

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

**CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**MINERAÇÃO EM FLUXOS CONTÍNUOS DE  
DADOS POR MEIO DE TÉCNICAS DE  
APRENDIZADO 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

**MINERAÇÃO EM FLUXOS CONTÍNUOS DE  
DADOS POR MEIO DE TÉCNICAS DE  
APRENDIZADO 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

Existe hoje uma variedade de sistemas que produzem grande quantidade de dados em curto espaço de tempo, como redes de sensores, mercado financeiro e sistemas de segurança. Estes conjuntos de dados têm tamanho indefinido, potencialmente infinito, e podem gerar exemplos com distribuição estatística mutável de acordo com o tempo. As fontes que geram esse tipo de conjuntos são conhecidas como Fluxo Contínuo de Dados (FCD). Métodos de Aprendizado de Máquina (AM) podem auxiliar a tomada de decisão no contexto de diversas aplicações pela aquisição de conhecimento utilizável a partir de um conjunto de dados. Abordagens mais clássicas de AM não são capazes de lidar com algumas características particulares de FCD. Para que seja possível realizar o aprendizado com FCD, é necessário incorporar uma variedade de mecanismos aos métodos. Desafios relacionados especificamente a FCD incluem a inviabilidade de armazenamento de todo o conjunto de dados em memória, a impossibilidade de processar um mesmo exemplo muitas vezes durante a tarefa de aprendizagem e a necessidade de manutenção para que o modelo continue representando os dados mais recentes do FCD. Métodos de aprendizado semissupervisionado vêm sendo aplicados com sucesso em conjuntos com tamanho e distribuição fixa e podem oferecer uma alternativa ao aprendizado de FCD onde existe um grande volume de dados não rotulados e uma pequena parte de dados rotulados ou alguma informação prévia sobre a relação entre esses dados. A proposta de trabalho apresentada aqui tem por objetivo investigar o aprendizado em FCD por meio de métodos semissupervisionados. Inicialmente são apresentados conceitos gerais de AM, questões específicas relacionadas ao aprendizado com dados de fluxo contínuo, e uma coletânea de abordagens que representam o estado-da-arte quanto a pesquisas no campo de aprendizado com FCD. Ao final do documento é apresentada a proposta de trabalho para investigação, desenvolvimento e análise de um método de aprendizado semissupervisionado a partir de dados de fluxo contínuo.

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

# ABSTRACT

Nowadays there is a variety of systems that produce great quantity of data in a short time space, such as sensors networks, financial market and security systems. These data sets have undefined, potentially infinite size, and may generate examples with changing distribution through time. The sources that generate these types of sets are known as Data Streams (DS). Machine Learning (ML) methods can help decision making in several applications by acquiring usable knowledge from data sets. More classical approaches for ML are not capable of dealing with some particular characteristics of DS. For making DS learning possible, it is necessary to incorporate several mechanisms to the methods. Challenges related specifically to DS include the unviability of storing the complete data set in memory, the impossibility of processing on example multiple times during the learning task and the need for maintenance so the model continues to represent the most recent data in the DS. Semi-supervised learning methods have been applied with success to data sets of fixed size and distribution and may offer an alternative to DS learning where there is a great volume of unlabeled data and a small part of data that is labeled or some previous information about the relation between data examples. The work proposal presented here aims to investigate DS learning by means of semi-supervised methods. Initially concepts on ML, matters relating to DS learning, and a collection of approaches that represent the state-of-the-art as for research in the field of DS learning are presented. A work proposal for investigating, developing and analyzing a method of semi-supervised learning from data in Streams is presented at the end of this document.

**Keywords:** data streams, semi-supervised learning, clustering, classification

## LISTA DE FIGURAS

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 <i>learning/mining</i> e <i>data streams/streaming data</i> . . . . .	19
2.2	<i>Framework</i> para agrupamento baseado em atributo (múltiplos FCD) (SILVA et al., 2013) . . . . .	20
2.3	Exemplos ilustrativos de modelos de janela temporal (SILVA et al., 2013) . . . . .	22
2.4	<i>Framework online-offline</i> (SILVA et al., 2013) . . . . .	26
3.1	Exemplos de árvores construídas a partir de um mesmo conjunto de exemplos (HUANG; GEDEON; NIKRAVESH, 2008) . . . . .	32

# **LISTA DE TABELAS**

4.1 Cronograma de atividades a partir de março de 2015 . . . . . 50

## ACRÔNIMOS E SIGLAS

---

**AH** – *Árvores de Hoeffding*

**AM** – *Aprendizado de Máquina*

**AP** – *Affinity 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*

**VFDT** – *Very Fast Decision Tree*

**cmc** – *core-micro-cluster*

**pcmc** – *potential core-micro-cluster*

# SUMÁRIO

<b>CAPÍTULO 1 – INTRODUÇÃO</b>	<b>9</b>
1.1 Motivação e Objetivos . . . . .	11
1.2 Organização . . . . .	12
<b>CAPÍTULO 2 – CONCEITOS GERAIS</b>	<b>13</b>
2.1 Aprendizado Semissupervisionado . . . . .	13
2.1.1 Aprendizado Supervisionado e Não Supervisionado . . . . .	14
2.1.2 Abordagens de Aprendizado Semissupervisionado . . . . .	16
2.2 Aprendizado em Fluxos Contínuos de Dados . . . . .	18
2.2.1 Janelas Temporais . . . . .	21
2.2.2 Desvios de Conceito . . . . .	23
2.2.3 Árvores de Hoeffding . . . . .	24
2.2.4 Agrupamento em Fluxos Contínuos de Dados . . . . .	25
2.2.5 Ferramentas . . . . .	28
2.3 Considerações Finais . . . . .	29
<b>CAPÍTULO 3 – ABORDAGENS PARA APRENDIZADO EM FLUXOS CONTÍNUOS DE DADOS</b>	<b>30</b>
3.1 Técnicas de Classificação em Fluxos Contínuos de Dados . . . . .	31
3.2 Técnicas de Agrupamento em Fluxos Contínuos de Dados . . . . .	33
3.2.1 Agrupamento Particional em FCD . . . . .	34



3.2.2	Agrupamento Baseado em Densidade em FCD . . . . .	35
3.3	Técnicas de Aprendizado Semissupervisionado em Fluxos Contínuos de Dados	39
3.3.1	Classificação Semissupervisionada em FCD . . . . .	39
3.3.2	Agrupamento Semissupervisionado em FCD . . . . .	40
3.3.3	Aprendizado Semissupervisionado Híbrido em FCD . . . . .	41
3.3.4	<i>Ensemble</i> de Modelos para Aprendizado Semissupervisionado em FCD	43
3.4	Considerações Finais . . . . .	44
<b>CAPÍTULO 4 – PROPOSTA DE TRABALHO</b>		<b>46</b>
4.1	Tópicos Principais . . . . .	47
4.1.1	Aprendizado em Fluxos Contínuos de Dados . . . . .	47
4.1.2	Aprendizado Semissupervisionado . . . . .	48
4.2	Cronograma de Atividades . . . . .	49
4.3	Contribuições Esperadas . . . . .	50
<b>REFERÊNCIAS</b>		<b>52</b>

# Capítulo 1

## INTRODUÇÃO

---

Avanços recentes nas tecnologias de *hardware* e *software* vêm favorecendo a obtenção de dados em larga escala para uma variedade de domínios. Esses dados podem ser gerados continuamente e em grandes velocidades, gerando um fluxo de dados. Existe hoje uma variedade de sistemas que produzem grande quantidade de dados em curto espaço de tempo, como:

**Redes de Sensores** Conjunto de pequenos sensores distribuídos geograficamente para a extração de informações do ambiente onde se encontram, como em rede elétricas e pesquisas meteorológicas.

**Redes de Computadores** Especialmente na análise de tráfego em redes, com identificação de padrões não usuais como detecção de endereço IP de invasores.

**Mercado Financeiro** A análise de dados da bolsa de valores deve ser rápida a fim de trazer resultados relevantes a investidores.

**Sistemas de Segurança** Sistemas capazes de identificar fraudes de cartão de crédito ou rastreamento visual de um ambiente, buscando identificação de pessoas suspeitas.

