UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APRENDIZADO EM FLUXOS CONTÍNUOS DE DADOS POR MEIO DE TÉCNICAS SEMISSUPERVISIONADAS

PRISCILLA DE ABREU LOPES

ORIENTADORA: PROFA. DRA. HELOISA DE ARRUDA CAMARGO

São Carlos – SP

Março/2015

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APRENDIZADO EM FLUXOS CONTÍNUOS DE DADOS POR MEIO DE TÉCNICAS SEMISSUPERVISIONADAS

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

Março/2015

RESUMO

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

ABSTRACT

....

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

LISTA DE FIGURAS

Lista de Algoritmos

LISTA DE TABELAS

ACRÔNIMOS E SIGLAS

AH – *Árvores de Hoeffding*

AM – *Aprendizado de Máquina*

AP – *Affinuty Propagation*

CVFDT – *Concept-adapting Very Fast Decision Tree*

EC – *Ensemble de Classificadores*

FCD – *Fluxos Contínuos de Dados*

FCM – *Fuzzy C-Means*

FPT – *Fuzzy Pattern Trees*

PA – *Passive-Agressive*

SVM – *Support Vector Machine*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

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

Capítulo 2

CONCEITOS GERAIS

O Aprendizado de Máquina (AM) refere-se à investigação de métodos computacionais capazes de adquirir conhecimento de forma automática. Desde a formalização do surgimento desta área de pesquisa, na década de 80 (LANGLEY, 2011), distintas abordagens foram propostas para a realização do processo de aprendizagem.

Aspectos como a evolução e ampliação do acesso a novas tecnologias e a internet tornaram propício o surgimento e desenvolvimento de diferentes e novos domínios. Para as novas características e desafios que despontaram neste contexto, as técnicas mais clássicas de AM já não obtiveram o mesmo sucesso e, então, começaram a surgir novas abordagens na tentativa de encontrar métodos capazes de lidar com novas peculiaridades desses domínios.

Neste capítulo são apresentados conceitos gerais que fundamentam a compreensão do problema tratado neste trabalho, bem como a proposta de pesquisa apresentada. Tais conceitos relacionam-se principalmente a aprendizado semissupervisionado e aprendizado em fluxos contínuos de dados.

2.1 Aprendizado Semissupervisionado

No contexto de AM, a inferência indutiva é um dos principais mecanismos utilizados para derivar conhecimento novo e predizer escritos futuros. No aprendizado indutivo o conhecimento é aprendido por meio de inferência indutiva sobre um conjunto de dados: objetos (também chamados de exemplos ou instâncias) que são descritos por um conjunto de atributos (MITCHELL, 1997). O aprendizado indutivo pode ser dividido em três abordagens: supervisionada, não supervisionada e semissupervisionada.

2.1.1 Aprendizado Supervisionado e Não Supervisionado

Abordagens supervisionadas são aquelas que realizam a extração de conhecimento pelo desenvolvimento de um modelo geral baseado em um conjunto de dados que possuem um atributo especial, chamado classe, que representa o conceito que se deseja aprender. Um exemplo de um conjunto de dados é dito rotulado se a classe à qual pertence é conhecida. Métodos conhecidos como de classificação tipicamente utilizam-se de conjuntos totalmente rotulados e, portanto, pertencem à categoria de aprendizado supervisionado. Estes métodos são amplamente utilizados por produzirem bons resultados (WITTEN; FRANK, 2005).

A maioria dos métodos de classificação utilizam-se de um conjunto de exemplos de treinamento para a construção de um classificador. Tais classificadores são constituídos de um conjunto de regras ou uma estrutura da qual possam ser extraídas regras de classificação. Um conjunto de exemplos 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 com classe desconhecida. Alguns métodos podem requerer um ajuste do classificador após um período de tempo ou o aumento do volume de dados.

Aplicações de árvores de decisão (QUINLAN, 1986), redes neurais (BISHOP, 1995), métodos estatísticos (DUDA; HART, 1973) e genéticos (GOLDBERG, 1989) fazem parte do conjunto de paradigmas 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 baseados na teoria de conjuntos *fuzzy* (ZADEH, 1965) podem realizar a indução de regras que permitem a representação de conhecimento impreciso a partir de um conjunto de dados (PEDRYCZ; GOMIDE, 1998). Sistemas neuro-*fuzzy* (KLOSE et al., 2001) se utilizam de algoritmos de aprendizado derivados da teoria de redes neurais para gerar regras *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; CINTRA; MONARD; CAMARGO, 2012).

Estratégias evolutivas, como Algoritmos Genéticos, são utilizados na otimização e criação de sistemas *fuzzy*. Inicialmente, os chamados Sistemas *Fuzzy* Genéticos, possuíam grande foco na geração de sistemas com alta acurácia (CORDÓN, 2011). Este paradigma foi modificado e há

nas pesquisas mais recentes uma preocupação em aproveitar o potencial de interpretabilidade dos conjuntos *fuzzy* para a geração e otimização de sistemas que, além de alta acurácia, sejam mais claros e interpretáveis para seres humanos (CORDÓN, 2011; FAZZOLARI et al., 2013).

Apesar dos bons resultados produzidos por técnicas supervisionadas, é possível que as classes não estejam disponíveis 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 (JAIN; MURTY; FLYNN, 1999). Essa divisão dos dados é baseada em métricas que determinam a relação de dissimilaridade ou similaridade entre diferentes exemplos.

As diferentes técnicas de agrupamento podem ser divididas nas seguintes categorias (HAN; KAMBER; PEI, 2012):

Hierárquico: cria uma decomposição hierárquica de um conjunto de exemplos de acordo com algum critério (DAY; EDELSBRUNNER, 1984; KAUFMAN; ROUSSEEUW, 1990; ZHANG; RAMAKRISHNAN; LIVNY, 1996);

Particional: constrói uma partição inicial de um conjunto de exemplos e, por meio de um processo iterativo, busca melhorar a partição, mudando exemplos de grupo baseado, geralmente, em uma medida de distância (MACQUEEN, 1967; BEZDEK, 1981; KAUFMAN; ROUSSEEUW, 1990);

Baseado em Densidade: baseado em funções densidade, é capaz de criar uma partição ou uma decomposição hierárquica de um conjunto de exemplos (ESTER et al., 1996; HINNEBURG; KEIM, 1998; ANKERST et al., 1999);

Baseado em Grades: todas as operações de agrupamento são realizadas dentro de uma estrutura de grades (*grid*), que é uma divisão do espaço dos exemplos em um número finito de células (WANG; YANG; MUNTZ, 1997; SHEIKHOLESLAMI; CHATTERJEE; ZHANG, 1998).

É relevante mencionar que dentro dos conjuntos descritos é possível encontrar técnicas que utilizam conceitos da teoria de conjuntos *fuzzy*. O *Fuzzy C-Means* (FCM) (BEZDEK, 1981), por exemplo, é uma proposta pioneira, uma das primeiras extensões *fuzzy* do algoritmo *k-means* (MACQUEEN, 1967).

O algoritmo *k-means* é um dos mais populares e simples algoritmos de agrupamento, ainda sendo amplamente utilizado e, muitas vezes, servindo de base ao desenvolvimento de novos algoritmos. O objetivo do *k-means* é agrupar os dados em k grupos disjuntos, de maneira que a soma das distâncias entre os exemplos pertencentes a um grupo e seu respectivo centro seja mínima. O centro de grupo, ou protótipo, representa o ponto médio dos pontos pertencentes a um determinado grupo. No FCM a partição dos dados é realizada em grupos que podem ser não disjuntos, cada exemplo possuindo um grau de pertinência para cada k grupo.

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 crescimento acelerado de conjuntos de dados em muitos domínios torna a rotulação manual e total dos dados onerosa. A aplicação de técnicas supervisionadas pode ser prejudicada por utilizar apenas uma pequena quantidade de dados rotulados. Ao mesmo tempo, a utilização de técnicas não supervisionadas desconsideraria totalmente esse conhecimento prévio disponível no processo de aprendizagem. Nesse contexto, surge a ideia de aprendizado semissupervisionado, apresentada na Seção 2.1.2.

2.1.2 Abordagens de Aprendizado Semissupervisionado

A ideia de exploração de informações rotuladas e não rotuladas data da década de 80 (PEDRYCZ, 1985; BOARD; PITT, 1989), mas vem sendo mais explorada, principalmente, na última década (CHAPELLE; SCHÖLKOPF; ZIEN, 2006; SCHWENKER; TRENTIN, 2014).

O aprendizado semissupervisionado tem como base técnicas supervisionadas ou não supervisionadas, adaptadas a fim de realizar a aprendizagem utilizando conjuntos parcialmente rotulados e/ou algum outro tipo de informação prévia já disponível.

visão geral: colocar ideia geral, a separação entre classificação semissupervisionada e agrupamento semissupervisionado e os tipos baseado em sementes e restrições. Ver texto mestrado (SEEGER, 2001; ZHU, 2005; CHAPELLE; SCHÖLKOPF; ZIEN, 2006; ZHU; GOLDBERG, 2009; SCHWENKER; TRENTIN, 2014)

D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks **(SHAMSHIRBAND et al., 2014)**

Data Understanding Using Semi-Supervised Clustering **(BHATNAGAR et al., 2012)**

Label-based Semi-supervised Fuzzy Co-clustering for Document Categorization **(YAN; CHEN, 2011)**

On Semi-supervised Fuzzy c-Means Clustering with Clusterwise Tolerance by Opposite Criteria **(HAMASUNA; ENDO, 2011)**

As técnicas de aprendizado citadas e referenciadas nesta seção 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 propiciaram o surgimento de domínios para os quais as características assumidas pelas abordagens mais clássicas de aprendizado não são verdadeiras. Nesse contexto, teve origem uma nova abordagem de tratamento dessa forma de aprendizado, denominado de aprendizado em fluxo contínuo de dados.

2.2 Aprendizado em Fluxos Contínuos de Dados

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

O surgimento e crescimento deste tipo de sistemas impulsionaram a pesquisa por técnicas que pudessem realizar a aprendizagem considerando as características específicas por estes domínios, referidos como Fluxos Contínuos de Dados (FCD) (em inglês *Data Streams* ou *Streaming Data*). A Figura 2.1 traz um gráfico que mostra uma visão geral do crescimento no número de publicações sobre aprendizado/mineração em FCD.

