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

FCM – *Fuzzy C-Means*

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 Sistemas *Fuzzy*

2.2 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 geral baseado em um conjunto de dados totalmente rotulado. Um exemplo de um conjunto de dados é dito rotulado se a classe à qual pertence é conhecida. Métodos de classificação e regressão tipicamente pertencem a esta categoria e são amplamente utilizadas por produzirem bons resultados.

A maioria dos métodos de classificação utilizam-se de um conjunto de exemplos de treinamento para a construção de um classificador, comumente constituído de um conjunto de regras ou uma estrutura da qual possam ser extraídas regras de classificação. Um conjunto de exem-

plos de teste independente do conjunto de treinamento é aplicado ao classificador no intuito de verificar a qualidade do resultado obtido na etapa de construção. Se a avaliação for satisfatória, o classificador poderá ser aplicado a conjuntos de novos exemplos. Alguns métodos podem requerer um ajuste do classificador após um período de tempo ou o aumento do volume de dados. Esse ajuste é, por vezes, realizado pela aplicação dos passos de criação e verificação de um classificador.

Aplicações de árvores de decisão (QUINLAN, 1986), redes neurais (BISHOP, 1995) e métodos estatísticos (DUDA; HART, 1973) fazem parte do conjunto de tentativas para a resolução do problema de classificação (MITCHELL, 1997). Existem métodos, como o *K-Nearest Neightbors* (COVER; HART, 1967), que não geram classificadores, mas utilizam a informação de rótulos para classificar novos exemplos, atribuindo classes por meio de métricas de similaridade.

Variações de métodos de classificação podem realizar a indução de regras *fuzzy* a partir de um conjunto de dados. Sistemas neuro-*fuzzy* (KLOSE et al., 2001) se utilizam de algoritmos de aprendizado derivados da teoria de redes neurais para gerar regaras *fuzzy*. Outras bordagens são baseadas em árvores de decisão, que podem ser induzidas e, posteriormente, ter regras extraídas da estrutura resultante (QUINLAN, 1993). Propostas para extensões chamadas árvores de decisão *fuzzy* também podem ser encontradas na literatura (JANIKOW, 1998), outros?.

Estratégias evolutivas, como Algoritmos Genéticos, são utilizados na otimização e criação de sistemas *fuzzy*. Sua habilidade é otimizar a estrutura e parâmetros de modelos, enquanto grande parte das estratégias de otimização é capaz apenas de adaptar parâmetros de um modelo (KLOSE; KRUSE, 2005).

Apesar das técnicas supervisionadas produzirem bons resultados, 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.

2.2.1 Separar com subseção a partir daqui?

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 que determinam a relação de dissimilaridade ou similaridade entre diferentes exemplos. Problemas como forte dependência de medidas de distância e normalização dos dados, definição do número correto de grupos para a divisão são observados quando aplicadas técnicas de agrupamento não supervisionadas.

O algoritmo *k-means* (MACQUEEN, 1967) é uma das mais populares e simples técnicas de agrupamento. O *k-means* ainda é amplamente utilizado e, muitas vezes, serve de base e inspiração para o desenvolvimento de novos algoritmos. O objetivo deste algoritmo é agrupar os dados em k grupos disjuntos, de maneira que a distância entre os exemplos pertencentes a um grupo e seu respectivo centro seja mínima. O centro de grupo, ou protótipo, representa o ponto médio entre os exemplos de um grupo.

O *Fuzzy C-Means* (FCM) (BEZDEK, 1981) é um algoritmo que implementa uma extensão *fuzzy* do algoritmo *k-means*. continuar mais um pouco.

2.2.2 Separar com subseção a partir daqui também?

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. Abordagens conhecidas e referências

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.

A Seção 2 descreve o processo de agrupamento e características particulares dentro deste processo, relevantes para abordagens não supervisionadas e semissupervisionadas.

2.2.3 Processo de agrupamento

Passar esta parte para a de agrupamento não superisionado?

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

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 semissupervisionadas, que tenham como resposta uma partição do conjunto original;

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 é examinada a partição resultante 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.

2.3 Aprendizado em Fluxos Contínuos de Dados

As técnicas de aprendizado citadas e referenciadas nas Seções 2 e 2.1 consideram características particulares para os dados disponíveis. Para essas propostas assume-se que o conjunto de dados é finito, os exemplos seguem uma distribuição estática e estão disponíveis para acesso sempre que necessário durante o processo de aprendizagem.

A evolução da tecnologia, a internet e o aumento significativo de seu número de usuários permitiu o desenvolvimento de domínios para os quais as características assumidas pelas abordagens mais clássicas de aprendizado não são verdadeiras.

Existe hoje uma variedade de sistemas que produzem grande quantidade de dados em curto espaço de tempo. Estes conjuntos de dados têm tamanho indefinido, potencialmente infinito, e podem gerar exemplos com distribuição estatística mutável de acordo com o tempo.

(GAMA; GABER, 2007) (AGGARWAL, 2007) (AGGARWAL et al., 2007) (RAJARAMAN; ULLMAN, 2011)

Características gerais, aspectos a considerar, desafios gerais?

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

Classificação x agrupamento?

2.3.1 Agrupamento em fluxos contínuos de dados

desafios específicos?

2.3.1.1 Representação

2.3.1.2 Modelo de janela*

2.3.1.3 Detecção de *Outliers**

2.3.1.4 Tempo*

conclusão seção.

2.4 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. An introduction to data streams. In: AGGARWAL, C. C. (Ed.). *Data Streams - Models and Algorithms*. [S.l.]: Springer, 2007. p. 1–8.

AGGARWAL, C. C. et al. 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.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

BISHOP, C. M. *Neural Networks for Pattern Recognition*. [S.l.]: Oxford University Press, 1995.

COVER, T. M.; HART, P. E. Nearest neighbor pattern classification. *IEEE Transactions in Information Theory*, IT-13, n. 1, p. 21–27, 1967.

DUDA, R. O.; HART, P. E. *Pattern classification and scene analysis*. [S.l.]: John Wiley and Sons, 1973.

GAMA, J.; 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.

JANIKOW, C. Z. Fuzzy decision trees: issues and methods. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, v. 28, n. 1, p. 1–14, Feb 1998.

KLOSE, A.; KRUSE, R. Semi-supervised learning in knowledge discovery. *Fuzzy Sets and Systems*, v. 149, p. 209–233, 2005.

KLOSE, A. et al. Data mining with neuro-fuzzy models. In: KANDEL, A.; LAST, M.; BUNKE, H. (Ed.). *Data Mining and Computational Intelligence*. Heidelberg, Germany: Physica-Verlag GmbH, 2001. p. 1–35.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: *In Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.]: University of California Press, 1967. v. 1, p. 281–297.

MITCHELL, T. *Machine Learning*. [S.l.]: McGraw-Hill Education (ISE Editions), 1997.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

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