Estes conjuntos de dados têm tamanho indefinido, potencialmente infinito, e podem gerar exemplos com distribuição estatística mutável de acordo com o tempo (GAMA, 2010). Fontes de dados com essas características são chamadas de Fluxos Contínuos de Dados.

Por não possuir tamanho definido, o armazenamento da totalidade de dados de um fluxo é inviável. A característica de chegada constante de novos dados impõe certa urgência no processamento dos dados antigos. Estas restrições impossibilitam o uso de sistemas já existentes de armazenamento de dados, pois esses sistemas não foram desenvolvidos para operar com dados

gerados em fluxo rápido e contínuo (BABCOCK et al., 2002). O desenvolvimento de mecanismos que possam lidar com dados gerados em fluxo contínuo para atender as diversas aplicações neste contexto torna-se imprescindível.

O Aprendizado de Máquina refere-se à investigação de métodos computacionais capazes de adquirir conhecimento de forma automática. Por meio de métodos de Aprendizado de Máquina sistemas computacionais podem aprender e otimizar seu desempenho de forma a torná-lo mais preciso. Algoritmos de Aprendizado de Máquina cumprem um papel importante em um grande número de aplicações, transformando dados em informações úteis para a tomada de decisão em diversos ambientes.

A maioria dos algoritmos mais tradicionais de Aprendizado de Máquina, no entanto, consideram que o conjunto total de dados está disponível e pode ser acessado a qualquer momento. Dentro da área de Fluxo Contínuo de Dados, esses algoritmos tradicionais são incapazes de desempenhar seu trabalho.

Para conseguir fazer a aquisição de conhecimento útil em ambientes dinâmicos, métodos de Aprendizado de Máquina devem ser adaptados para incorporar novos dados de forma incremental.

A chegada de dados em ambientes de fluxo contínuo pode ser rápida e esses dados não podem ser armazenados por muito tempo, isso significa que não é possível ler e processar os dados diversas vezes para a construção de um modelo. Idealmente, um algoritmo de Aprendizado de Máquina incremental deve fazer uma varredura simples sobre o Fluxo Contínuo de Dados.

Outra característica intrínseca a um Fluxo Contínuo de Dados é o fato da distribuição dos dados não ser estacionária, ou seja, podem ocorrer mudanças e evoluções dos dados. Esse aspecto também deve ser considerado pelos métodos que objetivam extrair informação dos dados, uma vez que existe a possibilidade dos modelos geradores se tornarem obsoletos com o tempo. É necessária, então, a implementação da capacidade de evolução do modelo de forma a acompanhar o desenvolvimento do fluxo.

A evolução do modelo pode acontecer por meio de estratégias de esquecimento, que delimitam a porção dos dados utilizada na geração do modelo, na intenção de priorizar o conhecimento obtido por dados mais recentes e descartar dados mais antigos (GAMA, 2010).

Desde a última década, surgem cada vez mais métodos diferentes que aplicam processo de aprendizagem em dados de fluxo contínuo. Esses métodos seguem, principalmente, abordagens supervisionadas (que realizam a extração de conhecimento pelo desenvolvimento de um modelo geral baseado em conjuntos de dados que possuem um atributo especial, chamado classe,

que representa o conceito que se deseja aprender) e não supervisionadas (processo capaz de realizar aprendizagem a partir de um conjunto de dados onde a informação de classes não está disponível).

Também nesse período, cresceram os esforços para investigação de métodos de aprendizado semissupervisionado, especialmente em conjuntos de dados clássicos, i.e., não contínuos. A motivação principal para utilização de métodos de aprendizado semissupervisionado é a disponibilidade de grandes volumes de dados em diversas áreas do conhecimento e o alto custo de interpretação e rotulação (atribuição de classe) manual desses dados. Métodos semissupervisionados consideram tanto um pequeno conjunto de dados rotulados, que possuem informação sobre classe, quanto um grande conjunto de dados não rotulados, que não possuem informação sobre classe, para realizar o processo de aprendizagem, obtendo resultados mais interessantes do que se realizasse o aprendizado exclusivamente pelo pequeno conjunto de dados rotulados ou pelo grande conjunto de dados não rotulados.

Em algumas aplicações de Fluxo Contínuo de Dados, é possível que não haja a totalidade de dados rotulados, ou que apenas um conjunto inicial do fluxo possua rótulos. Mais recentemente, métodos de aprendizado semissupervisionado vêm sendo adaptados para a incorporação de mecanismos capazes de lidar com as características de Fluxo Contínuo de Dados.

## **1.1 Motivação e Objetivos**

As vantagens da utilização de métodos de aprendizado semissupervisionado já foram evidenciadas em trabalhos anteriores (LOPES; CAMARGO, 2011, 2012). O estudo realizado para combinação de agrupamento semissupervisionado com sistemas baseados em regras *fuzzy* propiciou a continuação das investigações que, finalmente, deu origem à proposta apresentada neste documento.

O aprendizado em Fluxos Contínuos de Dados é uma tendência dentro das pesquisas na área de Aprendizado de Máquina, constituindo um tópico promissor para futuras pesquisas, em função do grande interesse observado nos últimos anos, devido à necessidade de aquisição de conhecimento em domínios de características específicas.

O objetivo do trabalho proposto aqui é a investigação de técnicas de aprendizado semissupervisionado aplicados ao contexto de Fluxo Contínuo de Dados. Para tanto, são caracterizados neste documento alguns conceitos iniciais sobre aprendizado em conjunto clássico e com dados de fluxo contínuo, estudando algumas das principais características dentro desta segunda área. Também é apresentado um conjunto de métodos conhecidos e diferenciados dentro do apren-

dizado em Fluxo Contínuo de Dados, a fim proporcionar uma visão geral das abordagens que vêm sendo utilizadas.

O levantamento bibliográfico realizado serve de base para a proposta de trabalho para tese de doutoramento, ambos apresentados neste documento.

## 1.2 Organização

O restante deste documento está estruturado como segue:

**Capítulo 2 - Conceitos Iniciais:** Este capítulo descreve conceitos básicos a respeito de aprendizado supervisionado, não supervisionado e semissupervisionado. É colocada uma caracterização do domínio de aplicações baseadas em Fluxos Contínuos de Dados, além de conceitos iniciais particulares ao Aprendizado de Máquina realizado nessa área e algumas noções básicas que fundamentaram o desenvolvimento de métodos para o aprendizado com dados de fluxo contínuo.

**Capítulo 3 - Abordagens para Aprendizado em Fluxos Contínuos de Dados:** Após a contextualização apresentada no capítulo anterior, este capítulo traz uma revisão bibliográfica a fim de traçar uma visão geral do estado-da-arte da área de aprendizado em Fluxos Contínuos de Dados. Os métodos que realizam aprendizado semissupervisionado são de particular interesse para este trabalho.

**Capítulo 4 - Proposta de Trabalho** Com base em todo o levantamento e investigações realizadas, parte apresentada neste documento, é apresentada neste capítulo uma proposta de trabalho para a continuação e desenvolvimento de tese dentro da área de aprendizado semissupervisionado com dados de fluxo contínuo.

# 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 prever eventos 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 possui 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 Neighbors* (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 abordagens 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. A ideia geral é que para cada exemplo de um grupo, a vizinhança, dentro de um raio determinado, possui um mínimo de pontos, ou seja, a densidade na vizinhança deve exceder um limiar definido. O formato de uma vizinhança é determinado pela escolha da função de distância utilizada analisar pares 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; SHEIKHOESLAMI; 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 pelo mesmo processo de aprendizado, chamado aprendizado semissupervisionado, 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.

Um número crescente de publicações e conferências sobre aprendizado semissupervisionado pode ser observado, sendo que as técnicas propostas têm sido aplicadas com sucesso, especialmente, em processamento de imagens (BENSAID et al., 1996; GRIRA; CRUCIANU; BOUJEMAA, 2006; PEDRYCZ et al., 2008) e classificação de textos (LIU; HUANG, 2003; GENG et al., 2009).

