UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APRENDIZADO SEMISSUPERVISIONADO COMO TÉCNICA DE MINERAÇÃO EM FLUXOS CONTÍNUOS DE DADOS

PRISCILLA DE ABREU LOPES

ORIENTADORA: PROFA. DRA. HELOISA DE ARRUDA CAMARGO

São Carlos – SP

Junho/2014

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APRENDIZADO SEMISSUPERVISIONADO COMO TÉCNICA DE MINERAÇÃO EM FLUXOS CONTÍNUOS DE DADOS

PRISCILLA DE ABREU LOPES

Qualificação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Heloisa de Arruda Camargo

São Carlos – SP

Junho/2014

RESUMO

....

Palavras-chave: aprendizado semissupervisionado, fluxos contínuos de dados, agrupamento, fuzzy

ABSTRACT

....

Keywords: semi-supervised learning, data streams, clustering, fuzzy

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM – *Aprendizado de Máquina*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo introduz o contexto e a motivação que levaram à elaboração de uma proposta ...

1.1 Contexto e Motivação

Capítulo 2

CONCEITOS GERAIS

Breve introdução ao capítulo.

Neste capítulo são apresentados conceitos gerais a respeito de agrupamento de dados semissupervisionado e de aprendizado em fluxos contínuos de dados, este último com foco em características específicas que devem ser consideradas quando realizada aprendizagem por agrupamento.

2.1 Agrupamento Semissupervisionado

Aprendizado de máquina refere-se à investigação de métodos computacionais capazes de adquirir conhecimento de forma automática. Um dos mecanismos utilizados para derivar conhecimento novo é por meio de inferência indutiva sobre um conjunto de dados ou exemplos . O aprendizado indutivo pode ser dividido em três abordagens: supervisionada, não supervisionada e semissupervisionada.

Abordagens supervisionadas são aquelas que realizam a extração de conhecimento pelo desenvolvimento de um modelo baseado em um conjunto de dados totalmente rotulado. Um dado é dito rotulado se a classe à qual pertence é conhecida. Técnicas de classificação e regressão tipicamente pertencem a esta categoria e são amplamente utilizadas por produzirem bons resultados. Abordagens conhecidas e referências

Apesar dos bons resultados por técnicas supervisionadas, é possível que a informação de classes não esteja disponível para determinados domínios, impedindo sua aplicação. Neste contexto normalmente são aplicadas técnicas não supervisionadas de aprendizado.

Agrupamento de dados é uma típica técnica não supervisionada, ou seja, um processo capaz

de realizar aprendizagem a partir de um conjunto de dados não rotulado. A aplicação de agrupamento tem como objetivo definir uma possível partição dos dados em grupos, de forma que exemplos semelhantes pertençam a um mesmo grupo e exemplos distintos pertençam a grupos distintos. Essa divisão dos dados é baseada em métricas para determinar a relação entre diferentes exemplos. A Seção 2.1.1 descreve o processo de agrupamento e características particulares dentro deste processo.

O crescimento acelerado de conjuntos de dados em muitos domínios torna a rotulação manual e total dos dados onerosa. O aprendizado semissupervisionado tem como base técnicas supervisionadas ou não supervisionadas, adaptadas a fim de contornar o problema da falta de rótulos, sendo mais explorado nos últimos 10 anos.

O agrupamento semissupervisionado, em particular, é realizado por técnicas que incluem mecanismos para a consideração de informação pré-existente no processo de geração de grupos.

Existem duas abordagens para a incorporação de semissupervisão em técnicas de agrupamento, dependentes do conhecimento prévio disponível. A abordagem por sementes utiliza uma parte pequena do conjunto de dados rotulada, chamados exemplos sementes. A abordagem por restrições entre pares define duas relações entre pares de exemplos que podem ser utilizadas no processo de agrupamento: *must-link*, indicando que um par de exemplos deve pertencer ao mesmo grupo, ou *cannot-link*, indicando que os exemplos do par devem pertencer a grupos distintos.

2.1.1 Processo de agrupamento

Jain, Murty e Flynn (1999) definem que uma atividade de agrupamento segue os passos brevemente descritos:

1. **Preparação de exemplos**: determina a representação dos dados, podendo ser aplicada alguma transformação ao conjunto, como normalização de domínio e seleção/extração de atributos;

2. **Métrica de comparação**: passo para a escolha de uma métrica para comparação apropriada ao domínio da aplicação, geralmente fornecida por uma função de distância definida entre pares de exemplos (métrica de dissimilaridade);

3. **Agrupamento**: aplicação de um algoritmo com o objetivo de agrupar os dados. Nesta etapa podem ser aplicados inúmeros algoritmos, não supervisionadas ou semissupervisi-

onadas, que tenham como resposta uma partição rígida ou *fuzzy* do conjunto original; até agora nada foi falado de fuzzy, mas acho que pode incluir na primeira parte da seção (?)

4. **Validação**: este passo visa determinar se o resultado da partição é significativo, geralmente realizando o cálculo de valor para índice de validação;

5. **Interpretação dos resultados**: passo em que são examinados os resultados com relação a seus exemplos, com o objetivo de determinar a natureza dos grupos.

falar mais a respeito de métricas, algoritmos, validação? apenas citar dentro dos passos ou colocar uma subsubsection?

conclusão seção.

2.2 Aprendizado em Fluxos Contínuos de Dados

(GAMA; GABER, 2007) (AGGARWAL, 2007) (AGGARWAL JIAWEI HAN, 2007) (RAJARAMAN; ULLMAN, 2011) (LESKOVEC; RAJARAMAN; ULLMAN,)

Características gerais, aspectos a considerar, desafios gerais?

Baseados em técnicas comuns atentando aos aspectos específicos do contexto de Streams.

Classificação x agrupamento?

2.2.1 Agrupamento em fluxos contínuos de dados

desafios específicos?

2.2.1.1 Representação

2.2.1.2 Modelo de janela*

2.2.1.3 Detecção de *Outliers**

2.2.1.4 Tempo*

conclusão seção.

2.3 Considerações Finais

finalizar o capítulo.

Capítulo 3

AGRUPAMENTO EM FLUXOS CONTÍNUOS DE DADOS

Visão geral, baseados em técnicas comuns atentando aos aspectos específicos do contexto de Streams.

Ref abordagens de ”classificação”.

3.1 Árvore de Hoedding

3.2 something

3.3 something else

3.4 Considerações Finais

Capítulo 4

PROPOSTA DE TRABALHO

4.1 Atividades Principais

4.2 Cronograma de Atividades

4.3 Contribuições Esperadas

4.4 Considerações Finais

REFERÊNCIAS

AGGARWAL, C. C. Data streams - models and algorithms. In: ____. [S.l.]: Springer, 2007. cap. An Introduction to Data Streams, p. 1–8.

AGGARWAL JIAWEI HAN, J. W. P. S. Y. C. C. On clustering massive data streams: A summarization paradigm. In: AGGARWAL, C. C. (Ed.). *Data Streams - Models and Algorithms*. [S.l.]: Springer, 2007. p. 9–38.

GAMA, J. a.; GABER, M. M. (Ed.). *Learning from Data Streams: Processing Techniques in Sensor Networks*. [S.l.]: Springer, 2007.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 31, n. 3, p. 264–323, September 1999.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. Mining of massive datasets. 2ª Edição. Disponível em: <http://infolab.stanford.edu/ ullman/mmds.html>.

RAJARAMAN, A.; ULLMAN, J. D. *Mining of Massive Datasets*. New York, NY, USA: Cambridge University Press, 2011.