Data Stream/Very Large Data *Fuzzy c-Means Algorithms for Very Large Data* (HAVENS et al., 2012)

Learning Streams vs online learning

Trabalhos base para texto sobre aprendizado em FCD:

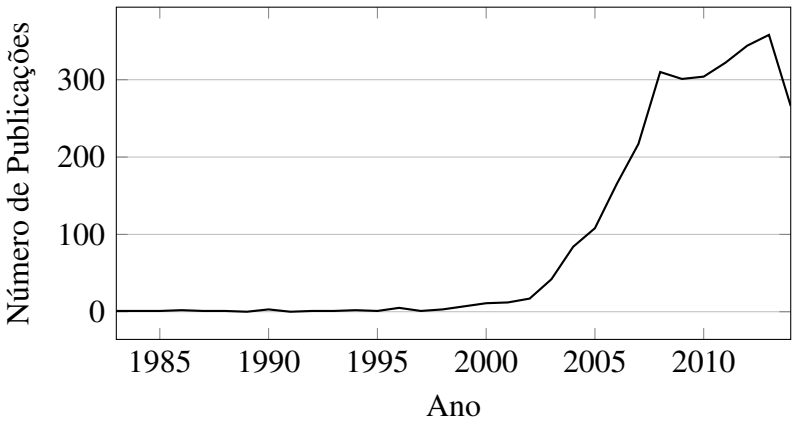

Figura 2.1: Progressão anual do número de publicações em inglês considerando o resultado de busca realizada na base Scopus, em 10 de janeiro de 2015, pela combinação dos termos *learning/mining* e *data streams/streaming data*

- *Models and issues in data stream systems* (BABCOCK et al., 2002)
- *Mining data streams: a review* (GABER; ZASLAVSKY; KRISHNASWAMY, 2005)
- *An Introduction to Data Streams* (AGGARWAL, 2007)
- *Learning from Data Streams: Processing Techniques in Sensor Networks* (GAMA; GABER, 2007)
- *Mining of Massive Datasets* (LESKOVEC; RAJARAMAN; ULLMAN, 2014) - capítulo sobre data streams

Outros trabalhos mais ligados a abordagens:

- *On Clustering Massive Data Streams: A Summarization Paradigm* (AGGARWAL et al., 2007)

2.2.1 Abordagem por Exemplo e por Atributo

Aprendizado em FCD por exemplo: um FCD, objetivo é gerar um modelo geral considerando cada exemplo/instância do FCD.

Aprendizado em FCD por atributo: múltiplos FCD, o objetivo é identificar um modelo geral considerando os diferentes comportamentos de diversos FCD ($\leftrightarrow$ relação com séries temporais)

2.2.1.1 Métodos Aprendizado FCD por Atributo – não precisa ser uma subseção

O *MINETRAC* (CASAS; MAZEL; OWEZARSKI, 2011) combina técnicas de aprendizado não supervisionado e semissupervisionado para identificação e classificação de diferentes classes de fluxos de tráfego de internet de características similares.

Uma proposta de potencial estrutura para a representação de exemplos de um FCD de forma compacta é apresentada por Chen, Chen e Sheng (2013), com o objetivo de, posteriormente, agrupar as estruturas de múltiplos FCD de forma não supervisionada.

Para a identificação de padrões em sequências de dados, como múltiplos FCD ou séries temporais, Li (2014) apresenta uma abordagem semissupervisionada baseada em grafo para propagação de rótulos e extensão do conjunto rotulado que realiza o treinamento de um classificador usando *Support Vector Machine* (SVM).

Patil, Fatangare e Kulkarni (2015) propõem um modelo de aprendizagem para o domínio de preços e demanda no fornecimento de eletricidade. A detecção e adaptação a mudanças em tendências e valores, capacidade de predição e adaptatividade do modelo são alguns dos desafios para os quais os autores buscam solução.

2.2.2 Árvores de Hoeffding

Apresentação das Árvores de Hoeffding (AH).

2.2.3 Aplicações

Relevant applications include network traffic monitoring, sensor network data analysis, Web click stream mining, power consumption measurement, dynamic tracing of stock fluctuations (ZHANG et al., 2012)

in real-world applications, such as credit fraud and intrusion detection (WU; LI; HU, 2012)

distribuição termal em transformadores (SOUZA et al., 2012)

Sistemas de recomendação, com teste em dados extraídos no The New York Times (MARIN et al., 2013) - dados implícitos em ações de usuários para adaptar o perfil de usuário de forma dinâmica

2.3 Considerações Finais

Este capítulo traz conceitos gerais relacionados a aprendizado semissupervisionado e em fluxos contínuos de dados, além das particularidades inerentes a estas abordagens. Esta síntese se faz necessária para situar o leitor, facilitando a compreensão do contexto investigado neste trabalho e permitindo entendimento mais claro do conteúdo que será apresentado no Capítulo 3.

Capítulo 3

Abordagens para Aprendizado em Fluxos Contínuos de Dados

Tendo em conta os aspectos envolvidos no aprendizado em FCD e o recente crescimento de esforços para encontrar soluções capazes de lidar com os desafios encontrados nesta variedade de domínios, evidenciado na Seção 2.2, julga-se interessante a investigação mais aprofundada das diversas propostas existentes na literatura para aprendizado em FCD.

As propostas de técnicas para aprendizado em FCD são, na maioria das vezes, adaptações de técnicas de aprendizado clássico para lidar com um ou mais desafios encontrados em domínios de FCD, e.g.: a necessidade de processar os dados logo que chegam, seja de forma *online* (exemplo a exemplo) ou considerando partes do conjunto de exemplos (processamento de *chunks*); a adaptação do modelo geral que representa o FCD e a otimização de suas estruturas; a detecção e tratamento de desvios de conceito.

Algumas técnicas de aprendizado incremental foram desenvolvidas para domínios específicos ou com foco em conjuntos de dados que, embora volumosos, não apresentam características de FCD, como, por exemplo, questões de desvio de conceito. Ainda que o foco principal das propostas não seja o aprendizado em FCD, certas abordagens podem ser aplicadas neste contexto, porém com algumas ressalvas, já que não possuem mecanismos para lidar com um ou outro aspecto intrínseco aos domínios FCD.

Este capítulo apresenta discussão sobre uma revisão bibliográfica de abordagens para aprendizado em FCD, desenvolvidas especificamente com este intuito ou não, considerando a abordagem de aprendizado por exemplos, em que o processo de aprendizagem tem seu foco nas instâncias de um único FCD (Seção 2.2.1).

3.1 Técnicas de Agrupamento em Fluxos Contínuos de Dados

Abordagens de agrupamento de dados são tipicamente utilizadas no aprendizado não supervisionado. Dentro de domínios de FCD é comum verificar a falta de informação de classe, seja por conta da natureza do domínio (não existem classes definidas) ou pela dificuldade em rotular exemplos de um FCD.

Data Stream Clustering: A Survey (SILVA et al., 2013)

Clustering Techniques for Streaming Data: A Survey (YOGITA; TOSHNIWAL, 2013)

On Density-Based Data Streams Clustering Algorithms: A Survey (AMINI; WAH; SABOOHI, 2014)

A proposta de Labroche (2014) está baseada no algoritmo *k*-medóides e realiza agrupamento *fuzzy* de forma incremental. O trabalho de Lemos, Caminhas e Gomide (2013) apresenta uma técnica para geração de um classificador *fuzzy* baseado em agrupamento incremental para a geração de regras que descrevem novos estados operacionais de um sistema de detecção e diagnóstico de falhas.

Hore, Hall e Goldgof (2007b) apresentam a proposta de uma abordagem genérica para agrupamento iterativo *fuzzy*/possibilístico em FCD, introduzindo equações objetivo transformadas para os algoritmo FCM (BEZDEK, 1981), *possibilitistc c-means* (KRISHNAPURAM; KELLER, 1996) e Gustafson-Kessel (GUSTAFSON; KESSEL, 1978). Outro trabalho (HORE; HALL; GOLDGOF, 2007a) traz uma nova variante do FCM para aprendizado em FCD, o *Streaming* FCM, que realiza adaptação à evolução de distribuições pela utilização de uma parte do histórico de protótipos/centróides no agrupamento de *chunks* de dados, conforme sua chegada. Em (HORE et al., 2008) é explorada uma extensão *online* para o FCM que mantém sumarização do agrupamento usando exemplos ponderados. Os exemplos ponderados obtidos pelo agrupamento de cada *chunk* de dados formam um *ensemble* que é transformado em um conjunto de grupos finais.

Uma característica das técnicas mencionadas até agora é o fato de serem baseadas em algoritmos de agrupamento particional em que um valor deve ser definido para o número de grupos, que servirá de entrada para o algoritmo. Para domínios em que o número de grupos é flutuante, i.e., alguns grupos podem desaparecer e pode ocorrer o surgimento de novos grupos, a definição de um valor fixo para número de grupos pode ser um problema.

Uma extensão do agrupamento Affinity Propagation (AP) (FREY; DUECK, 2007) para apren-

dizado em FCD é o algoritmo *Streaming* AP (ZHANG; FURTLEHNER; SEBAG, 2008). Usando método de passagem de mensagem, o AP escolhe, entre os exemplos disponíveis, aqueles que melhor representam o conjunto, os chamados *exemplars*, que indicam os diferentes grupos dentro do conjunto de exemplos. A proposta de extensão é dividida em dois passos, sendo que o objetivo do primeiro é encontrar os *exemplars* ponderados dentro de um *chunk* de dados por uma extensão do AP (*Weighted Affinity Propagation*), enquanto o segundo visa diminuir a complexidade do modelo pela aplicação do *Weighted Affinity Propagation* para o conjunto de *exemplars*. Em trabalho mais recente (ZHANG et al., 2014), o *Streaming* AP traz melhorias como mecanismo de detecção de mudanças e adaptação do modelo da distribução dos dados.

Os algoritmos de agrupamento baseados em densidade também são utilizados como alternativa para domínios onde determinar um valor fixo para número de grupos pode ser uma tarefa difícil.

A proposta de Cao et al. (2006) é o algoritmo de agrupamento em FCD baseado em densidade chamado DenStream, que utiliza duas estruturas de sumarização para lidar com novas distribuições no FCD, diferenciando-as de *outliers*. As estruturas nomeadas *core-micro-cluster*, referentes ao agrupamento em si, e *pontential core-micro-cluster*, distribuição de exemplos que aparenta representar *outliers*, mas podem ser um potencial *core-micro-cluster*, são mantidas e, de tempos em tempos, uma avaliação é realizada para determinar *core-micro-cluster* que deixaram de existir e *pontential core-micro-cluster* que são identificados como não-*outliers* e passam a fazer parte do modelo de agrupamento.

Forestiero, Pizzuti e Spezzano (2009) propõe uma técnica de agrupamento baseado em um sistema multi-agente que usa uma estratégia auto-organizadora para agrupar pontos semelhantes: os exemplos são associados a agentes e colocados em um espaço 2D, para aplicação de uma estratégia heurística baseada em um modelo de *flocking* (KENNEDY; EBERHART; SHI, 2001).

Embora seja uma técnica menos recente, o DenStream é visto como uma abordagem mais clássica para aprendizado em domínios FCD que gera bons resultados e, por isso, ainda serve de inspiração para outras técnicas e é utilizado para avaliações e comparações com novas técnicas. O algoritmo APDenStream (ZHANG et al., 2013) baseia-se nos métodos AP e DenStream para definição de um modelo geral que representa o FCD.

3.1.1 Aprendizado Não Supervisionado em FCD baseados em Densidade+Grid

3.1.1.1 *Clustering data streams using grid-based synopsis* (BHATNAGAR; KAUR; CHAKRAVARTHY, 2013)

Técnica: Agrupamento que utiliza estrutura de sinopse (grid-density inspired).