As publicações sugerem e analisam modificações de métodos já conhecidos a fim de considerar sua aplicação a um conjunto com maioria de dados não rotulados e uma pequena parte de dados rotulados. A obra de Zhu e Goldberg (2009) apresenta de forma resumida algumas tendências e características para classificação semissupervisionada, como *self-training*, *co-training* e *generative models* (CHAPELLE; SCHÖLKOPF; ZIEN, 2006), e apontamentos a respeito de outras formas de aprendizado semissupervisionado, como por agrupamento.

A utilização de métodos de agrupamento em aprendizado semissupervisionado pode ocorrer de duas formas: colaboração na rotulação do conjunto de dados ou agrupamento considerando informação prévia. No primeiro caso, algoritmos de agrupamento são aplicados ao conjunto de dados não rotulado para gerar grupos que, posteriormente, serão rotulados por algum outro método, com base na porção rotulada do conjunto. No segundo caso, métodos consagrados de agrupamento são modificados a fim de implementar a semissupervisão já no processo de geração de grupos e, em alguns casos, poder definir rótulos para estes grupos.

Chama-se de agrupamento semissupervisionado aquele realizado por métodos que incluem mecanismos para a consideração da informação dos rótulos pré-existentes (a informação pode vir na forma de rótulos ou não) no processo de geração de grupos. Os mecanismos utilizados incluem: modificação da função objetivo, para que inclua satisfação de restrições (PEDRYCZ; WALETZKY, 1997); Reforço de restrições durante o processo de agrupamento (WAGSTAFF et al., 2001; BASU; MOONEY, 2004; GRIRA; CRUCIANU; BOUJEMAA, 2005, 2008); Inicialização e restrição do agrupamento com base nos exemplos rotulados (BENSAID et al., 1996; BENSAID; BEZDEK, 1998; LABZOUR; BENSAID; BEZDEK, 1998; BASU; BANERJEE; MOONEY, 2002).

Os métodos desta classe podem ser divididos em duas abordagens para incorporação de semissupervisão, dependendo do conhecimento disponível: abordagem por sementes e abordagem por restrições entre pares. As sementes compõem uma parte do conjunto de dados que pode ser utilizada para estabelecer restrições ao algoritmo, restrições entre pares de dados e para definição de rótulos de grupos. As restrições entre pares podem ser da forma *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.

O agrupamento fuzzy semissupervisionado ocorre quando são incluídos mecanismos de semissupervisão em métodos de agrupamento fuzzy. A maior parte das publicações coloca a abordagem de Pedrycz (1985) como o primeiro trabalho na área de agrupamento *fuzzy* semissupervisionado. Anos após a primeira publicação, Pedrycz e Waletzky (1997) comentam sobre a falta de atenção dada aos mecanismos de supervisão parcial e discutem mais amplamente suas ideias.

Ainda hoje é possível identificar novas propostas de métodos semissupervisionados, uma vez que questões como o volume de dados e custo de rotulação manual de exemplos persiste. A proposta do trabalho (HAMASUNA; ENDO, 2011) introduz o conceito de tolerância entre grupos, utilizado em conjunto com restrições entre pares de exemplos para a construção de um novo algoritmo de agrupamento semissupervisionado baseado no FCM. Yan e Chen (2011) utilizam um conjunto de dados rotulados para inicialização e criação de restrições de pares de exemplos, extraídos a partir dos rótulos, durante o processo de agrupamento explorado dentro do contexto de categorização de documentos. O algoritmo *Data Understanding using Semi-Supervised Clustering* (BHATNAGAR et al., 2012) não necessita de parâmetros e processa os dados uma única vez, utilizando uma porção de exemplos rotulados para a identificação de pequenos grupos dentro das classes. Shamshirband et al. (2014) propõem o *D-FICCA*, um algoritmo de agrupamento que integra uma modificação, baseada em densidade e lógica *fuzzy*, para o algoritmo de competição imperialista (ATASHPAZ-GARGARI; LUCAS, 2007). Em (ZHENPENG et al., 2014) é proposto um algoritmo de agrupamento semissupervisionado baseado no *k-means* e ganho de informação para escolha dos protótipos iniciais. O trabalho de Schwenker e Trentin (2014) traz uma revisão atual de outros métodos de agrupamento semissupervisionado.

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.

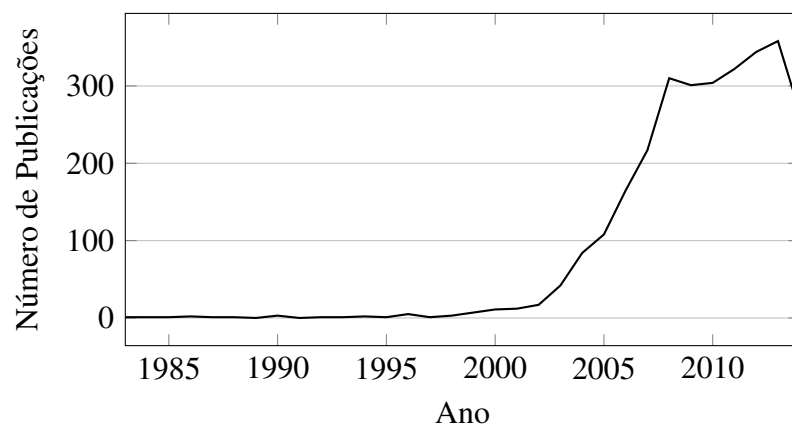
## 2.2 Aprendizado em Fluxos Contínuos de Dados

Existe hoje uma variedade de sistemas que produzem grande quantidade de dados em curto espaço de tempo, como monitoração de tráfego de rede (AGGARWAL; YU, 2008; YU et al., 2009; ZHANG et al., 2012; BREVE; ZHAO, 2013), redes de sensores (GAMA; GABER, 2007; PAN; YANG; PAN, 2007; ZHANG et al., 2012; BOUCHACHIA; VANARET, 2014), mineração de *clicks* na *web* (MARIN et al., 2013), medida de consumo de energia (SILVA et al., 2011; ZHANG et al., 2012), fraude de cartão de crédito (WU; LI; HU, 2012), mineração de textos da *web* (FDEZ-RIVEROLA

et al., 2007; CHENG et al., 2011; KMIECIAK; STEFANOWSKI, 2011; NAHAR et al., 2014a), rastreamento visual (LIU; ZHOU, 2014), olfação artificial (VITO et al., 2012), pesquisa meteorológica, mercado de ações e registros de supermercados (YOGITA; TOSHNIWAL, 2013).

Estes conjuntos de dados têm tamanho indefinido, potencialmente infinito, e podem gerar exemplos com distribuição estatística mutável de acordo com o tempo (GAMA, 2010).

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.



**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***

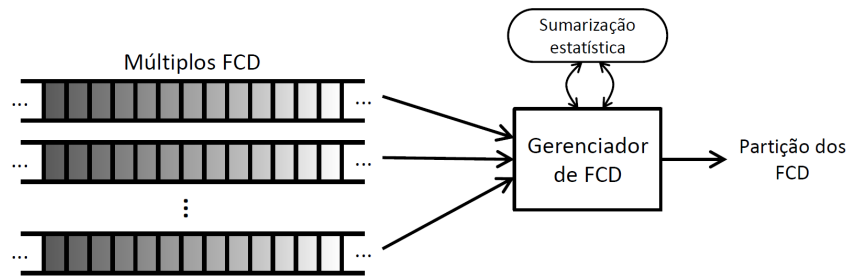
No modelo de FCD alguns ou todos os dados de entrada que serão utilizados não estão disponíveis em disco ou memória para acesso a qualquer momento, mas chegam de maneira contínua em um ou mais fluxos. FCDs diferem de conjuntos de dados convencionais em diversos aspectos (BABCOCK et al., 2002):

- Os exemplos no fluxo chegam de maneira contínua e constante;
- O sistema não possui controle sobre a ordem na qual os exemplos chegam para ser processados;
- Os fluxos tem tamanho potencialmente infinito;
- Uma vez que um exemplo do FCD foi processado, ele é descartado ou arquivado. Estes exemplos não podem ser recuperado de forma simples, pois guardá-los em memória ou

disco seria inviável.

No contexto de aprendizado em FCD, podemos identificar duas abordagens distintas: baseado em exemplos e baseado em atributo.

