dos autômatos e da tecnologia adaptativa. O desempenho obtido apresenta resultados comparáveis aos principais algoritmos para aprendizagem de máquina tornando-se o primeiro algoritmo de aprendizagem baseado em árvores de decisão adaptativas. O algoritmo baseia-se na utilização e combinação de técnicas sintáticas e estatísticas, além de ser proposta de um algoritmo incremental, ou seja, durante o processo de classificação de dados, novos exemplos podem ser incorporados à ele. A utilização de tecnologias adaptativas no trabalho, facilita a representação de dispositivos que sofrem constante alteração em resposta a estímulos externos. Os testes do *Adaptree* foram realizados sobre os conjuntos de dados da ferramenta WEKA (Waikato Environment for Knowledge Analysis) e comparados à outros algoritmos tradicionais não incrementais (ID3, Naive Bayes, C4.5, 5NN e Redes Neurais com *Backpropagation*).

Catae e Rocha (2011) através do uso de dispositivos adaptativos aplicados à árvores de decisão apresentam a possibilidade da modelagem adequada de operações de paridade, ou exclusivo (XOR) e multiplexação, abordagens que não apresentavam bons resultados em estruturas de árvores de decisão tradicionais. O algoritmo propõe que a árvore seja capaz de associar novos atributos com base em atributos já existentes e a utilização de dispositivos adaptativos possibilita a utilização de um nível independente da estrutura do núcleo possibilitando que a árvore seja construída, no entanto, mantendo-se

a estrutura do mesmo. A modelagem do algoritmo baseia-se, principalmente, no Princípio da Navalha de Occam e em conceitos da Complexidade Algorítmica (complexidade de Kolmogorov), que optam pela simplicidade como conceito de maior importância.

A tecnologia adaptativa e o aprendizado incremental são conceitos muito importantes quando a tarefa de aprendizagem precisa lidar com conjuntos de dados dinâmicos. O trabalho de Pistori e Neto, além de ser uma proposta de aprendizagem incremental, é o primeiro algoritmo a utilizar a estrutura das árvores de decisão e técnicas adaptativas para seu funcionamento, dando uma base fundamental para a elaboração do trabalho proposto. Como os testes do *Adaptree* foram realizados apenas em comparação com algoritmos não incrementais, existe a possibilidade de testes futuros serem realizadas apenas em comparação a ele, sem a necessidade de serem feitos com os algoritmos tradicionais. Analogamente, o trabalho de Catae e Rocha também utilizou técnicas adaptativas para a construção de árvores de decisão, no entanto, partindo do pressuposto de que a simplicidade é um dos conceitos mais importantes. Nesse contexto, o princípio e conceito da Navalha de Occam e Complexidade de Kolmogorov também podem contribuir para o trabalho proposto, uma vez que o mesmo também necessita de métricas em comum para a elaboração e determinação de melhores modelos. Além disso, seguindo a ideia da simplicidade o algoritmo proposto pode apresentar resultados mais expressivos.

## **2.4. Diferencial Tecnológico**

O algoritmo proposto será exposto a testes de comparação tanto com algoritmos não incrementais quanto com algoritmos de aprendizagem incremental.

## **2.5. Metodologia**

1. Estudar os conceitos gerais de aprendizagem de máquina e adaptatividade.
  - Estudar o comportamento de dispositivos adaptativos aplicados à árvores de decisão.



- Estudar o funcionamento dos algoritmos de aprendizagem de máquina clássicos (ex. ID3, C4.5, ID5R, Adaptree).

2. Propor um algoritmo para árvores de decisão adaptativas baseado em aprendizagem incremental.

- A proposta do algoritmo será baseada no conceito de adaptatividade que permitirá que o conjunto de regras da árvores de decisão seja dinamicamente modificado.

3. Implementar e testar o algoritmo de criação de árvores de decisão adaptativas.

- O algoritmo implementado será integrado com a ferramenta WEKA<sup>1</sup> (*Waikato Environment for Knowledge Analysis*), que é um ambiente para execução de algoritmos de aprendizagem de máquina.

- Os testes do algoritmo serão realizados utilizando medidas de precisão e taxas de erro que, serão calculados pela própria ferramenta.

4. Comparar o algoritmo proposto com os demais algoritmos de indução de árvores decisão incrementais e não incrementais.

- Serão aplicados diferentes algoritmos de aprendizagem de máquina em conjuntos de dados disponíveis no UCI<sup>2</sup>, que é um repositório de *data sets* para testes de aprendizagem de máquina.