Resumo: Proposta do algoritmo Exclusive and Complete Clustering (ExCC) captura clusters não sobrepostos em FCDs com atributos mistos, sendo que cada ponto pertence a um cluster ou é um outlier/ruído. O algoritmo é robusto, adaptando-se a mudanças na distribuição dos dados e detectando "succint outliers" (?) *on-the-fly*. Utiliza uma estrutura fixa de grid como sinopse e realiza agrupamento verificando regiões densas no grid. speed-based pruning é aplicada à sinopse antes do agrupamento para garantir o curso de novos clusters descobertos.

Comentário: Abordagem não supervisionada, baseada em densidade e grid

3.1.1.2 *A Study of Density-Grid based Clustering Algorithms on Data Streams* (AMINI et al., 2011)

Técnica: Revisão de algoritmos de agrupamento baseados em grid que contam com conceitos de densidade (density-grid clustering algorithms).

DUCStream (GAO et al., 2005) D-Stream I (ESTER et al., 1996) DD-Stream (JIA; TAN; YONG, 2008) D-Stream II (TU; CHEN, 2009) PKS-Stream (REN; CAI; HU, 2011)

3.1.1.3 *A Clustering Algorithm Based on Density-Grid for Stream Data* (ZHANG et al., 2012)

Técnica: Proposta de um algoritmo de agrupamento baseado em densidade e grid, PKS-Stream-I, para aprendizado em FCD. Otimização do algoritmo PKS-Stream.

3.1.1.4 *Discovering Clusters with Arbitrary Shapes and Densities in Data Streams* (MAGDY; YOUSRI; EL-MAKKY, 2011)

Técnica: Proposta de um novo algoritmo de agrupamento density-grid based para aprendizado em FCD. O algoritmo utiliza um componente online para mapear os dados de entrada em células grade. Um componente offline é, então, utilizado para agrupar as grades baseado na informação de densidade.

Aplicação: KDD99

Fechamento da seção.

3.2 Técnicas de Classificação em Fluxos Contínuos de Dados

A indução de árvores de decisão é uma forma de aprendizado supervisionado amplamente utilizada e tem sido bastante explorada dentro do contexto de FCDs. Muitas das propostas envolvendo árvores de decisão utilizam ideias gerais das AH, apresentadas na Seção 2.2.2.

O algoritmo *Incremental Extremely Random Forest* (WANG et al., 2009) considera o aprendizado, feito por árvore de decisão baseada em AH, em FCDs com baixo volume de exemplos, mas em domínios onde seja necessária a adaptação do modelo geral de classificação.

A *Very Fast Decision Tree* (DOMINGOS; HULTEN, 2000), uma proposta de aprendizado incremental baseada em AH, serviu de fundamento para outros métodos. Uma das abordagens, *Concept-adapting Very Fast Decision Tree* (CVFDT) (HULTEN; SPENCER; DOMINGOS, 2001), tem como foco a detecção e adaptação a desvios de conceito em FCDs. Liu, Li e Zhong (2009) apresentam uma proposta de um mecanismo para integração de ambiguidades à CVFDT, modificando a divisão de nós pela exploração de múltiplas opções. A técnica visa garantir que o conhecimento mais novo seja utilizado na divisão dos nós, mas também é capaz de reaprender conceitos ressurgentes.

Tsai, Lee e Yang (2009) apresentam uma proposta diferenciada para mineração de regras de desvios de conceitos, buscando encontrar a regra que governa o desvio identificado. A técnica é baseada em AH e conta com estratégias para diminuir a complexidade das regras de desvio de conceito mineradas.

Dentro do conjunto de abordagens para aprendizado supervisionado em FCD também encontramos alguns métodos híbridos que se utilizam de conceitos *fuzzy*.

A proposta de Shaker, Senge e Hüllermeier (2013) é um método de classificação adaptativo baseado em *Fuzzy Pattern Trees* (FPT) (HUANG; GEDEON; NIKRAVESH, 2008), com mecanismo de adaptação que identifica a antecipação de possíveis mudanças locais no modelo atual e confirma essas mudanças por teste estatístico de hipótese.

Wang, Ji e Jin (2013) expõem a proposta de um *framework* geral para a utilização de pesos *fuzzy* para cada exemplo do FCD. De forma incremental, conforme a chegada de novos exemplos, o cálculo da pertinência, baseado na informação de rótulo, é efetuado, levando em conta as pertinências já calculados para exemplos antigos. O cálculo de distância utilizado pode identificar possíveis *outliers*. A priori, o *framework* pode ser utilizado em conjunto com qualquer algoritmo de classificação que faça uso da informação obtida na forma de pesos. Os autores optaram por vincular, de forma direta, um algoritmo baseado em redes neurais, o *Passive-Agressive* (PA)

(CRAMMER et al., 2006), constituindo a técnica neuro-*fuzzy* (JANG; SUN; MIZUTANI, 1997) chamada de *Fuzzy Passive-Agressive*.

A ideia da construção de um modelo preditivo pela integração de múltiplos modelos (WITTEN; FRANK, 2005; ROKACH, 2010) também pode ser encontrada no aprendizado em FCD. Abordagens que se utilizam deste sistema serão tratadas, doravante, pelo termo geral *Ensemble* de Classificadores (EC).

Tsymbal et al. (2008) propõe uma técnica de integração dinâmica de EC para auxiliar no trabalho com desvios de conceito, onde cada classificador recebe um peso proporcional a sua acurácia local. Para a classificação final, o melhor classificador base, quando houver, é selecionado ou é realizada uma votação ponderada entre os classificadores.

A mudança de conceitos e contexto é o foco no trabalho de Gomes, Menasalvas e Sousa (2011). A proposta baseada em EC realiza a detecção de mudanças de conceito e, dinamicamente, adiciona e remove classificadores ponderados de acordo com o que foi identificado. Conceitos estáveis são detectados por método baseado na taxa de erro do processo de aprendizado. A informação de contexto é utilizada na adaptação a conceitos recorrentes (ou ressurgentes) e no gerenciamento de conhecimento aprendido previamente.

A dificuldade de classificação de instâncias incompletas também é uma preocupação quando tratamos de aprendizado em FCD. A maior parte das abordagens assume que todos os exemplos do FCD possuem valores para um determinado conjunto de atributos, no entanto existem esforços (MILLÁN-GIRALDO; SÁNCHEZ; TRAVER, 2011) para que os exemplos sem um ou mais atributos possam ser aproveitados no processo de aprendizagem.

Fechamento da seção.

3.3 Abordagens que utilizam Técnicas Semissupervisionadas no Processo de Aprendizado

3.3.0.5 *Learning to Group Web Text Incorporating Prior Information* (CHENG et al., 2011)

Técnica: Proposta de um framework unificado capaz de incorporar informação prévia sobre pertinência de ckusters para análise de agrupamento de textos web e desevolvimento de umnovo modelo de agrupamento semissupervisionado. Este framework oferece várias vantagens sobre outras abordagens semissupervisionada existentes. A abordagem proposta é capaz de lidar com restrições entre pares de exemplos e rótulos de forma simultânea, além de permitir a obtenção de informação prévia de forma automática ou com pouco esforço humano.

Aplicação: News-500 (Google News) e fbs-500 (facebook comments)

3.3.1 Aprendizado Semissupervisionado em FCD baseado em Agrupamento

3.3.1.1 *C-DenStream: Using Domain Knowledge on a Data Stream* (RUIZ; MENASALVAS; SPILIOPOULOU, 2009)

Técnica: agrupamento semissupervisionado (informação no formato de restrições), baseado em *densidade*, para aprendizado em FCD.

Aplicação: Conjuntos reais e sintéticos.

Relevância: *To our knowledge, this is the first approach to* ***include domain knowledge in clustering for data streams***.

Resumo: 1ª (?) extensão do paradigma de aprendizado semissupervisionado (agrupamento) estático para FCD. Apresentação do C-DenStream, algoritmo de agrupamento baseado em densidade para FCD, que usa informação de domínio em formato de restrições. Proposta de novo método para utilização de conhecimento prévio em FCD. Estudo de performance em conjuntos reais e sintéticos demonstra efetividade e eficiência do método.

Comentário: Este método deve ser melhor investigado, pois utiliza uma extensão semissupervisionada (formato de restrições) para realizar agrupamento em FCD.

3.3.1.2 *Clustering evolving data stream with affinity propagation algorithm* (ATWA; LI, 2014)

Técnica: extensão semissupervisionada do agrupamento *Affinity Propagation* (FREY; DUECK, 2007) para uso em FCD.

Resumo: Proposta de um algoritmo de agrupamento semissupervisionado que extende *Affinity Propagation (AP)* para lidar com FCD. Um conjunto de instâncias rotuladas é incorporado para detecção de mudança in the generative process underlying the data stream (desvio de conceito ?), que requer a atualização do modelo o mais rápido possível. Experimentos comparativos com outros métodos de agrupamento em FCD demonstram efetividade e eficiência do método proposto.

Comentário: Este método deve ser melhor investigado, pois utiliza uma extensão semissupervisionada (formato de rótulos/sementes) para realizar agrupamento em FCD.

3.3.1.3 *CE-Stream: Evaluation-based Technique for Stream Clustering with Constraints* (SIRAMPUJ; KANGKACHIT; WAIYAMAI, 2013)

Técnica: Agrupamento semissupervisionado com informação na forma de restrições entre pares de instâncias, adaptado para aplicação em FCD

Relevância: Implementação de mecanismos para lidar com restrições dinâmicas.

Resumo: Proposta de um algoritmo para agrupamento em FCD com uso de conhecimento prévio em forma de restrições *must-link*: CE-Stream, extensão para E-stream (não supervisionado). Inclui mecanismos para lidar com restrições dinâmicas, que mudam de acordo com o tempo (constraint activation, fading and outdating). Comparedo ao E-Stream, resultados experimentais mostram que CE-Stream obtém melhor performance de agrupamento, considerando qualidade de grupos (*f-measure*) e tempo de execução.

Comentário: Esta abordagem parece interessante por utilizar semissupervisão em forma de restrições e incluir mecanismos para esquecimento, ativação e atualização dessas restrições. Trata-se de um método de agrupamento semissupervisionada em FCD

3.3.1.4 *GT2FC: An online growing interval type-2 self-learning fuzzy classifier* (BOUCHACHIA; VANARET, 2014)

Técnica: Agrupamento semissupervisionado para encontrar partições fuzzy tipo-2 por projeção dos grupos encontrados, também gerando regras fuzzy tipo-2

Aplicação: Ambient Intelligence (sensores), sintéticos e clássicos (UCI).

Resumo: Proposta de um método de aprendizado de regras fuzzy tipo-2 em FCD (Growing Type-2 Fuzzy Classifier - GT2FC). Utiliza algoritmo semissupervisionado Growing Gaussian Mixture Model para gerar as partições fuzzy tipo-2 para construir as regras fuzzy tipo-2. GT2FC concilia aprendizado por conjunto parcialmente rotulado e utiliza técnicas de otimização e seleção de atributo para manter baixa complexidade do classificador. A proposta também contato com mecanismos de online learning, e pode ser utilizada de forma offline ou online.