Os algoritmos de aprendizado em FCD baseado em atributos consideram um conjunto de múltiplos FCD e o objetivo do aprendizado é identificar padrões de comportamento entre os diferentes conjuntos, remetendo ao aprendizado de séries temporais. A Figura 2.2 apresenta um *framework* popular para o agrupamento de múltiplos FCD. Algoritmos de classificação de múltiplos FCD seguem esquema parecido.



**Figura 2.2:** *Framework* para agrupamento baseado em atributo (múltiplos FCD) (SILVA et al., 2013)

Alguns trabalhos na literatura seguem a abordagem baseada em atributos. O algoritmo *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.

Para este trabalho, no entanto, o foco está nos métodos que seguem a abordagem de aprendizado baseado em exemplos, na qual o objetivo é criar um modelo correspondente aos exemplos de um FCD. As estratégias descritas no restante deste documento referem-se ao aprendizado baseado em exemplos.

Um dos objetivos do aprendizado em FCD é encontrar de forma rápida um modelo alternativo válido, mesmo que seja apenas aproximado, do que seria obtido com abordagens clássicas de aprendizado, devido às limitações de tempo e espaço, típicas dessa forma de aprendizado. A estratégia mais frequentemente utilizada para representar dados do FCD, para contornar a impossibilidade de armazenamento de todos os dados, é a criação de sumários ou sinopses da informação encontrada nos dados. Uma grande variedade de técnicas tem sido desenvolvidas para o armazenamento de sumários ou sinopses da informação histórica encontrada em FCD. (GAMA; GABER, 2007).

É possível manter estatísticas simples de FCD, que podem ser computadas de forma incremental. Para definir a média de um FCD, por exemplo, precisamos manter o número de observações ( $i$ ) e a soma dos valores encontrados até o momento ( $\sum x_i$ ). Assim, com a chegada de um novo dados, a média pode ser calculada de forma incremental, como na equação 2.1.

$$\bar{x}_i = \frac{(i-1) \times \bar{x}_{i-1} + x_i}{i} \quad (2.1)$$

De maneira semelhante podem ser definidas outras estatísticas, como desvio padrão e coeficiente de correlação entre dois fluxos. O interessante nessas fórmulas é poder manter estatísticas exatas sobre uma sequência de dados potencialmente infinita sem ter que armazenar todos os dados (GAMA; GABER, 2007).

Contudo, este tipo de estatística tem uso limitado dentro do contexto de FCD, já que, na maior parte das aplicações, os dados recentes são os mais relevantes. Para contornar esse problema, uma proposta popular consiste de definir uma janela temporal que cubra os dados mais recentes.

### 2.2.1 Janelas Temporais

As janelas temporais são uma abordagem bastante utilizada para resolver a questão de conjuntos abertos (infinitos) como FCD. Ao invés do considerar todo o conjunto de exemplos de um fluxo, são considerados subconjuntos de exemplos ao longo do tempo. Neste modelo, uma marca temporal está associada a cada exemplo, a fim de determinar se o exemplo é válido ou não, ou seja, se está dentro ou fora de uma determinada janela temporal.

Existem diferentes modelos de janelas que podem ser encontrados na literatura. Os modelos *Sliding Windows*, *Damped Windows* e *Landmark Windows* são os mais relevantes (ZHU; SHASHA, 2002).

**Modelo *Sliding Window*:** neste modelo apenas a informação mais recente do FCD é armazenada em uma estrutura de dados cujo tamanho pode ser variável ou fixo. Esta é uma estrutura tipo *First In, First Out*, que considera os exemplos de um determinado ponto no tempo atual até um ponto no passado. A Figura 2.3a traz um exemplo do modelo *Sliding Window*. Algoritmos (REN; MA, 2009) que utilizam este modelo apenas atualizam os sumários estatísticos dos exemplos dentro da janela.

**Modelo *Damped Window*:** também conhecido como *time-fading*, este modelo considera a informação mais recente pela associação de pesos aos exemplos do FCD (JIANG; GRUNWALD, 2006): exemplos mais recentes tem peso maior que exemplos mais antigos e o peso dos exemplos diminui de acordo com o tempo. Um exemplo pode ser visualizado na Figura 2.3b, que mostra o decaimento do peso de acordo com o degradê dos exemplos. Algoritmos (CAO et al., 2006; CHEN; TU, 2007; ISAKSSON; DUNHAM; HAHSLER, 2012) baseados nesse modelo usualmente adotam uma função exponencial de decaimento para o peso dos exemplos.

**Modelo *Landmark Window*:** o processamento, neste modelo, se faz por porções disjuntas do FCD, nomeadas *chunks*, que são separadas de acordo com a ocorrência de *landmarks* (aparecimento de exemplos relevantes). Os *landmarks* podem ser definidos de acordo com o tempo, e.g., diário ou semanal, ou quanto ao número de elementos observados desde o *landmark* anterior. Quando um novo *landmark* é alcançado, todos os exemplos da janela são removidos e novos são adicionados a partir desse momento. Na Figura 2.3c há um exemplo para o modelo. Estratégias possíveis usando este modelo são baseadas na utilização dos modelos obtidos pelos diversos *chunks* em conjunto ou como guias para próximos modelos.

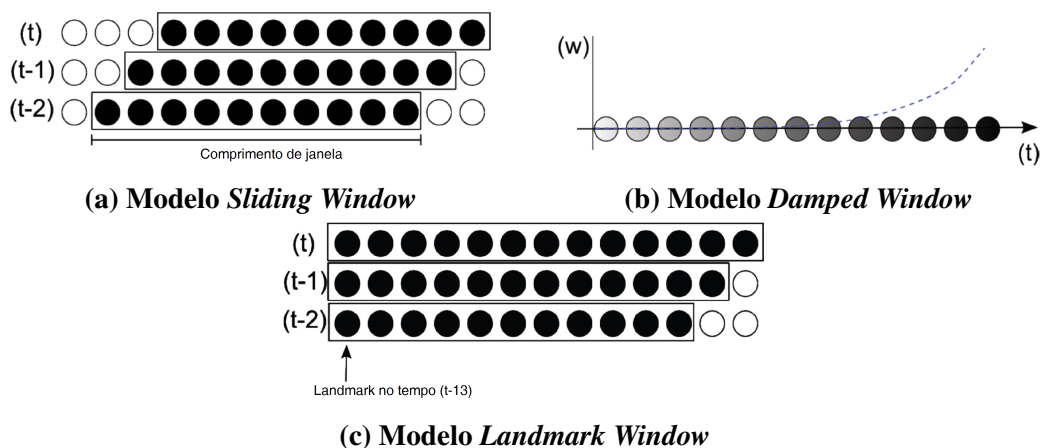


Figura 2.3: Exemplos ilustrativos de modelos de janela temporal (SILVA et al., 2013)

Além das características mencionadas a respeito do volume de exemplos, há também um componente temporal inerente a aprendizado em FCD. Os dados podem evoluir de acordo com o tempo e assim a distribuição do conjunto pode ser alterada. Deste modo, algoritmos que apenas sugerem adaptações para contornar as questões de volume de exemplos podem não ser soluções efetivas neste contexto. Algoritmos de aprendizado em FCD devem ter foco claro na evolução dos dados (AGGARWAL, 2007).

O uso de modelos de janela temporal podem auxiliar no tratamento de um dos aspectos da evolução dos exemplos do FCD, por permitir uma avaliação de acordo com o tempo. No entanto, existem outros aspectos que podem ser explorados. A próxima seção fala sobre a difícil tarefa de identificação de desvios de conceito.

### **2.2.2 Desvios de Conceito**

Na maioria das aplicações do mundo real, os dados são coletados durante um período de tempo. Para longos períodos, é plausível considerar que os exemplos não são independentes ou não possuem mesma distribuição. Em domínios complexos, é provável que a distribuição de classes mude de acordo com o tempo (GAMA et al., 2004). Essas mudanças são conhecidas como desvios de conceito.

Desvios de conceito podem ser graduais, onde há uma transição suave entre as distribuições, ou abruptas, quando a distribuição muda repentinamente.