<sup>1</sup> Disponível em <http://www.cs.waikato.ac.nz/ml/weka/>

<sup>2</sup> <http://archive.ics.uci.edu/ml/>

## 2.6. Planejamento do Trabalho

Atividades	TCC 1					TCC 2				
	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1. Revisão bibliográfica										
2. Escrita da proposta										
3. Escrita do projeto										
4. Elaboração do algoritmo										
5. Implementação do algoritmo e integração com a ferramenta										
6. Realização dos testes e comparação dos resultados										
7. Escrita da Monografia de TCC										
8. Elaboração da apresentação final e defesa.										

## 2.7. Recursos Necessários

Os recursos que serão utilizados no trabalho são: um computador, um Ambiente de Desenvolvimento Integrado(IDE – *Integrated Development Environment, Intellij IDEA, Eclipse*), ferramenta WEKA(*Waikato Environment for Knowledge Analysis*) e o repositório UCI. Todos os recursos estão disponíveis e não precisarão ser fornecidos.

## 2.8. Horário de Trabalho

Horário	Seg	Ter	Qua	Qui	Sex	Sab
7h30 - 8h20						
8h20 - 9h10						
9h10 - 10h	TCC	TCC	TCC	TCC	Orientação	
10h10 - 11h	TCC	TCC	TCC	TCC	Orientação	
11h - 11h50	TCC	TCC	TCC	TCC	Orientação	
13h - 13h50						
13h50 - 14h40						
14h40 - 15h30						
15h40 - 16h30						

16h30 - 17h20						
17h20 - 18h10						
18h50 - 19h40						
19h40 - 20h30						
20h30 - 21h20				TCC		
21h30 - 22h15					TCC	

## REFERÊNCIAS

ALPAYDIN, Ethem. **Introduction to Machine Learning**. 2ª Edição. MIT Press, 2010. ISBN 978-0-262-01243-0.

CATAE, Fabrício S.; ROCHA, Ricardo L. A. Introdução a Árvores de Decisão Adaptativas. In: WORKSHOP DE TECNOLOGIA ADAPTATIVA, 5. São Paulo. **Anais eletrônicos...** São Paulo: 2011. Disponível em: <<http://lta.poli.usp.br/lta/publicacoes/artigos/2011/catae-e-rocha-2011-introducao-a-arvores-de-decisao-adaptativas/view>>. Acesso em: 01 abr. 2011.

CORMEN, Thomas H.; LEISERSON, Charles E.; RIVEST, Ronald L.; STEIN, Clifford. **Introduction to Algorithms**. 3ª Edição. MIT Press, 2009. ISBN 978-0-262-03384-8.

MITCHELL, Tom M. **Machine Learning**. 1ª Edição. McGraw-Hill, 1997. ISBN: 0070428077.

NETO, J. J.: **Solving complex problems with Adaptive Automata**. Lecture Notes in Computer Science. S. Yu, A. Paun (Eds.): Implementation and Application of Automata 5th International Conference, CIAA 2000, Vol.2088, London, Canada, Springer-Verlag, pp.340, 2000.

NETO, João J. Um levantamento da pesquisa em técnicas adaptativas na eusp. **Sistemas e Computação**, Salvador, v.1, n.1, p. 23-47, jan./jun. 2011.

PISTORI, Hemerson; NETO, João J. **Adaptree – Proposta de um Algoritmo para Indução de Árvores de Decisão Baseado em Técnicas Adaptativas**. 2002. Anais Conferência Latino Americana de Informática - CLEI. Montevideo, Uruguai, 2002.

PISTORI, Hemerson. **Tecnologia Adaptativa Em Engenharia De Computação: Estado Da Arte e Aplicações**. 2003. 174p. Tese - Escola Politécnica da Universidade de São Paulo, São Paulo, 2003.

STANGE, Renata L. **Adaptatividade em aprendizagem de máquina: conceitos e**

**estudo de caso.** 2011. 98p. Dissertação - Escola Politécnica da Universidade de São Paulo, São Paulo, 2011.

UTGOFF, P. E.: *Incremental Induction of Decision Trees*. Machine Learning, Machine Learning, 4, 161-186, 1989.

YOSHIDA, Murilo L. **Aprendizado supervisionado incremental de Redes Bayesianas para mineração de dados.** 2007. 132p. Dissertação - São Carlos : UFSCar, 2007.