Comentário: Trata-se de um classificador que gera regras fuzzy tipo-2 e conta com mecanismos para realizar online learning (gerenciamento de concept drift, por exemplo). Não sei se seria possível adaptar para fuzzy tipo-1

3.3.1.5 *A Semi-supervised Incremental Clustering Algorithm for Streaming Data* (HALKIDI; SPILIOPOULOU; PAVLOU, 2012)

Técnica: Agrupamento semissupervisionado, que utiliza um stream de restrições, introduzindo o conceito de multi-clusters (regiões densas e overlapping) para um stream de restrições. Implementa técnica para identificação de outliers.

3.3.1.6 *Semi-supervised classification for reducing false positives* (HUANG; WANG; LI, 2012)

Técnica: Algoritmo EM semissupervisionado para aprendizado em FCD

3.3.1.7 *A Framework for Clustering Uncertain Data Streams* (AGGARWAL; YU, 2008)

Técnica: Proposta de método para agrupamento de FCD incertos. Utiliza-se um modelo geral de incerteza, no qual assume-se que algumas estatísticas de incerteza estão disponíveis. Utilização de informação prévia sobre incerteza pode melhorar a qualidade dos resultados do aprendizado. Utilizado em domínios onde há ruído e dados incompletos e estatísticas a respeito desse tipo de falha. Utiliza o conceito de micro-cluster, mantendo um conjunto dessas estruturas que podem ser substituídas por novos micro-clusters conforme a chegada de novos dados.

Aplicação: KDD99, KDD98, Forest Covertype

3.3.1.8 *Learnable Topical Crawler Through Online Semi-supervised Clustering* (WU; YE; FU, 2009)

Técnica: Proposta de um método de agrupamento semissupervisionado para construção de um learnable topical crawler. A proposta aplica um agrupamento k-means com restrições para

detectar novas amostras de páginas que são enviadas a um classificador de páginas e preditor de links para atualização de modelos aprendidos.

Aplicação: Web Topic Crawler

3.3.1.9 *An Incremental Affinity Propagation Algorithm and Its Applications for Text Clustering* (SHI et al., 2009)

Técnica: Proposta de um esquema semissupervisionado para utilização do agrupamento Affinity Propagation. Informação prévia é representada pelo ajuste da matriz de similaridade. Um estudo é aplicado para ampliar o conhecimento prévio.

Aplicação: benchmark data set Reuters-21578

3.3.1.10 *Semi-supervised Classification Method for Dynamic Applications* (MOUCHAWEH, 2010)

Técnica: Proposta de um método de classificação semissupervisionada baseada em fuzzy pattern matching (FPM). O objetivo é aprender funções de pertinência com um conjunto de dados inicial limitado. A função de pertinência das classes é aprendida e atualizada usando uma abordagem incremental ou recurssiva. Este método não requer informação prévia sobre a natureza ou número de classes.

3.3.1.11 *Using correlation based subspace clustering for multi-label text data classification* (AHMED; KHAN; RAJESWARI, 2010)

Técnica: Proposta de extensão de proposta anterior, abordagem SISC (Semi-supervised Impurity based Subspace Clustering) e sua variação multi-label SISC-ML. O novo algoritmo é chamado de H-SISC (Hierarchical SISC). H-SISC captura a correlação implícita que existe entre cada par de rótulos de classes em um ambiente multi-label, desenvolvendo um classificador multi-label robusto.

3.3.1.12 *Application of Compound Gaussian Mixture Model clustering in the data stream* (GAO; LIU; GAO, 2010)

Técnica: Proposta do Compound Gaussian Mixture Model (CGMM), baseado no agrupamento GMM, utilizando amostra de dados rotulados para melhorar o agrupamento.

3.3.1.13 *Incremental Semi-supervised Clustering In A Data Stream With A Flock Of Agents* (BRUNEAU; PICAROUGNE; GELGON, 2009)

Técnica: Proposta de adaptação de algoritmo de agrupamento bio-inspired para aprendizado em FCD, que cria e visualiza grupos de dados de forma dinâmica. Introdução de um operador de mescla para a sumarização de um grupo de dados e um operador de divisão que usa informação de um pequeno conjunto de dados rotulados e permite a adaptação do agrupamento para mudanças no FCD.

3.3.2 Aprendizado Semissupervisionado em FCD baseado em Classificador

3.3.2.1 *Active Learning Framework Combining Semi-Supervised Approach for Data Stream Mining* (KHOLGHI; KEYVANPOUR, 2011)

Técnica: active learning (seleção de instâncias "importantes" para rotulação) + influência de instâncias não rotuladas para aprendizado semissupervisionado em FCD. $\longrightarrow$ não está muito claro como ocorre a integração pelos trechos disponíveis do artigo, então talvez seja interessante ler outros trechos do artigo.

Aplicação: conjuntos reais.

Relevância: *To the best of our knowledge, no* ***semi-supervised Active Learning*** *combination exists for* ***data streams***.

Resumo: Construir um modelo para predição de rótulos de instâncias futuras com grande acurácia. Proposta de um framework que combina semissupervisão por meio de active learning e consideração de influência de instâncias não rotuladas a fim de melhorar a performance de aprendizagem. Visa solucionar os "problemas" de active learning (desconsideração de influência de instâncias não rotuladas) e desafios de aprendizado em FCD. Experimentos provam efetividade do framework proposto.

Comentário: Trata-se de um método de classificação adaptado para semissupervisão (influência de dados não rotulados). Aparentemente, não é uma adaptação "fake" como em 3.3.5.2, mas é uma solução baseada em classificadores, não agrupamento.

3.3.2.2 *Mining Recurring Concept Drifts with Limited Labeled Streaming Data* (LI; WU; HU, 2012)

Técnica: Árvore de Hoeffding, utilizando agrupamento k-means para split de folhas e rotulação automática de exemplos.

Resumo: Extensão de *Mining Recurring Concept Drifts With Limited Labeled Streaming Data* (LI; WU; HU, 2010b). O trabalho propõe um algoritmo de classificação semissupervisionada em FCD, chamado REDLLA, utilizando uma árvore de decisão como modelo de classificação. Para o crescimento da árvore, utiliza-se de um algoritmo de agrupamento baseado no *k-means* para a produção de "concept clusters" e rotulação automática de dados não rotulados. Potenciais desvios de conceito são identificados conceitos recorrentes são mantidos. Estudos mostram as vantagens da proposta com relação a outros algoritmos de classificação em FCD e algoritmos semissupervisionados em FCD, para até 90% de dados não rotulados.

Comentário: Esta proposta utiliza Hoeffding Tree, que já foi considerada uma abordagem potencialmente interessante, pois possui algumas vantagens das árvores de decisão produzidas por algoritmos estáticos.

3.3.2.3 *On Achieving Semi-supervised Pattern Recognition By Utilizing Tree-based SOMs* (ASTUDILLO; OOMMEN, 2013)

Técnica: Proposta semissupervisionada baseada em SOM

Resumo: Proposta de um algoritmo para classificação semisssupervisionada de padrões por Topology Oriented SOM baseada em árvore (TTOSOM) (ASTUDILLO; John Oommen, 2011). Primeiramente realiza o treinamento de uma TTOSOM em que os neurônios obedecem a distribuição das classes. Em seguida, a informação de rótulo do conjunto é utilizada, atribuindo rótulos a cada nó da Rede Neural, que, por sua vez, particiona o espaço em regiões Voroni (?). Com a chegada de dados de teste, o exemplo recebe a classe correspondente ao neurônio mais próximo.

Comentário: Proposta baseada em Redes Neurais

3.3.2.4 *Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition* (KASABOV et al., 2013)

Técnica: ??

Aplicação: ??

Resumo: Proposta de evolving spiking neural networks dinâmica com mecanismos para aprendizado não supervisionado, supervisionado e semissupervisionado.

Comentário: Um grande mistério esta proposta... Só fisguei que usa redes neurais.

3.3.2.5 *A comparison of two purity-based algorithms when applied to semi-supervised streaming data classification* (BERTINI; ZHAO, 2013)

Técnica: Classificação com algoritmo baseado em grafo, utilizando mecanismos de semissupervisão também baseados em grafos

Resumo: Proposta de algoritmo que realiza aprendizado semissupervisionado considerando distribuição não estacionária de dados, o KAOGINCSSL. Análise do algoritmo considerando duas diferentes estratégias de propagação de rótulos para treinamento de classificadores. A primeira consite em aplicar o algoritmo indutivo KAOGSS para construir um classificador e a segunda utiliza o algoritmo de transdução PMTLA para propagar rótulos previamente à construção do classificador. Resultados quanto acurácia e tempo de processamento de ambos os algoritmos são apresentados para problemas não estacionários.

Comentário: Classificação baseada em grafo com a opção de utilização de dois mecanismos distintos para incorporar semissupervisão.

3.3.2.6 *Semi-supervised learning with concept drift using particle dynamics applied to network intrusion detection data* (BREVE; ZHAO, 2013)

Técnica: extensão de método semissupervisionado baseado em competição de partículas para lidar com FCD e desvio de conceito

Aplicação: KDD Cup 1999 Data of network intrusion

Resumo: Proposta de extensão de método semissupervisionado baseado em competição de partículas para lidar com FCD e desvio de conceito. O algoritmo de classificação é naturalmente adaptado para mudanças de conceitos sem um mecanismo explícito para detecção de desvio de conceito. Possui mecanismo de esquecimento.

Comentário: classificação por competição de partículas???

3.3.2.7 *On-line laplacian one-class support vector machines* (FRANDINA et al., 2013)

Técnica: One-Class SVM estendido para lidar com FCD, semissupervisionado.

Resumo: Proposta do algoritmo On-line Laplacian One-Class SVM, que considera exemplos rotulados positivamente e não rotulados para atualização do classificador. Utiliza buffer para lidar com FCD.

Comentário: Classificação por SVM

3.3.2.8 *Particle competition and cooperation in networks for semi-supervised learning with concept drift* (BREVE; ZHAO, 2012)

Técnica: Extensão de proposta usando competição e cooperação entre partículas para realizar aprendizado semissupervisionado baseado em grafos em FCD. O classificador modelado se adapta naturalmente a modificações de conceito, sem qualquer mecanismo explícito de detecção de desvio de conceito. Conta com mecanismo de esquecimento.

3.3.2.9 *Partially labeled data stream classification with the semi-supervised K-associated graph* (BERTINI; LOPES; ZHAO, 2012)

Técnica: Proposta de uma abordagem semissupervisionada baseada em grafo para extensão do classificador K-associated Optimal Graph para realizar aprendizado em FCDs. A adaptação consiste no "espalhamento" dos rótulos no conjunto de treinamento.

3.3.2.10 *Learning very fast decision tree from uncertain data streams with positive and unlabeled samples* (LIANG et al., 2012)

Técnica: Proposta do algoritmo puuCVFDT (CVFDT for positive and unlabeled uncertain data), baseado na árvore de decisão de rápida adaptação de conceito (CVFDT), para realizar aprendizado semissupervisionado em FCD.

3.3.2.11 *Online co-localization in indoor wireless networks by dimension reduction* (PAN; YANG; PAN, 2007)

Técnica: Proposta de um método semissupervisionado para aprendizado em FCD, baseado em técnicas manifold.