Abordagens que lidam com desvios de conceitos podem ser classificadas em duas categorias: aquelas que adaptam o modelo em intervalos regulares sem considerar que mudanças ocorreram e aquelas que primeiro detectam desvios de conceitos e, então, adaptam o modelo a essas mudanças. As abordagens da primeira categoria são aquelas que utilizam modelos de janelas temporais.

As estratégias que realizam a detecção de mudanças para posterior adaptação do modelo mantêm um monitoramento, realizado pela definição de indicadores baseados no modelo. Se um desvio é detectado durante o monitoramento, são aplicadas ações para a adaptação do modelo de aprendizado.

Os trabalhos (GAMA et al., 2004; LI; WU; HU, 2012; WU; LI; HU, 2012) trazem maiores informações sobre algumas estratégias de detecção de desvios de conceito.

A seção a seguir descreve um método para indução de árvores de decisão a partir de dados que chegam de forma contínua, que serve de inspiração para variadas técnicas de aprendizado



em FCD.

### 2.2.3 Árvores de Hoeffding

Um dos métodos mais conhecidos e utilizados para classificação de exemplos é o aprendizado por árvores de decisão. Esses métodos induzem modelos na forma de árvores a partir dos dados disponíveis, onde cada nó contém um teste para um atributo, cada ramo uma possibilidade de valor para o teste e cada folha a predição de uma classe.

Uma árvore de decisão é aprendida pela recursiva troca de folhas por nós de teste. O atributo relacionado ao teste é escolhido pela comparação dos atributos disponíveis, de acordo com alguma métrica.

Métodos clássicos de aprendizado de árvores de decisão (QUINLAN, 1986, 1993) consideram que todos os exemplos de treinamento podem ser armazenados simultaneamente na memória principal, por isso, são limitados no número de exemplos dos quais podem aprender. Outros métodos consideram que os dados estão disponíveis em disco e realizam o aprendizado acessando sequencial e repetidamente os dados.

Domingos e Hulten (2000) propõem uma árvore de decisão capaz de aprender em domínios *online*, onde o conjunto de dados é potencialmente infinito, mas sua distribuição é estática. Esse método é conhecido como Árvore de Hoeffding (AH).

Uma AH requer que cada exemplo seja lido e processado apenas uma vez. A escolha do atributo para um nó da árvore é baseada em um pequeno subconjunto de exemplos treinamento. Dado um fluxo de exemplos, os primeiros serão usados para escolher o atributo da raiz. Escolhido o atributo raiz, os próximos exemplos são passados às folhas correspondentes e usados para escolher os atributos apropriados para a substituição por nós de teste e assim em diante. Em cada nó folha, o rótulo é escolhido de acordo com a maioria de exemplos da mesma classe presentes na folha.

Para definir o momento de criação de um nó teste, o limiar de Hoeffding ( $\epsilon$ ) é utilizado, sendo desnecessário definir um número fixo de exemplos. O objetivo ao utilizar esse índice é garantir que, com alta probabilidade, o atributo escolhido usando um pequeno conjunto de exemplos é o mesmo que seria escolhido a partir dos infinitos exemplos.

No mesmo trabalho, os autores sugerem uma implementação de um sistema de árvore de decisão baseado em AH, chamado *Very Fast Decision Tree* (VFDT.) O sistema VFDT constrói árvores de decisão usando memória e tempo por exemplo constantes, podendo incorporar

dezenas de milhares de exemplos por segundo.

Extensões de AH consideram outros métodos para determinar os rótulos nas folhas, podendo ser construído um modelo dentro de cada folha, a partir dos exemplos contidos na folha, para classificação de novos exemplos. Também são implementadas outras formas para determinar o momento de divisão de um nó folha e detecção e adaptação a desvios de conceito. Algumas dessas extensões são apresentadas no Capítulo 3.

O algoritmo de AH original considera que todos os exemplos do conjunto contínuo são rotulados. Essa realidade não é verdadeira para todas as aplicações FCD. A próxima seção apresenta algumas questões associadas especificamente a métodos de agrupamento em FCD, que consideram exemplos não rotulados no processo de aprendizagem.

## 2.2.4 Agrupamento em Fluxos Contínuos de Dados

Pela rápida e contínua chegada dos exemplos de um FCD, é natural inferir que grande parte das áreas onde algoritmos de aprendizado em FCD pode ser aplicado encaram a dificuldade da falta de rótulos disponíveis para a execução de métodos supervisionados. Devido a isso, cresce o interesse por abordagens de agrupamento em FCD.

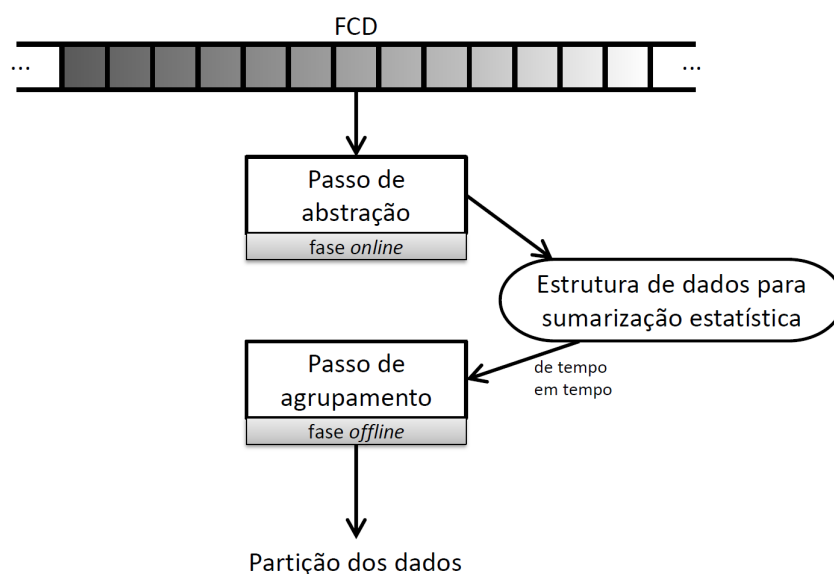
### 2.2.4.1 Um *Framework* para Agrupamento em FCD

Algoritmos de agrupamento baseados em exemplos podem ser resumidos em dois passos (CAO et al., 2006; YANG; ZHOU, 2006): abstração dos dados (componente *online*) e agrupamento (componente *offline*), ilustrados na Figura 2.4.

A fase *online*, abstração dos dados, sumariza os dados do FCD com o auxílio de estruturas particulares para lidar com restrições de espaço e memória das aplicações FCD. Essas estruturas de sumarizam os dados para preservar o significado dos objetos originais sem a necessidade de armazená-los. Estruturas frequentemente utilizadas são vetores de atributos, arranjos de protótipos e grades de dados. Essas estruturas são melhor detalhadas na Seção 2.2.4.2.

Para a contínua sumarização dos exemplos que chegam e dar maior importância aos exemplos mais recentes, uma abordagem popular é a definição de janelas temporais, como apresentado na Seção 2.2.1.

Durante o passo de abstração, algoritmos de agrupamento em FCD devem utilizar mecanismos para detecção de *outliers* que sejam capazes de diferenciar verdadeiros *outliers* de evolução de grupos (Seção 2.2.2), uma vez que a distribuição dos dados pode variar de acordo



**Figura 2.4: Framework online-offline (SILVA et al., 2013)**

com o tempo.

Na fase *offline* é possível obter uma partição dos dados pelo passo de agrupamento. Neste momento, pode ser necessária a definição de alguns valores de entrada (número de grupos, por exemplo) para que seja possível ter uma visão geral dos grupos do FCD. Algoritmos de agrupamento tradicionais podem ser utilizados considerando o conjunto de estruturas de sumarização para encontrar uma partição dos dados. O formato dos grupos encontrados está ligado ao algoritmo de agrupamento empregado, por exemplo, o *k-means* (MACQUEEN, 1967) gera grupos hipersféricos enquanto o DBSCAN (ESTER et al., 1996) é capaz de descobrir grupos de formatos aleatórios.

O *framework* apresentado nesta seção é frequentemente utilizado para o desenvolvimento de novas técnicas de agrupamento em FCD. Algumas dessas propostas são discutidas no Capítulo 3.

