**Mining High-Speed Data Streams**

[P. Domingos, G. Hulten, Mining high speed data streams, in: Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, 2000, pp. 71-80.]

Aplicação do algoritmo VFDT para mineração de fluxo contínuo de acesso de dados via web do Campus principal da Universidade de Washington

O algoritmo VFDT utiliza o limite Hoeffding para garantir que a saída é assintoticamente(numa curva plana) proximamente idêntica do aprendiz convencional.

A árvore Hoeffding pode ser aprendida no tempo constante. O pior caso é proporcional ao número de atributos.

Aplicação do VFDT(Very Fast Decision Tree learner ) para um fluxo de dados de alta velocidade de dados de log da web é em curso.

**Pontos positivos**

O modelo de aprendizado é similar ao aprendizado obtido no modo batch

Construção de árvores complexas com custo computacional aceitável

Projeto de árvore de decisão para Base de dados extremamente grande (potencialmente infinita)

**Pontos negativos**

O algoritmo ID5R não é viável para aprendizado de fluxo de dados em alta velocidade

**Accurate Decision Trees for Mining High-speed Data Streams**

[J. Gama, R. Rocha, P. Medas, Accurate decision trees for mining high-speed data streams, in: Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2003, pp. 523-528.]

Estudo do problema da construção da acurácia da decisão do modelo em árvore para fluxo de dados.

Comparação dos algoritmos VFDT com C4.5

Modelo em árvore tem alto grau de interpretabilidade

Na configuração de classificação, a constante que minimiza o 0-1 função de perda é o modo do atributo etiqueta dos exemplos que caem nesta folha

Comparação da performance do VFDTc (algoritmo de Classificação) usando duas estratégias diferentes de classificação Naive Bayes e Majority Class.

**Pontos positivos**

**Pontos negativos**

**Estimating online vacancies in real-time road traffic monitoring with traffic sensor data stream**

[F. Wang, L. Hu, D. Zhou, R. Sun, Estimating online vacancies in real-time road traffic monitoring with traffic sensor data stream, Ad Hoc Networks 35 (2015) 1-11.]

Aplicação de Regressão Linear Múltipla

**IDUTC: An Intelligent Decision-Making System for Urban Traffic-Control Applications**

[M. Patel, N. Ranganathan, IDUTC: an intelligent decision-making system for urban traffic-control applications, IEEE Trans. Veh. Technol. 50 (3)(2001)816-829.]

**Mining Time-Changing Data Streams**

[G. Hulten, L. Spencer, P. Domingos, Mining time-changing data streams, in: Proceedings of 7th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2001, pp. 97-106.]

Algoritmos CVFDT, VFDT

Propor um algoritmo eficiente mineração de árvore de decisão de fluxo de dados com mudança constante, CVFDT.

**Learning model trees from envolving data streams**

[E. Ikonomovska, J. Gama, S. Džeroski , Learning model trees from envolving data streams, Data Min. Knowl. Discovery J. 23 (1)(2010) 1-41.]

Utiliza o algoritmo FIMTDD devido à sua facilidade de interpretação e bom desempenho preditivo, contudo, regressão sobre fluxo de dados de mudança de tempo (time-changing) é relativamente inexplorada e os problemas não são triviais.

O processamento de imagens de modelo de fluxo de fluxos de dados é, em geral, essencial para o cenário de fluxo contínuo. Primeiro, o conjunto de dados não é mais finito e disponível antes do aprendizado. Como resultado, é impossível armazenar todos os dados na memória e aprender com eles como um todo. Em segundo lugar, as varreduras sequenciais múltiplas sobre os dados do treinamento não são permitidas. Um algerar deve-se para coletar informação relevante na velocidade que chega e decidir incremental sobre decisões de rachadura. Em terceiro lugar, o conjunto de dados de treinamento pode consistir em dados de várias distribuições diferentes. Assim, o modelo precisa de monitoramento contínuo e atualização sempre que uma mudança é detectada. Desenvolveu-se um ganho incremental para o estudo de modelos para endereçar essas questões, denominadas árvores de modelo incremental rápido com detecção de deriva (FIMT-DD). O pseudo-código do algoritmo é dado na Fig. 1. O algoritmo começa com uma folha vazia e lê exemplos na ordem de chegada. Cada exemplo é atravessado para uma folha onde as estatísticas necessárias são atualizadas. Dada a primeira porção de instâncias 1, o algoritmo encontra a melhor divisão para cada atributo e, em seguida, classifica os atributos de acordo com alguma medida de avaliação. Se o critério de divisão é satisfeito, ele faz um cuspe no melhor atributo, criando duas folhas novas, uma para cada ramo da divisão. A partir da chegada de novas instâncias, a divisão criada é transmitida ao longo dos ramos correspondentes ao resultado do teste Na divisão por seus valores. Os testes de detecção de alterações são atualizados com cada exemplo do fluxo. Se for detectada uma alteração, será realizada uma adaptação da estrutura em árvore. Os detalhes do algoritmo são dados nas próximas seções.