Aplicação: online co-localization (recuperação de coais de dispositivos móveis e pontos de acesso a partir de sinais de rádio)

3.3.2.12 *Applying lazy learning algorithms to tackle concept drift in spam filtering* (FDEZ-RIVEROLA et al., 2007)

Técnica: Proposta de duas novas técnicas para identificar desvio de conceito no modelo SpamHunting (proposto em trabalho anterior), sistema para rotulação e filtragem automática de spam.

Aplicação: Detecção de spam

3.3.2.13 *Mining data streams with labeled and unlabeled training examples* (ZHANG; ZHU; GUO, 2009)

Técnica: Proposta de um framework para construção de modelos de predição a partir de FCD com exemplos rotulados e não rotulados. Para a construção do modelo, os dados do stream são associados a 4 categorias distintas, cada qual correspondendo à situação do desvio de conceito podendo existir ou não nos dados rotulados e não rotulados. Em seguida, é utilizado um framework de aprendizado transfer semi-supervised SVM baseado em relational k-means.

Aplicação: http://www.cse.fau.edu/ xqzhu/stream.html

3.3.2.14 *Semi Supervised Multi Kernel (SeSMiK) graph embedding: Identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy* (TIWARI; KURHANEWICZ, 2010)

Técnica: Apresentação de um framework unificado de fusão de dados, Semi Supervised Multi Kernel Graph Embedding (SeSMiK- GE). O método permite a representação de modalidades individuais de dados por um framework de multi-kernel combinados, seguido de uma redução semissupervisionado de dimensionalidade, onde é utilizada a informação de rótulos parciais.

3.3.2.15 *Evolving granular neural network for semi-supervised data stream classification* (LEITE; COSTA; GOMIDE, 2010)

Técnica: Introdução de um framework neuro-fuzzy para a classificação em FCD usando um algoritmo de aprendizado parcialmente supervisionado. O framework consiste de um a rede neural granular evolutiva capaz de processar FCD não estacionários usando um algoritmo incremental "one-pass". A rede neural granular gera fuzzy hyperboxes e usa neurônios nullnorm para classificar os dados. O algoritmo realizada adaptações estruturais e paramétricas.

3.3.2.16 *Classifying evolving data streams with partially labeled data* (BORCHANI; nAGA; BIELZA, 2011)

Técnica: Proposta de uma nova abordagem (genérica) semissupervisionada para lidar com FCD com desvios de conceito e dados parcialmente rotulados. Realiza monitoração de três tipos de drift: feature, conditional ou dual drift. Se desvio é detectado, um novo classificador aprende por meio dos dados mais recentes, usando EM.

Aplicação: rotating hyperplane, mushroom e malware detection

3.3.2.17 *Semi-supervised classification of characterized patterns for demand forecasting using smart electricity meters* (SILVA et al., 2011)

Técnica: Proposta de extensão de um abordagem de aprendizado incremental para caracterização de padrões em um ambiente de FCD de "medidores espertos", com a incoporação de classificação semissupervisionada. O algoritmo incremental pattern characterization learning (IPCL) aprende de forma autônoma a partir de "medidores espertos" e acumula padrões em uma estrutura de coluna. A introdução de classificação semissupervisionada melhora a qualidade e usabilidade do modelo de aprendizado.

Aplicação: dados de "medidores espertos" de energia

3.3.3 Aprendizado Semissupervisionado em FCD baseado em Agrupamento e Classificação

3.3.3.1 *Concurrent Semi-supervised Learning with Active Learning of Data Streams* (NGUYEN; NG; WOON, 2013)

Técnica: Atividades de aprendizado concorrente (agrupamento e classificação se complementam) em FCD

Aplicação: Conjuntos de dados sintéticos e reais

Resumo: Extensão de *Concurrent Semi-supervised Learning of Data Stremas* (NGUYEN et al., 2011). É investigado o potencial de aprendizado semissupervisionado concorrente e proposta de um algoritmo incremental chamado CSL-Stream (Concurrent Semi-supervised Learning of Data Streams) que realiza agrupamento e classificação ao mesmo tempo. Experimentos mostram que CSL-Stream tem melhor performance que algoritmos de agrupamento e classificação (D-Stream and SmSCluster) quanto a acurácia, velocidade e escalabilidade. Uma técnica melhorada de aprendizado ativo faz com que o CSL-Stream funcione bem com conjunto de

dados parcialmente rotulados.

Comentário: Abordagem concorrente com mecanismo de semissupervisão. Trabalha com agrupamento e classificação de forma que o resultado de um complementa o outro

3.3.3.2 *Learning From Concept Drifting Data Streams with Unlabeled Data* (WU; LI; HU, 2012)

Técnica: Proposta de um algoritmo de classificação semissupervisionada para FCDs com desvios de conceito e dados não rotulados (SUN). É utilizado agrupamento baseado no *k-modes* para produzir grupos de conceito (e rotulação) em folhas de uma árvore de decisão incremental. Possui mecanismo de distinção entre novos conceitos e ruído. Este artigo é uma extensão de *Learning from concept drifting data streams with unlabeled data* (LI; WU; HU, 2010a).

3.3.3.3 *Clustering Feature Decision Trees for Semi-supervised Classification from High-speed Data Streams* (XU; QIN; CHANG, 2011)

Técnica: Proposta de construção de um modelo de árvore de decisão de clustering feature, a partir de FCD parcialmente rotulado. CFDT aplica um algoritmo de micro-clustering que acessa os dados apenas uma vez para prover sumários estatísticos do dados para indução de uma árvore de decisão. Micro-clusters também são utilizados para classificação em folhas de árvore para melhorar a acurácia de classificação.

Aplicação: GAUSS, hyperplane, random RBF, random Tree, SEA, Waveform, KDD99, forest covertype

3.3.4 Aprendizado Semissupervisionado em FCD baseado em *Ensemble* de Classificadores

3.3.4.1 *Semi-supervised ensemble learning of data streams in the presence of concept drift* Ahmadi2012 $\longrightarrow$ SPRINGER

Técnica: aprendizado ensemble para rotulação de instâncias não rotuladas e posterior atualização dos modelos.

Aplicação: ???

Resumo: Apresentação de um novo algoritmo de aprendizado ensemble (que traduz para combinação? comitê?) semissupervisionado em FCD. "Voto da maioria" para rotular instâncias não rotuladas. Estudos demonstram que o algoritmo proposto é comparável a outros algoritmos

de aprendizado semissupervisionado online (em FCD).

Comentário: Esta pesquisa não prioriza métodos ensemble classifier.

3.3.4.2 *A new semi-supervised learning based ensemble classifier for recurring data stream* Zhang2014a ⟶ SPRINGER

Técnica: Classificador ensemble, com algoritmos base semissupervisionados.

Aplicação: ???

Resumo: Proposta de um classificador ensemble para aprendizado semissupervisionado para solução do problema de desvio de conceito em FCD. Algoritmos base utilizam instâncias rotuladas e não rotuladas como conjunto de treinamento para obter melhor aprendizado, informação histórica é mantida como parte de peso no fator de decisão quando construído o classificador ensemble. Nova abordagem melhor que o "modelo ensemble geral" e pode ser utilizado em FCD.

Comentário: Esta pesquisa não prioriza métodos ensemble classifier.

3.3.4.3 *Detecting cyberbullying in social networks using multi-agent system* Nahar2014a ⟶ IOS PRESS

Técnica: ensemble classifier que considera instâncias não rotuladas.

Aplicação: Texto (Detecção de cyberbulling em redes sociais.

Resumo: Proposta de um framework baseado em sessão para detecção automática de cyberbulling dentro do grande volume de FC de texto não rotulado. Com a incorporação de um classificador ensemble de uma classe ao framework. O processamento do FCD é feito em ambiente distribuído multi-agente para processar multiplas fontes de redes sociais. Apenas algumas instâncias positivas de cyberbulling estão disponíveis para treinamento inicial. Contribuição maior é detecção de cyberbulling quando não há rótulos disponíveis. Experimentos indicam que a proposta obtém melhor resultados que outros métodos.

Comentário: Esta pesquisa não prioriza métodos ensemble classifier.

3.3.4.4 *A semi-supervised ensemble approach for mining data streams* (LIU et al., 2013)

Técnica: Modelo semissupervisionado que utiliza classificador ensemble e modelos não supervisionados.

Resumo: Proposta de uma abordagem semissupervisionada baseada em ensemble para mineração em FCD. O classificador é treinando utilizando apenas dados rotulados. Novos exemplos rotulados são utilizados para atualização do classificador, enquanto dados não rotulados são utilizados para construir modelos não supervisionados. As classes são previstas por um modelo semissupervisionado que consiste do classificador e dos modelos não supervisionados de forma a maximizar o consenso, permitindo que melhor performance seja alcançada pela aplicação de restrições obtidas pelos modelos não supervisionados com um número limitado de exemplos rotulados.

Comentário: A proposta parte da construção de modelos supervisionado e não supervisionado, aplicados conjuntamente para obter melhor performance considerando restrições ao classificador obtidas pelos modelos não supervisionados.

3.3.4.5 *Semi-supervised Data Stream Ensemble Classifiers Algorithm Based On Cluster Assumption* (XUEJUN, 2012)

Técnica: Classificador Ensemble semissupervisionado baseado em cluster assumption para aprendizado em FCM.

3.3.4.6 *A framework for application-driven classification of data streams* (ZHANG et al., 2012)

Técnica: Proposta de framework para categorização inicial dos exemplos de treinamento em quatro tipos, atribuindo prioridade de aprendizado. Para cada um dos quatro tipos de dados é aplicado um classificador baseado em SVM: classical SVM, semi-supervised SVM, transfer semi-supervised SVM, and relational k-means transfer semi-supervised SVM.

Aplicação: wireless sensor stream, power supply stream and intrusion detection stream (no mention of sources)

3.3.4.7 *Facing the Reality of Data Stream Classification: Coping with Scarcity of Labeled Data* (MASUD et al., 2012)

Técnica: Cada modelo de classificação é construindo como uma coleção de micro-clusters usando agrupamento semissupervisionado e um ensemble desses modelos é utilizado para classificar dados não rotulados.

3.3.4.8 *A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data* (MASUD et al., 2008a)

Técnica: Proposta de técnica semissupervisionada baseada na construção de micro-clusters a partir de agrupamento semissupervisionado e classificação realizada pelo algoritmo KNN. Para cada chunk de dados, um modelo de classificação é construído, os L melhores modelos (de acordo com acurácia individual) são utilizados em um ensemble. Maior detalhamento no Tech Report (MASUD et al., 2008b).

3.3.4.9 *Learning from Testing Data: A New View of Incremental Semi-Supervised Learning* (CAO; HE, 2008)

Técnica: Proposta de um algoritmo iterativo que pode recuperar, de forma adaptativa, os rótulos dos dados de treinamento, de acordo com os níveis de confiança, e, em seguida, estender a população de treinamento por esses dados recuperados. O objetivo é melhorar o sistema aprendido offline por meio de aprendizado semissupervisionado incremental. Gera vários modelos de classificação a partir dos chunks, utilizando um ensemble destes modelos.