#### 2.2.4.2 Estruturas de Sumarização de Exemplos

A necessidade de utilizar estruturas de sumarização do FCD já foi expressada anteriormente neste documento. Esta seção traz algumas das estruturas de sumarização frequentemente utilizadas no passo de abstração dos dados de um FCD.

##### Vetor de Atributos

O uso de vetores de atributos para sumarização de grandes volumes de dados foi introduzido no algoritmo *BIRCH* (ZHANG; RAMAKRISHNAN; LIVNY, 1996). Este vetor, chamado de *Clus-*

tering Feature (CF), conta com três componentes: o número de exemplos ( $N$ ), a soma linear dos exemplos ( $LN$ ) e a soma quadrática dos exemplos ( $SS$ ), sendo que  $LN$  e  $SS$  são estruturas  $n$ -dimensionais, de acordo com o número de atributos do FCD. Essas componentes permitem o cálculo de métricas de grupo, como média, raio e diâmetro do grupo.

O vetor CF possui propriedades de incrementais e aditivas, ou seja, é possível inserir um novo exemplo em um CF pela atualização das estatísticas e dois CF podem ser mesclados em um terceiro vetor CF de forma simples.

Algumas abordagens de aprendizado em FCD utilizam o vetor CF como descrito nesta descrição, por vezes incluindo pesos para ponderar os grupos (CAO et al., 2006; KRANEN et al., 2011). Entretanto, há outras abordagens que utilizam variações do CF, a fim de produzir outras estatísticas.

A estrutura nomeada microgrupo, usada primeiramente no algoritmo *CluStream* (AGGARWAL et al., 2003), estende o conceito do vetor CF, adicionando mais duas componentes ao CF original: a soma de marcas temporais ou *timestamps* ( $LST$ ) e a soma quadrática de *timestamps* ( $SST$ ). As duas novas componentes tem o objetivo de incluir aspecto temporal na descrição de grupos, que pode ser utilizado para identificar *outliers* ou desvios de conceito.

A proposta do algoritmo *SWClustering* (ZHOU et al., 2008) também sugere uma extensão para o vetor CF, chamada de *Temporal CF*, que adiciona uma nova componente ao CF original: a *timestamp* do exemplo mais recente a ser inserido no grupo.

Algoritmos que fazem uso de microgrupos ainda podem manter um histórico dessas estruturas para determinar *snapshots* do FCD, i.e., recuperar a situação da partição de grupos em um determinado momento no tempo. (AGGARWAL et al., 2007)

### Arranjos de Protótipos

Alguns algoritmos de agrupamento utilizam uma estrutura simplificada chamada Arranjos de Protótipos, que consiste em um conjunto de protótipos (medóides, centróides, etc) que resumizam a partição dos dados.

O algoritmo *STREAM* (GUHA et al., 2000) divide o FCD em partes (*chunks*) e, para cada uma das partes, são definidos  $2k$  exemplos representativos obtidos por uma variante do algoritmo  $k$ -medóides (KAUFMAN; ROUSSEEUW, 1990). Esse processo é repetido até que seja completado um conjunto de  $m$  exemplos e, então, o agrupamento é aplicado aos protótipos com o objetivo de reduzir esse conjunto.

Estratégia similar é utilizada para o algoritmo *Stream LSearch* (O'CALLAGHAN et al., 2002),

que os protótipos em memória. Quando a memória está cheia, o conjunto de protótipos é agrupado a fim de manter na memória apenas um subconjunto de protótipos.

### Grades de Dados

A sumarização dos exemplos de um FCD também pode ser feita por meio de grades (CAO et al., 2006; CHEN; TU, 2007; GAMA; RODRIGUES; LOPES, 2011), ou seja, pelo particionamento do espaço  $n$ -dimensional de atributos em células grade de densidade.

Uma estratégia (CHEN; TU, 2007) para a utilização de grades é a associação de um coeficiente de densidade que decresce com o tempo. A densidade de uma célula de grade é determinada pela soma das densidades de cada exemplo inserido na grade. Cada célula é representada por uma tupla  $\langle tg, tm, D, label, status \rangle$ , onde  $tg$  é a última vez que a célula foi atualizada,  $tm$  é a última vez que a célula foi removida do conjunto de células válidas (não *outliers*),  $D$  é a densidade desde a última atualização,  $label$  é o rótulo de classe da célula e  $status$  indica se é uma célula normal ou esporádica (células com poucos objetos, *outliers*).

A manutenção das células de grade é realizada durante a fase *online*. Uma célula pode se tornar esparsa se não receber exemplos por muito tempo e uma célula esparsa pode se tornar densa se receber muitos exemplos. Quando um novo exemplo chega, é verificado a célula a qual pertence e estrutura da célula é atualizada. Células com o *status* esporádico são removidas periodicamente.

### 2.2.5 Ferramentas

Com o crescimento da pesquisa sobre aprendizado em FCD, é interessante o investimento em ferramenta de software para a aplicação das diversas técnicas propostas. Existem hoje duas ferramentas principais com este objetivo que estão disponíveis gratuitamente.

1. **MOA (Massive On-line Analysis)** (BIFET et al., 2010) é um *framework* de código aberto que disponibiliza implementação de uma série de algoritmos e métricas para classificação e agrupamento em FCD. A ferramenta também conta com recursos para visualização dos processos de aprendizado. Está disponível pelo endereço: <http://moa.cms.waikato.ac.nz/>.
2. **VFML (Very Fast Machine Learning)** (HULTEN; DOMINGOS, 2003) é um pacote de implementações para mineração de FCD de alta velocidade e conjuntos de dados *very large*. Está disponível pelo endereço: <http://www.cs.washington.edu/dm/vfml/>.

## 2.3 Considerações Finais

Este capítulo apresenta conceitos gerais relacionados a aprendizado semissupervisionado e em fluxos contínuos de dados, além de 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 *on-line* (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 organizadas de acordo com as suas principais características, desenvolvidas especificamente com este intuito ou não. As técnicas colocadas neste capítulo consideram a abordagem de aprendizado por exemplos, em que o processo de aprendizagem tem seu foco nas instâncias de um único FCD.

## 3.1 Técnicas de Classificação em Fluxos Contínuos de Dados

No caso em que o FCD envolve dados rotulados, o aprendizado pode ser feito por meio de técnicas supervisionadas. Encontra-se na literatura um volume significativo de trabalhos que utilizam a indução de modelos baseados em árvores. Outros trabalhos abordam o assunto propondo a criação de modelos híbridos utilizando conceitos *fuzzy*. Outra abordagem bastante explorada nesse contexto é a integração de múltiplos modelos. A seguir, alguns trabalhos representativos dessa categoria são apresentados.

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 2.2.3.

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* (VFDT) (DOMINGOS; HULTEN, 2000), uma proposta de aprendizado incremental baseada em AH, serviu de fundamento para outros métodos. A abordagem utiliza limiares de Hoeffding para garantir que a saída de suas árvores seja o mais próxima de um classificador convencional, que aprende a partir de conjuntos estáticos.

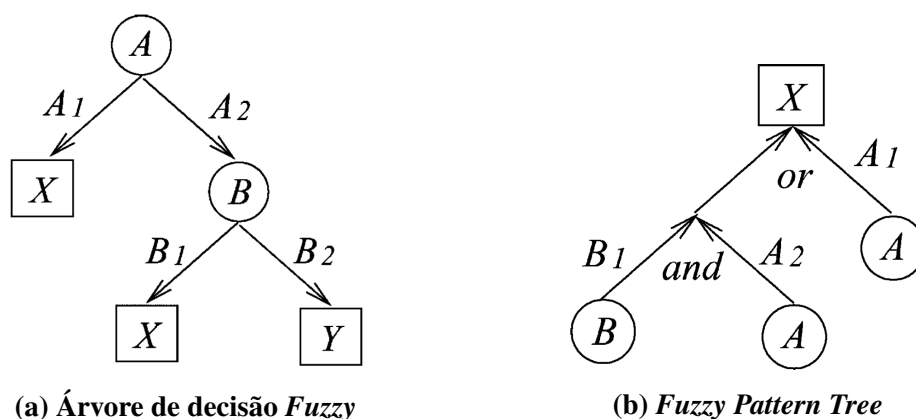
Uma das propostas que utiliza VFDT é a *Concept-adapting Very Fast Decision Tree* (CVFDT) (HULTEN; SPENCER; DOMINGOS, 2001), que tem como foco a detecção e adaptação a desvios de conceito em FCDs. Liu, Li e Zhong (2009) apresentam uma proposta de um mecanismo para integração de ambiguidades à CVFDT, 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 mineração de regras de conceito é feita pelo algoritmo CDR-Tree, que considera dois blocos de dados de um FCD que dizem respeito aos mesmos indivíduos em tempos distintos. A partir dos exemplos de cada bloco, o algoritmo forma pares que serão utilizados para a construção de uma árvore de decisão tradicional. As regras de desvio de conceito são mineradas a partir da árvore construída com os pares de exemplos velhos (bloco de dados no tempo  $\delta$ ) e novos (bloco de dados no tempo  $\delta + t$ ). No mesmo trabalho são propostas 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*.