**Pontos positivos**

**Pontos negativos**

**Traffic big data prediction and visualization using Fast Incremental Model Trees-Drift Detection (FIMT-DD)**

**Algoritmos de Regressão**

Regressão Linear

FIMTDD

<http://www.alelopesquisa.com.br/mobilidade.html>

pesquisa-de-mobilidade-alelo.pdf na pasta Artigos

**Reunião do dia 08/02**

O algoritmo VFDT é classificador

Sugestão

* Trocar o algoritmo VFDT pelo FIMTDD porque é **regressor**

Foram encontrados 2 artigos utilizando o algoritmo FIMTDD

Carcterísticas:

**MELHORAR A TRADUÇÃO**

Ver se faz sentido juntando

Árvore de modelo e regressão(Regression and model trees) tem a facilidade de interpretação e bom desempenho preditivo, contudo, regressão sobre fluxo de dados de mudança de tempo (time-changing) é relativamente inexplorada e os problemas não são triviais.(**Learning model trees from evolving data streams, Elena Ikonomovska, João Gama, Saso Dzeroski**)página 129

O processamento de imagens de modelo de fluxo de fluxos de dados é, em geral, essencial para o cenário de fluxo contínuo. Primeiro, o conjunto de dados não é mais finito e disponível antes do aprendizado. Como resultado, é impossível armazenar todos os dados na memória e aprender com eles como um todo. Em segundo lugar, as varreduras sequenciais múltiplas sobre os dados do treinamento não são permitidas. Um algoritmo deve-se para coletar informação relevante na velocidade que chega e decidir incremental sobre decisões de rachadura. Em terceiro lugar, o conjunto de dados de treinamento pode consistir em dados de várias distribuições diferentes. Assim, o modelo precisa de monitoramento contínuo e atualização sempre que uma mudança é detectada. Desenvolveu-se um ganho incremental para o estudo de modelos para endereçar essas questões, denominadas árvores de modelo incremental rápido com detecção de deriva (FIMT-DD). O pseudo-código do algoritmo é dado na Fig. 1.

O algoritmo começa com uma folha vazia e lê exemplos na ordem de chegada. Cada exemplo é atravessado para uma folha onde as estatísticas necessárias são atualizadas. Dada a primeira porção de instâncias 1, o algoritmo encontra a melhor divisão para cada atributo e, em seguida, classifica os atributos de acordo com alguma medida de avaliação. Se o critério de divisão é satisfeito, ele faz um cuspe no melhor atributo, criando duas folhas novas, uma para cada ramo da divisão. A partir da chegada de novas instâncias, a divisão criada é transmitida ao longo dos ramos correspondentes ao resultado do teste Na divisão por seus valores. Os testes de detecção de alterações são atualizados com cada exemplo do fluxo. Se for detectada uma alteração, será realizada uma adaptação da estrutura em árvore. (**Learning model trees from evolving data streams, Elena Ikonomovska, João Gama, Saso Dzeroski**)página 135

A ferramenta MOA(Massive **Online** Analysis) trabalha com o algoritmo FIMTDD (Fast Incremental Model Trees-Drift Detection)

O FIMTDD trabalha com grande volume de dados predizendo condição de fluxo de tráfico

**Tarefas**

1. **Mandar email para José (ok)**
2. **Ver uma base de dados no MOA para fazer testes (ok)**
3. **Verificar nos artigos como é definido o tamanho da amostra ou janela ou window**
4. **Pegar o livro de Inteligência Artificial para buscar informações de Fluxo de Dados (ok)**

**O link dos datasets** [**http://moa.cs.waikato.ac.nz/datasets/**](http://moa.cs.waikato.ac.nz/datasets/)

**Agradecimentos**

**Agradeço também a minha orientadora, Profª. Flávia Bernardini e co-orientador Patrick Moratori, pela oportunidade dada, pela confiança em meu trabalho, e especialmente, por ter me ensinado tudo o que sei sobre pesquisa. Deixo parte desse mérito também ao Prof. Basílio, ao aluno de graduação de Ciência da Computação, José Santos, cuja colaboração foi inestimável dentro de minha pesquisa. Não tenho palavras para agradecer a ambos professores por suas respectivas contribuições e amizades**

**Dicas de apresentação e preparação da dissertação**

* **Apresentação verbal formal, sem gírias**
* **Introdução ou resumo(eu não lembro) contar uma historinha até chegar o problema, breve definições(descrições) do algoritmos que serão utilizados**
* **aumentar a figura, aproveitando melhor o espaço em branco**
* **colocar fontes que ao imprimir fique legível**
* **Atentar para o título estar em sincronia com o conteúdo da dissertação**

**Delimitação do problema é falar que o tempo não será considerado como variável para predição(?)**

**Reunião 24/02/2017**

**No artigo Traffic big data prediction and visualization using Fast Incremental Model Trees-Drift Detection (FIMT-DD)**

**Tudo estava dando a entender que era um problema de regressão porque neste artigo foi utilizado o algoritmo FIMT-DD para analisar e predizer uma base de dados muito grande de tráfego.**

**Existe possibilidade de fazer predição em um problema considerado de classificação?**