3.3.4.10 *Semi-supervised learning in nonstationary environments* (DITZLER; POLIKAR, 2011)

Técnica: Proposta de um ensemble de classificadores baseado em abordagem semissupervisionada que tira aproveito de dados rotulados e não rotulados, visando a detecção de desvios de conceito. Os dados rotulados produzem classificadores, cujos pesos de votação são determinados pela distância entre componentes de um Gaussuiam Mixture Model (GMM) treinado com o conjunto completo de dados.

3.3.4.11 Boosting Semissupervisionado

Online detection of concept drift in visual tracking (LIU; ZHOU, 2014)

Técnica: detecção de desvio de conceito com método de boosting semissupervisionado

Aplicação: Visual Tracking

Resumo: Apresentação de um framework que combina detecção de desvio de conceito com método de boosting semissupervisionado para construir um visual tracker.

Comentário: Esta pesquisa atualmente prioriza métodos aplicados a mineração de dados ou texto, devido peculiaridades de trabalho em outros tipos de domínios. Ainda, trata-se de

uma proposta que utiliza mecanismo de boosting semissupervisionado, não priorizada neste momento.

***Semi-supervised learning techniques in artificial olfaction: A novel approach to classification problems and drift counteraction* (VITO et al., 2012)**

Técnica: Mecanismo de boosting semissupervisionado para melhorar acurácia de algoritmos de classificação supervisionada. Bem orientado à aplicação: Artificial Olfaction.

***A semi-supervised boosting algorithm for mining time-changing data streams* (HUANG; SHA; MA, 2011)**

Técnica: Proposta de um algoritmo semissupervisionada para aprendizado em FCDs parcialmente rotulados. Para monitar mudanças de conceito, é utilizado um método que não necessita de rótulos. Assim que ocorre a detecção de desvio, um algoritmo de boosting semissupervisionado é utilizado para ajustar o modelo ultrapassado.

Aplicação: hyperplane, Adult (UCI)

Ao final desta seção há um pagebreak, que deverá ser removido.

3.3.5 Outras Técnicas de Aprendizado Semissupervisionado em FCD

3.3.5.1 *Semi-supervised approach to handle sudden concept drift in Enron data* (KMIECIAK; STEFANOWSKI, 2011)

Técnica: Proposta de abordagem que detecta desvios de conceito em dados não rotulados e retreina o classificador usando um número limitado de exemplos rotulados.

Aplicação: Enron corpora (e-mails)

3.3.5.2 Semissupervisão em etapa estática do aprendizado

coloquei assim apenas para organizar, não será uma subseção necessariamente...

Anomaly Intrusion Detection for Evolving Data Stream Based on Semi-supervised Learning **(YU et al., 2009)**

Técnica: agrupamento semissupervisionado para *estender o conjunto de dados rotulado* e *aprendizado supervisionado em FCD* (?), com mecanismo de "esquecimento" de dados mais antigos.

Aplicação: *Sistemas de Detecção de Intrusos* em redes de computadores.

Resumo: Proposta de um algoritmo baseado em aprendizado semissupervisionado para detecção de anomalias em FCD, dentro do contexto de segurança em redes de computadores. O algoritmo se utiliza de técnicas de atenuação para resolver o problema de mudança no padrão de tráfego, diminuindo a importância dos dados mais antigos, e um conjunto rotulado estendido, gerado por aprendizado semissupervisionado, para treinar o modelo de detecção. O algoritmo manifesta resultados experimentais de melhor acurácia comparado a algoritmos baseados em histórico completo e totalmente rotulado.

Comentário: A semissupervisão, aparentemente, é utilizada como parte de rotulação automática e estática para extensão de conjunto de treinamento para construção supervisionada de modelo em FCD. O objetivo desta pesquisa é encontrar métodos que incorporem semissupervisão no processo de aprendizagem em FCD de forma mais intrínseca.

Semi-supervised learning for cyberbullying detection in social networks **(NAHAR et al., 2014)**

Técnica: ensemble classifier supervisionado + abordagem semissupervisionada para aumentar conjunto de treinamento rotulado.

Aplicação: Detecção de cyberbulling em redes sociais

Resumo: Proposta de uma abordagem semissupervisionada para aumentar a quantidade de instâncias de treinamento e aplicação de um algoritmo de SVM fuzzy. Um conjunto de treinamento rotulado inicial é utilizado para rotulação automática de novas instâncias do stream, por meio de um ensemble classifier. O algorithm de SVM fuzzy é utilizado para ponderar o espaço de atributos. A avaliação mostra a superioridade da proposta em diferentes cenários.

Comentário: ensemble classifiers não são foco da pesquisa e a proposta não utiliza semissupervisão no processo de aprendizagem

Compose: A semisupervised learning framework for initially labeled nonstationary streaming data (DYER; CAPO; POLIKAR, 2014)

Técnica: Aprendizado SS para aumentar o conjunto de dados rotulados + aprendizado supervisionado

Aplicação: conjuntos de dados reais (National Oceanic and Atmospheric Administration weather) e sintéticos (carefully designed)

Resumo: Proposta do framework COMPOSE (extração compacta de amostra de objeto), utilizado para aprender desvios de conceitos em ambiente de FCD, onde há apenas um conjunto inicial de dados rotulados e, após a inicialização, apenas dados não rotulados. COMPOSE segue 3 passos: 1) combinação dos dados rotulados iniciais aos dados não rotulados atuais para treinar um classificador semissupervisionado e rotular de forma automática o conjunto de dados; 2) para cada classe, construção de formas que "embrulham" os dados, representando a distribuição atual da classe; 3) compactação das formas e extração de instâncias representantes (core supports), que servirão como conjunto rotulado inicial para os próximos novos dados não rotulados. A performance da proposta é comparada com três outras propostas, sendo duas supervisionadas e com acesso a todos os rótulos. A execução em conjunto de dados reais demonstra que a proposta é competitiva até mesmo contra método supervisionado que recebe apenas dados rotulados.

Comentário: A proposta inclui uma etapa estática de aprendizado semissupervisionado para extensão do conjunto de dados rotulados. Pela descrição inicial, o fluxo dos experimentos possui apenas exemplos não rotulados, mas o framework poderia ser utilizado com fluxo parcialmente rotulado.

***Clustering-training for data stream mining* (WU; YANG; ZHOU, 2006)**

Técnica: Proposta de um algoritmo de aprendizado semissupervisionado baseado em treinamento por agrupamento, para seleção de exemplos confiáveis a serem utilizados no re-treinamento de um classificador.

3.3.5.3 Com supervisão de especialista

***A stream-based semi-supervised active learning approach for document classification* (BOUGUELIA; BELAID; BELAID, 2013)**

Técnica: adaptive incremental neural gas algorithm (AING) para aprendizado em FCD, gerando um classificador a partir de especialista rotulador

Aplicação: FCD de documentos não rotulados

Resumo: Proposta de extensão do adaptive incremental neural gas algorithm (AING) com mecanismos de semissupervisão e para aprendizado em FCD, para classificação de documentos. Baseado em uma métrica de incerteza, o classificador ativamente pede (a um humano) rótulos de documentos que são mais informativos para o aprendizado. Um modelo é mantido na forma de topologia de grafo dinamicamente evolutivo dos documentos rotulados, chamados de neurônios.

Comentário: Realiza aprendizado semissupervisionado com etapa de supervisão de especialista para rotulação de documentos mais relevantes para o processo de aprendizagem.

Ao final desta seção há um pagebreak, que deverá ser removido.

3.4 Considerações Finais

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.

AGGARWAL, C. C.; YU, P. S. A Framework for Clustering Uncertain Data Streams. In: *2008 IEEE 24th International Conference on Data Engineering*. [S.l.]: IEEE, 2008. v. 00, p. 150–159.

AHMED, M. S.; KHAN, L.; RAJESWARI, M. Using Correlation Based Subspace Clustering for Multi-label Text Data Classification. In: *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*. [S.l.]: IEEE, 2010. p. 296–303.

AMINI, A.; WAH, T. Y.; SABOOHI, H. On Density-Based Data Streams Clustering Algorithms: A Survey. *Journal of Computer Science and Technology*, v. 29, n. 1, p. 116–141, 2014.

AMINI, A. et al. A study of density-grid based clustering algorithms on data streams. In: *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. [S.l.]: IEEE, 2011. p. 1652–1656.

ANKERST, M. et al. Optics: Ordering points to identify the clustering structure. In: *ACM Sigmod Record*. [S.l.: s.n.], 1999. p. 49–60.

ASTUDILLO, C. A.; John Oommen, B. Imposing tree-based topologies onto self organizing maps. *Information Sciences*, v. 181, n. 18, p. 3798–3815, 2011.

ASTUDILLO, C. A.; OOMMEN, B. J. On achieving semi-supervised pattern recognition by utilizing tree-based SOMs. *Pattern Recognition*, v. 46, n. 1, p. 293–304, 2013.

ATWA, W.; LI, K. Clustering Evolving Data Stream with Affinity. In: *Database and Expert Systems Applications*. [S.l.]: Springer International Publishing, 2014. p. 446–453.

BABCOCK, B. et al. Models and issues in data stream systems. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02*. New York, New York, USA: ACM Press, 2002. p. 1–16.

BERTINI, J. a. R.; LOPES, A. D. A.; ZHAO, L. Partially labeled data stream classification with the semi-supervised K-associated graph. *Journal of the Brazilian Computer Society*, v. 18, n. 4, p. 299–310, 2012.

BERTINI, J. R.; ZHAO, L. A Comparison of Two Purity-Based Algorithms When Applied to Semi-supervised Streaming Data Classification. In: *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*. [S.l.]: IEEE, 2013. p. 21–27.

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

BHATNAGAR, V. et al. Data Understanding using Semi-Supervised Clustering. In: *2012 Conference on Intelligent Data Understanding*. [S.l.]: IEEE, 2012. p. 118–123.

BHATNAGAR, V.; KAUR, S.; CHAKRAVARTHY, S. Clustering data streams using grid-based synopsis. *Knowledge and Information Systems*, v. 41, n. 1, p. 127–152, jun. 2013.

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

BOARD, R.; PITT, L. Semi-supervised learning. *Machine Learning*, Kluwer Academic Publishers, v. 4, n. 1, p. 41–65, 1989.

BORCHANI, H.; nAGA, P. L.; BIELZA, C. Classifying evolving data streams with partially labeled data. *Intelligent Data Analysis*, v. 15, n. 5, p. 655–670, 2011.

BOUCHACHIA, A.; VANARET, C. GT2FC: An Online Growing Interval Type-2 Self-Learning Fuzzy Classifier. *IEEE Transactions on Fuzzy Systems*, v. 22, n. 4, p. 999–1018, ago. 2014.

BOUGUELIA, M.-R.; BELAID, Y.; BELAID, A. A Stream-Based Semi-supervised Active Learning Approach for Document Classification. In: *2013 12th International Conference on Document Analysis and Recognition*. [S.l.]: IEEE, 2013. p. 611–615.

BREVE, F.; ZHAO, L. Particle competition and cooperation in networks for semi-supervised learning with concept drift. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. [S.l.]: IEEE, 2012. p. 1–6.