*Fuzzy Pattern Trees* (HUANG; GEDEON; NIKRAVESH, 2008) são árvores induzidas por processo diferenciado, utilizando operadores de agregação distintos de árvores de decisão *fuzzy* tradicionais, podendo gerar regras que utilizam t-conormas (OU) além de t-normas (E). A fim de ilustrar a diferença entre as estruturas, a Figura 3.1 traz as árvores de decisão construídas a partir de um mesmo pequeno conjunto de dados exemplo utilizando um algoritmo tradicional de árvore de decisão *fuzzy* (Figura 3.1a) e o algoritmo proposto para a construção de *Fuzzy Pattern Trees* (Figura 3.1b).



**Figura 3.1:** Exemplos de árvores construídas a partir de um mesmo conjunto de exemplos (HUANG; GEDEON; NIKRAVESH, 2008)

A proposta apresentada em (SHAKER; SENGE; HÜLLERMEIER, 2013) é um método de classificação adaptativo baseado na indução de *Fuzzy Pattern Trees*. Esta proposta implementa um mecanismo de adaptação que identifica, de forma preditiva, possíveis mudanças locais no modelo atual. A ideia básica do método é manter um conjunto de *Fuzzy Pattern Trees* composto pelo modelo atual (ativo) e um conjunto de modelos vizinhos. A predição é realizada pelo modelo ativo, enquanto os modelos vizinhos funcionam como adaptações antecipadas. Os modelos vizinhos são mantidos até que a performance do modelo ativo tenha queda significativa, causada, por exemplo, por um desvio de conceito. Neste momento, o modelo atual é substituído pelo modelo vizinho mais adequado.

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 proposto e utilizado no trabalho, que é baseado na descoberta de centróides de classe e distância entre pares de exemplos, pode colaborar para a identificação de 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-Aggressive* (PA) (CRAMMER et al., 2006), constituindo a técnica neuro-fuzzy (JANG; SUN; MIZUTANI, 1997) chamada de *Fuzzy Passive-Aggressive*.

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 de identificação e tratamento de 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.

Desvios 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.

Pesquisadores têm investido na área de classificação em FCD, porém é importante ressaltar que nem sempre existem rótulos disponíveis para realizar esse tipo de tarefa de aprendizado e a rotulação manual dos exemplos de um FCD é inviável considerando seu tamanho potencialmente infinito. A próxima seção descreve técnicas de agrupamento em FCD.

## 3.2 Técnicas de Agrupamento em Fluxos Contínuos de Dados

Abordagens de agrupamento de dados são tipicamente utilizadas no aprendizado não supervisionado. Dentro da área 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. A Seção 2.2.4 discute algumas características específicas do agrupamento de dados em FCD e mecanismos para lidar com tais características.

Independentemente dos métodos adotados, é desejável que algoritmos de agrupamento em FCDs possuam a capacidade de (AMINI; WAH; SABOOHI, 2014): *a)* descobrir grupos de formatos arbitrários; *b)* lidar com ruído; *c)* realizar o agrupamento sem informação prévia sobre o número de grupos. Há diferentes técnicas para agrupamento em FCD e elas podem ser divididas de acordo com a abordagem de agrupamento que seguem.

O *ClusTree* (KRANEN et al., 2011), por exemplo, é a proposta de uma estrutura hierárquica compacta e auto-adaptativa para manter sumários de FCD, construindo uma hierarquia de microgrupos em diferentes níveis. Outras técnicas que seguem a abordagem hierárquica de agrupamento são o *E-Stream* (UDOMMANETANAKIT; RAKTHANMANON; WAIYAMAI, 2007), que possui suporte para cinco tipos de evolução em grupos (aparecimento, desaparecimento, evolução própria, mescla e divisão), e suas extensões para suporte de incerteza (valores faltantes) em FCDs heterogêneos (conjuntos que combinam atributos numérico-contínuos e categóricos) (MEESUKSABAI; KANGKACHIT; WAIYAMAI, 2011) e FCDs de alta dimensão (CHAIRUKWATTANA et al., 2014).

As subseções a seguir apresentam técnicas já existentes baseadas em agrupamento particional e baseado em densidade, onde é possível identificar focos de pesquisas mais recentes.

### 3.2.1 Agrupamento Particional em FCD

Os algoritmos de agrupamento particional para FCD, em sua maioria, são propostos como extensões de algoritmos de agrupamento particionais conhecidos, como *k-means*, *k-medóides* e *Affinity Propagation*.

O algoritmo *CluStream* (AGGARWAL et al., 2003) é baseado no algoritmo *k-means* e introduz um *framework online-offline* para agrupamento em FCD que vem sendo adotado para grande parte dos algoritmos propostos recentemente. Yang e Zhou (2006) propõem uma extensão chamada *HCluStream* para lidar com FCDs heterogêneos.

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), *possibilitistic 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.

Alguns trabalhos propõem algoritmos baseados no agrupamento *Affinity Propagation* (AP) (FREY; DUECK, 2007). 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. Com os *exemplars* não é necessário definir o número de grupos inicialmente.

Uma extensão do AP para aprendizado em FCD é o algoritmo *Streaming AP* (ZHANG; FURTLEHNER; SEBAG, 2008). Esta proposta 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 distribuição dos dados.

### 3.2.2 Agrupamento Baseado em Densidade em FCD

Os algoritmos baseados em densidade também são utilizados como alternativa para a tarefa de agrupamento, sendo duas de suas vantagens a alta tolerância a ruído ou *outliers* e a habilidade em descobrir grupos de formatos arbitrários.

As técnicas de agrupamento baseado em densidade seguem, comumente, duas abordagens que são descritas nas próximas seções e incluem exemplos de algoritmos que se encaixam nessas categorias.

### 3.2.2.1 Algoritmos de Microgrupos de Densidade

Em algoritmos de agrupamento de microgrupos de densidade, microgrupos mantêm a informação de sumarização dos exemplos e o agrupamento é realizado usando as estruturas de sinopse.

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* (cmc), referentes ao agrupamento em si, e *potential core-micro-cluster* (pcmc), distribuição de exemplos que representa regiões menos densas que são mantidas. O aprendizado da estrutura do FCD é realizado em duas fases. Na fase *online* do algoritmo, cada novo exemplo pode ser associado a um microgrupo já existente (cmc ou pcmc), de acordo com cálculo de métrica de dissimilaridade (distância Euclidiana) ou será criado um novo pcmc para o novo exemplo. Na fase *offline* é aplicado o algoritmo *DBSCAN* (ESTER et al., 1996) para determinar o grupos finais, de acordo com o conjunto de cmc. De tempos em tempos, um método de poda avalia o conjunto de pcmc para garantir que são *outliers*, de acordo com a densidade, e descartá-los. O *DenStream* não possui mecanismos para a eliminação de microgrupos ou para fundir dois ou mais microgrupos, o que pode ser problemático conforme o crescimento do conjunto de exemplos e limitações de espaço para armazenamento.