BREVE, F.; ZHAO, L. Semi-supervised Learning with Concept Drift Using Particle Dynamics Applied to Network Intrusion Detection Data. In: *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*. [S.l.]: IEEE, 2013. p. 335–340.

BRUNEAU, P.; PICAROUGNE, F.; GELGON, M. Incremental semi-supervised clustering in a data stream with a flock of agents. In: *2009 IEEE Congress on Evolutionary Computation*. [S.l.]: IEEE, 2009. p. 3067–3074.

CAO, F. et al. Density-Based Clustering over an Evolving Data Stream with Noise. In: *Proceedings of the 6th SIAM International Conference on Data Mining*. [S.l.: s.n.], 2006. p. 328–339.

CAO, Y.; HE, H. Learning from testing data: A new view of incremental semi-supervised learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. [S.l.]: IEEE, 2008. p. 2872–2878.

CASAS, P.; MAZEL, J.; OWEZARSKI, P. MINETRAC: Mining flows for unsupervised analysis & semi-supervised classification. In: *Proceedings of the 23rd International Teletraffic Congress*. [S.l.: s.n.], 2011. p. 87–94.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. *Semi-Supervised Learning*. [S.l.]: MIT Press, 2006. 523 p.

CHEN, J.; CHEN, P.; SHENG, X. A Sketch-based Clustering Algorithm for Uncertain Data Streams. *Journal of Networks*, v. 8, n. 7, p. 1536–1542, jul. 2013.

CHENG, Y. et al. Learning to Group Web Text Incorporating Prior Information. In: *2011 IEEE 11th International Conference on Data Mining Workshops*. [S.l.]: IEEE, 2011. p. 212–219.

CINTRA, M. E.; MONARD, M. C.; CAMARGO, H. A. FuzzyDT - A Fuzzy Decision Tree Algorithm Based on C4. 5. In: *CBSF - Brazilian Congress on Fuzzy Systems*. [S.l.: s.n.], 2012. p. 199–211.

CORDÓN, O. A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *International Journal of Approximate Reasoning*, v. 52, n. 6, p. 894–913, set. 2011.

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

CRAMMER, K. et al. Online Passive-Aggressive Algorithms. *The Journal of Machine Learning Research2*, v. 7, p. 551–585, 2006.

DAY, W. H. E.; EDELSBRUNNER, H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, v. 1, n. 1, p. 7–24, 1984.

DITZLER, G.; POLIKAR, R. Semi-supervised learning in nonstationary environments. In: *The 2011 International Joint Conference on Neural Networks*. [S.l.]: IEEE, 2011. p. 2741–2748.

DOMINGOS, P.; HULTEN, G. Mining high-speed data streams. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00*. [S.l.: s.n.], 2000. p. 71–80.

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

DYER, K. B.; CAPO, R.; POLIKAR, R. COMPOSE: A semisupervised learning framework for initially labeled nonstationary streaming data. *IEEE transactions on neural networks and learning systems*, v. 25, n. 1, p. 12–26, jan. 2014.

ESTER, M. et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Second International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 1996. p. 226–231.

FAZZOLARI, M. et al. A Review of the Application of Multiobjective Evolutionary Fuzzy Systems : Current Status and Further Directions. *Fuzzy Systems, IEEE Transactions on*, v. 21, n. 1, p. 45–65, 2013.

FDEZ-RIVEROLA, F. et al. Applying lazy learning algorithms to tackle concept drift in spam filtering. *Expert Systems with Applications*, v. 33, n. 1, p. 36–48, jul. 2007.

FORESTIERO, A.; PIZZUTI, C.; SPEZZANO, G. FlockStream: A Bio-Inspired Algorithm for Clustering Evolving Data Streams. In: *2009 21st IEEE International Conference on Tools with Artificial Intelligence*. [S.l.]: IEEE, 2009. p. 1–8.

FRANDINA, S. et al. On-Line Laplacian One-Class Support Vector Machines. In: *Artificial Neural Networks and Machine Learning (ICANN2013)*. [S.l.]: Springer Berlin Heidelberg, 2013. p. 186–193.

FREY, B. J.; DUECK, D. Clustering by Passing Messages Between Data Points. *Science*, v. 315, n. 5814, p. 947–949, fev. 2007.

GABER, M. M.; ZASLAVSKY, A.; KRISHNASWAMY, S. Mining data streams: a review. *ACM SIGMOD Record*, v. 34, n. 2, p. 18, 2005.

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

GAO, J. et al. An incremental data stream clustering algorithm based on dense units detection. *Advances in Knowledge Discovery and Data Mining*, v. 3518, p. 420–425, 2005.

GAO, M. M.; LIU, J. Z.; GAO, X. X. Application of Compound Gaussian Mixture Model clustering in the data stream. In: *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*. [S.l.]: IEEE, 2010. p. V7–172–V7–177.

GOLDBERG, D. *Genetic Algorithms in Search, Optimization, and Machine Learning*. [S.l.]: Addison-Wesley, 1989. 432 p.

GOMES, J. a. B.; MENASALVAS, E.; SOUSA, P. a. C. Learning recurring concepts from data streams with a context-aware ensemble. In: *Proceedings of the 2011 ACM Symposium on Applied Computing - SAC '11*. New York, New York, USA: ACM Press, 2011. p. 994.

GUSTAFSON, D. E. G. D. E.; KESSEL, W. C. K. W. C. Fuzzy clustering with a fuzzy covariance matrix. *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, v. 17, n. 2, p. 761–766, 1978.

HALKIDI, M.; SPILIOPOULOU, M.; PAVLOU, A. A semi-supervised incremental clustering algorithm for streaming data. *Advances in Knowledge Discovery and Data Mining*, v. 7301, p. 578–590, 2012.

HAMASUNA, Y.; ENDO, Y. On semi-supervised fuzzy c-means clustering with clusterwise tolerance by opposite criteria. In: *2011 IEEE International Conference on Granular Computing*. [S.l.]: IEEE, 2011. p. 225–230.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann Publishers, 2012. 744 p. (Data Management Systems Series).

HAVENS, T. C. et al. Fuzzy c-Means Algorithms for Very Large Data. *IEEE Transactions on Fuzzy Systems*, v. 20, n. 6, p. 1130–1146, dez. 2012.

HINNEBURG, A.; KEIM, D. A. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*. [S.l.: s.n.], 1998. v. 5865, p. 58–65.

HORE, P. et al. Online fuzzy c means. In: *NAFIPS 2008 - 2008 Annual Meeting of the North American Fuzzy Information Processing Society*. [S.l.]: IEEE, 2008. p. 1–5.

HORE, P.; HALL, L. O.; GOLDGOF, D. B. A fuzzy c means variant for clustering evolving data streams. In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. [S.l.]: IEEE, 2007. p. 360–365.

HORE, P.; HALL, L. O.; GOLDGOF, D. B. Creating Streaming Iterative Soft Clustering Algorithms. In: *NAFIPS 2007 - 2007 Annual Meeting of the North American Fuzzy Information Processing Society*. [S.l.]: IEEE, 2007. p. 484–488.

HUANG, S.; SHA, A.; MA, S. A Semi-Supervised Boosting Algorithm for Mining Time-Changing Data Streams. *Journal of Information & Computational Science*, v. 13, p. 2807–2814, 2011.

HUANG, S.; WANG, K.; LI, T. Semi-supervised Classification for Reducing False Positives. *Journal of Computational Information Systems*, v. 13, n. 8, p. 5327–5334, 2012.

HUANG, Z.; GEDEON, T. D.; NIKRAVESH, M. Pattern trees induction: A new machine learning method. *IEEE Transactions on Fuzzy Systems*, v. 16, n. 4, p. 958–970, 2008.

HULTEN, G.; SPENCER, L.; DOMINGOS, P. Mining time-changing data streams. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2001. p. 97–106.

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, 1999.

JANG, J.-S. R.; SUN, C.-T.; MIZUTANI, E. *Neuro-Fuzzy and Soft Computing*. [S.l.: s.n.], 1997. 614 p.

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, 1998.

JIA, C.; TAN, C.; YONG, A. A grid and density-based clustering algorithm for processing data stream. In: *Proceedings - 2nd International Conference on Genetic and Evolutionary Computing, WGEC 2008*. [S.l.: s.n.], 2008. p. 517–521.

KASABOV, N. et al. Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition. *Neural networks : the official journal of the International Neural Network Society*, Elsevier Ltd, v. 41, n. 1995, p. 188–201, maio 2013.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. [S.l.]: John Wiley and Sons, 1990. 368 p.

KENNEDY, J.; EBERHART, R. C.; SHI, Y. *Swarm Intelligence*. [S.l.]: Morgan Kaufmann, 2001. 512 p.

KHOLGHI, M.; KEYVANPOUR, M. Active Learning Framework Combining Semi-Supervised Approach for Data Stream Mining. In: *Intelligent Computing and Information Science*. [S.l.]: Springer Berlin Heidelberg, 2011, (Communications in Computer and Information Science, v. 135). p. 238–243.

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.

KMIECIAK, M. R.; STEFANOWSKI, J. Semi-supervised approach to handle sudden concept drift. *Control and Cybernetics*, v. 40, n. 3, p. 667–695, 2011.

KRISHNAPURAM, R.; KELLER, J. M. The possibilistic C-means algorithm: Insights and recommendations. *IEEE Transactions on Fuzzy Systems*, v. 4, n. 3, p. 385–393, 1996.

LABROCHE, N. Online fuzzy medoid based clustering algorithms. *Neurocomputing*, Elsevier, v. 126, p. 141–150, fev. 2014.

LANGLEY, P. The changing science of machine learning. *Machine Learning*, v. 82, n. 3, p. 275–279, 2011.

LEITE, D.; COSTA, P.; GOMIDE, F. Evolving granular neural network for semi-supervised data stream classification. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. [S.l.]: IEEE, 2010. p. 1–8.

LEMOS, A.; CAMINHAS, W.; GOMIDE, F. Adaptive fault detection and diagnosis using an evolving fuzzy classifier. *Information Sciences*, Elsevier Inc., v. 220, p. 64–85, jan. 2013.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. *Mining of Massive Datasets*. [S.l.]: Cambridge University Press, 2014. 476 p.

LI, F. A Pattern Query Strategy Based on Semi-supervised Machine Learning in Distributed WSNs. *Journal of Information and Computational Science*, v. 11, n. 18, p. 6447–6459, dez. 2014.

LI, P.; WU, X.; HU, X. Learning from Concept Drifting Data Streams with Unlabeled Data. In: *Proceedings of the National Conference on Artificial Intelligence*. [S.l.: s.n.], 2010. p. 1945–1946.

LI, P.; WU, X.; HU, X. Mining Recurring Concept Drifts with Limited Labeled Streaming Data. In: *JMLR: Workshop and Conference Proceedings 13*. [S.l.: s.n.], 2010. v. 3, n. 2, p. 241–252.

LI, P.; WU, X.; HU, X. Mining Recurring Concept Drifts with Limited Labeled Streaming Data. *ACM Transactions on Intelligent Systems and Technology*, v. 3, n. 2, p. 1–32, fev. 2012.

LIANG, C. et al. Learning very fast decision tree from uncertain data streams with positive and unlabeled samples. *Information Sciences*, Elsevier Inc., v. 213, p. 50–67, dez. 2012.

LIU, J.; LI, X.; ZHONG, W. Ambiguous decision trees for mining concept-drifting data streams. *Pattern Recognition Letters*, Elsevier B.V., v. 30, n. 15, p. 1347–1355, nov. 2009.

LIU, J. et al. A Semi-supervised Ensemble Approach for Mining Data Streams. *Journal of Computers*, v. 8, n. 11, p. 2873–2879, nov. 2013.

LIU, Y.; ZHOU, Y. Online Detection of Concept Drift in Visual Tracking. In: *Neural Information Processing*. [S.l.]: Springer International Publishing, 2014. p. 159–166.

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.

MAGDY, A.; YOUSRI, N. a.; EL-MAKKY, N. M. Discovering Clusters with Arbitrary Shapes and Densities in Data Streams. In: *2011 10th International Conference on Machine Learning and Applications and Workshops*. [S.l.]: IEEE, 2011. p. 279–282.

MARIN, L. et al. On-line dynamic adaptation of fuzzy preferences. *Information Sciences*, Elsevier Inc., v. 220, p. 5–21, jan. 2013.

MASUD, M. M. et al. A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. In: *2008 Eighth IEEE International Conference on Data Mining*. [S.l.]: IEEE, 2008. p. 929–934.

MASUD, M. M. et al. *A Practical Approach To Classify Evolving Data Streams: Training With Limited Amount Of Labeled Data*. [S.l.], 2008. 11 p.

MASUD, M. M. et al. Facing the reality of data stream classification: coping with scarcity of labeled data. In: *Knowledge and Information Systems*. [S.l.: s.n.], 2012. v. 33, n. 1, p. 213–244.

MILLÁN-GIRALDO, M.; SÁNCHEZ, J. S.; TRAVER, V. J. On-line learning from streaming data with delayed attributes: a comparison of classifiers and strategies. *Neural Computing and Applications*, v. 20, n. 7, p. 935–944, jun. 2011.

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

MOUCHAWEH, M. S. Semi-supervised classification method for dynamic applications. *Fuzzy Sets and Systems*, Elsevier, v. 161, n. 4, p. 544–563, fev. 2010.

NAHAR, V. et al. Semi-supervised Learning for Cyberbullying Detection in Social Networks. In: *Databases Theory and Applications*. [S.l.]: Springer International Publishing, 2014. p. 160–171.

NGUYEN, H.; NG, W.; WOON, Y. Concurrent Semi-supervised Learning with Active Learning of Data Streams. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems VIII*. [S.l.]: Springer Berlin Heidelberg, 2013. v. 7790, p. 113–136.

NGUYEN, H.-l. et al. Concurrent Semi-supervised Learning of. In: *Data Warehousing and Knowledge Discovery*. [S.l.]: Springer Berlin Heidelberg, 2011. p. 445–459.

PAN, J.; YANG, Q.; PAN, S. Online co-localization in indoor wireless networks by dimension reduction. In: *Proceedings of the National Conference on Artificial Intelligence*. [S.l.: s.n.], 2007. p. 1102–1107.

PATIL, P.; FATANGARE, Y.; KULKARNI, P. Semi-supervised Learning Algorithm for Online Electricity Data Streams. In: *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*. [S.l.]: Springer India, 2015. p. 349–358.

PEDRYCZ, W. Algorithms of fuzzy clustering with partial supervision. *Pattern Recognition Letters*, v. 3, n. January, p. 13–20, 1985.

PEDRYCZ, W.; GOMIDE, F. *An Introduction to Fuzzy Sets: Analysis and Design*. [S.l.]: MIT Press, 1998. (A Bradford book).

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.

REN, J.; CAI, B.; HU, C. Clustering over Data Streams Based on Grid Density and Index Tree. *Journal of Convergence Information Technology*, v. 6, n. 1, p. 83–93, 2011.

ROKACH, L. Ensemble-based classifiers. *Artificial Intelligence Review*, v. 33, p. 1–39, 2010.

RUIZ, C.; MENASALVAS, E.; SPILIOPOULOU, M. C-DenStream: Using domain knowledge on a data stream. In: *Discovery Science*. [S.l.]: Springer Berlin Heidelberg, 2009. p. 287–301.

SCHWENKER, F.; TRENTIN, E. Partially supervised learning for pattern recognition. *Pattern Recognition Letters*, v. 37, fev. 2014.

SEEGER, M. *Learning with labeled and unlabeled data*. [S.l.], 2001. 62 p.

SHAKER, A.; SENGE, R.; HüLLERMEIER, E. Evolving fuzzy pattern trees for binary classification on data streams. *Information Sciences*, Elsevier Inc., v. 220, p. 34–45, jan. 2013.

SHAMSHIRBAND, S. et al. D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks. *Measurement*, Elsevier Ltd, v. 55, p. 212–226, set. 2014.

SHEIKHOLESLAMI, G.; CHATTERJEE, S.; ZHANG, A. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: *Proceedings of the International Conference on Very Large Data Bases*. [S.l.: s.n.], 1998. p. 428–439.

SHI, X. et al. An incremental affinity propagation algorithm and its applications for text clustering. In: *2009 International Joint Conference on Neural Networks*. [S.l.]: IEEE, 2009. p. 2914–2919.

SILVA, D. et al. Semi-supervised classification of characterized patterns for demand forecasting using smart electricity meters. In: *2011 International Conference on Electrical Machines and Systems*. [S.l.]: IEEE, 2011. p. 1–6.

SILVA, J. A. et al. Data stream clustering: A survey. *ACM Computing Surveys*, v. 46, n. 1, p. 1–31, out. 2013.

SIRAMPUJ, T.; KANGKACHIT, T.; WAIYAMAI, K. CE-Stream : Evaluation-based technique for stream clustering with constraints. In: *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. [S.l.]: IEEE, 2013. p. 217–222.

SOUZA, L. et al. Thermal modeling of power transformers using evolving fuzzy systems. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 25, n. 5, p. 980–988, ago. 2012.

TIWARI, P.; KURHANEWICZ, J. Semi supervised multi kernel (SeSMiK) graph embedding: identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2010)*. [S.l.]: Springer Berlin Heidelberg, 2010. p. 666–673.

TSAI, C.-J.; LEE, C.-I.; YANG, W.-P. Mining decision rules on data streams in the presence of concept drifts. *Expert Systems with Applications*, Elsevier Ltd, v. 36, n. 2, p. 1164–1178, mar. 2009.

TSYMBAL, A. et al. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, v. 9, n. 1, p. 56–68, jan. 2008.

TU, L.; CHEN, Y. Stream data clustering based on grid density and attraction. *ACM Transactions on Knowledge Discovery from Data*, v. 3, n. 3, p. 12:1–12:27, 2009.

VITO, S. et al. Semi-Supervised Learning Techniques in Artificial Olfaction: A Novel Approach to Classification Problems and Drift Counteraction. *IEEE Sensors Journal*, v. 12, n. 11, p. 3215–3224, nov. 2012.

WANG, A. et al. An incremental extremely random forest classifier for online learning and tracking. In: *2009 16th IEEE International Conference on Image Processing (ICIP)*. [S.l.]: IEEE, 2009. p. 1449–1452.

WANG, L.; JI, H.-B.; JIN, Y. Fuzzy Passive-Aggressive classification: A robust and efficient algorithm for online classification problems. *Information Sciences*, Elsevier Inc., v. 220, p. 46–63, jan. 2013.

WANG, W.; YANG, J.; MUNTZ, R. STING: A statistical information grid approach to spatial data mining. In: *Proceedings of International Conference on Very Large Data*. [S.l.: s.n.], 1997. p. 1–18.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. [S.l.: s.n.], 2005. 560 p.

WU, Q.-Y.; YE, Y.; FU, J. Learnable topical crawler through online semi-supervised clustering. In: *2009 International Conference on Machine Learning and Cybernetics*. [S.l.]: IEEE, 2009. p. 231–236.

WU, S.; YANG, C.; ZHOU, J. Clustering-training for Data Stream Mining. In: *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*. [S.l.]: IEEE, 2006. p. 653–656.

WU, X.; LI, P.; HU, X. Learning from Concept Drifting Data Streams with Unlabeled Data. *Neurocomputing*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 92, p. 145–155, 2012.

XU, W.-h.; QIN, Z.; CHANG, Y. Clustering feature decision trees for semi-supervised classification from high-speed data streams. *Journal of Zhejiang University SCIENCE C*, SP Zhejiang University Press, v. 12, n. 8, p. 615–628, 2011.

XUEJUN, W. Semi-supervised Data Stream Ensemble Classifiers Algorithm Based on Cluster Assumption. In: *Software Engineering and Knowledge Engineering: Theory and Practice*. [S.l.: s.n.], 2012. v. 1, p. 713–721.

YAN, Y.; CHEN, L. Label-based semi-supervised fuzzy co-clustering for document categoraization. In: *2011 8th International Conference on Information, Communications & Signal Processing*. [S.l.]: IEEE, 2011. p. 1–5.

YOGITA, Y.; TOSHNIWAL, D. Clustering techniques for streaming data - a survey. In: *2013 3rd IEEE International Advance Computing Conference (IACC)*. [S.l.]: IEEE, 2013. p. 951–956.

YU, Y. et al. Anomaly intrusion detection for evolving data stream based on semi-supervised learning. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [S.l.: s.n.], 2009. v. 5506 LNCS, p. 571–578.

ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965.

ZHANG, D. et al. A Clustering Algorithm Based on Density-Grid for Stream Data. In: *2012 13th International Conference on Parallel and Distributed Computing, Applications and Technologies*. [S.l.]: IEEE, 2012. p. 398–403.

ZHANG, J.-p. et al. Online stream clustering using density and affinity propagation algorithm. In: *2013 IEEE 4th International Conference on Software Engineering and Service Science*. [S.l.]: IEEE, 2013. p. 828–832.

ZHANG, P. et al. A framework for application-driven classification of data streams. *Neurocomputing*, Elsevier, v. 92, p. 170–182, set. 2012.

ZHANG, P.; ZHU, X.; GUO, L. Mining Data Streams with Labeled and Unlabeled Training Examples. In: *2009 Ninth IEEE International Conference on Data Mining*. [S.l.]: IEEE, 2009. p. 627–636.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In: *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 1996. (SIGMOD '96), p. 103–114.

ZHANG, X. et al. Data Stream Clustering With Affinity Propagation. *IEEE Transactions on Knowledge and Data Engineering*, v. 26, n. 7, p. 1644–1656, jul. 2014.

ZHANG, X.; FURTLEHNER, C.; SEBAG, M. Data streaming with affinity propagation. In: *Machine Learning and Knowledge ...* . [S.l.]: Springer Berlin Heidelberg, 2008. p. 628–643.

ZHU, X. *Semi-Supervised Learning Literature Survey*. [S.l.], 2005. 60 p. Disponível em: <http://pages.cs.wisc.edu/ jerryzhu/pub/ssl_survey.pdf>.

ZHU, X.; GOLDBERG, A. B. *Introduction to Semi-Supervised Learning*. [S.l.: s.n.], 2009. 130 p.