O *DenStream* serviu de inspiração para outras técnicas que consideram situações particulares e implementam adaptações para lidar com contextos diversos. Li-xiong et al. (2009) desenvolveram um algoritmo baseado no *DenStream* para aplicações com grande volume de outliers. O algoritmo *SDStream* (REN; MA, 2009) tem a habilidade de descobrir grupos de formatos arbitrários dentro de um modelo de janela deslizante, que permite o esquecimento progressivo dos dados antigos. *HDenStream* (JINXIAN; HUI, 2009) é um algoritmo adaptado para aprendizado em FCDs com atributos heterogêneos pela inclusão de um atributo bidimensional para manter a frequência de atributos categóricos. O *HDDStream* (NTOUTSI et al., 2012) traz adaptações ao original *DenStream* para melhorar o agrupamento de FCDs de alta dimensão. O *PreDeConStream* (HASSANI et al., 2012) melhora a eficiência da fase *offline* do *HDDStream*.

Alguns métodos híbridos utilizam conceitos do algoritmo *DenStream* aliado a outras abordagens. O *StreamOptics* (TASOULIS; ROSS; ADAMS, 2007) é um *framework* baseado nos conceitos de cmc e pcmc, que utiliza o algoritmo *OPTICS* (ANKERST et al., 1999) para produzir visualização gráfica da estrutura do FCD e sua evolução com o passar do tempo. No entanto, em nenhum momento é gerada uma partição do conjunto, então a análise do agrupamento deve ser realizada manualmente. Isaksson, Dunham e Hahsler (2012) propõem o algoritmo *SOS-*

*tream*, que detecta estrutura de FCDs de rápida evolução pela adaptação automática de limiar para o agrupamento baseado em densidade. O limiar é individual por grupo e é definido automaticamente dentro do processo de agrupamento, baseado na ideia de construir vizinhanças com um mínimo de pontos, como parte da análise para criação, remoção, mescla e divisão de grupos. O algoritmo utiliza aprendizado competitivo como em *Self Organizing Maps* (KOHONEN, 1982), o que pode tornar o processo mais oneroso.

O algoritmo *APDenStream* (ZHANG et al., 2013) baseia-se nos métodos AP e *DenStream* para definição de um modelo geral que representa o FCD. O algoritmo AP substitui o *DBSCAN* na fase *offline* do *DenStream*. Baseado em trabalho anterior dos autores (FORESTIERO; PIZZUTI; SPEZZANO, 2009), FlockStream (FORESTIERO; PIZZUTI; SPEZZANO, 2013) utiliza um sistema multi-agente baseado em um modelo de *flocking* (KENNEDY; EBERHART; SHI, 2001). Nesta técnica os agentes são microgrupos que trabalham de forma independente mas formam grupos juntos.

### 3.2.2.2 Algoritmos Baseados em Densidade e Grades

Uma tendência que pode ser observada nos trabalhos mais recentes é a integração entre diferentes abordagens de agrupamento, especificamente agrupamento baseado em grades e baseado em densidade. Há estudos (AMINI et al., 2011; AMINI; WAH; SABOOHI, 2014) sobre técnicas que seguem esta abordagem.

Métodos de agrupamento baseados em grade e densidade mapeiam os pontos de dados em grades (células do espaço de dados) que são agrupados de acordo com suas densidades. Uma das primeiras tentativas de associar os dois métodos foi o trabalho de Gao et al. (2005), que propõe um algoritmo de agrupamento incremental de passagem única usando unidades densas (em forma de grades), que são consideradas em uma fase de agrupamento caso tenham densidade acima de um limiar pré-definido.

*D-Stream I* (CHEN; TU, 2007) é a proposta de um *framework* para agrupamento de FCD em tempo real, que conta com fases *online* e *offline*. Na fase *online* acontece a leitura de um novo exemplo e seu mapeamento na grade. Na fase *offline* os grupos são ajustados em intervalos de tempo de acordo com o procedimento: no primeiro intervalo de tempo, cada grade densa é associada a um grupo distinto; a cada novo intervalo de tempo os grupos são ajustados pela identificação de grades densas e esparsas; se a grade for densa, será mesclada a outras grades vizinhas, formando um grupo, caso contrário a grade é removida do grupo. O *D-Stream I* agrupa os exemplos em tempo real de um FCD utilizando conceitos de agrupamento por densidade e grades, incluindo mecanismos para decaimento de densidade, detecção de evolução

de comportamento e para lidar com *outliers*.

Algumas extensões para o *D-Stream I* são propostas. Jia, Tan e Yong (2008) propuseram um modelo que melhora a qualidade dos grupos pela detecção dos pontos limites de uma grade. A extensão *D-Stream II* (TU; CHEN, 2009) inclui uma restrição de atração entre grades para a mesclagem e composição de um grupo na fase *offline*. A atração entre grades é verificada pela construção de hipercubos centrados nos exemplos de uma grade, considerando cada atributo do FCD, a fim de estabelecer um volume para a grade, que será utilizada para verificar se duas grades devem ou não ser mescladas em um mesmo grupo.

Ao realizar agrupamento baseado em grades, quanto maior a dimensão do conjunto de dados, maior o número de grades vazias. Pensando nisso, Ren, Cai e Hu (2011) propõem o algoritmo *PKS-Stream* para conjuntos de alta dimensão. A técnica possui uma estrutura de árvore para manter as grades não-vazias e suas relações. À chegada de um novo exemplo, o algoritmo verifica as estruturas de sumarização de grade para ver se o exemplo pertence a alguma das grades existentes na estrutura de árvore. Caso não seja verdadeiro, um novo elemento de sumarização de grade é criado. De tempo em tempo, a árvore é ajustada, pela remoção de grades esparsas, e os grupos são formados de acordo com a densidade das grades vizinhas. O *PKS-Stream I* (ZHANG et al., 2012) é uma versão otimizada do *PKS-Stream* quanto à seleção de período para detecção de densidade, identificação e remoção esporádica de grades.

*DCUStream* (YANG et al., 2012) é um algoritmo baseado em densidade e grades para aprendizado em FCDs com dados incertos. A proposta introduz o conceito de *core dense grid* que é uma grade densa com vizinhos esparsos, usados no momento de agrupamento na fase *offline*, quando os vizinhos esparsos são considerados ruído. O processo de busca pelos *core dense grids* e seus vizinhos pode ser bastante lento.

A proposta do algoritmo *DENGRIS-Stream* (AMINI; WAH, 2012) realiza o agrupamento dentro de um modelo de janela deslizante, na tentativa de capturar a distribuição mais recente dos dados. É a primeira proposta para agrupamento baseado em densidade e grades que considera o modelo de janela deslizante.

*ExCC* (BHATNAGAR; KAUR; CHAKRAVARTHY, 2013) é um algoritmo de agrupamento baseado em densidade e grades que tem como foco o aprendizado em FCDs heterogêneos. O algoritmo é robusto, adaptando-se a mudanças na distribuição dos dados e detectando *outliers* com rapidez. O algoritmo implementa um mecanismo para garantir o curso de novos grupos descobertos.

Pela revisão aqui exposta, pode-se verificar que é crescente o número de novas propostas

para agrupamento em FCDs, em especial aquelas que utilizam abordagem baseada em densidade e outras abordagens capazes de lidar com o surgimento e desaparecimento de grupos de forma simples.

No entanto, neste tipo de aprendizado, é ignorada qualquer informação prévia que possa existir a respeito da distribuição dos dados. O investimento em novas propostas para aprendizado semissupervisionado em FCD também cresceu nos últimos anos, como pode ser inferido pelas técnicas apresentadas na próxima seção.

### **3.3 Técnicas de Aprendizado Semissupervisionado em Fluxos Contínuos de Dados**

A busca por melhores resultados no aprendizado em FCD impulsionou o desenvolvimento de técnicas semissupervisionadas para trabalhar neste contexto. As abordagens de aprendizado semissupervisionado para conjuntos estáticos, juntamente com as abordagens de aprendizado supervisionado e não supervisionado em FCD servem de inspiração para as propostas descritas nas próximas seções.

#### **3.3.1 Classificação Semissupervisionada em FCD**

As técnicas de aprendizado semissupervisionado baseadas em classificadores assumem que o conjunto de dados é parcialmente rotulado. No caso desse tipo de aprendizado em FCD, a parte rotulada dos exemplos pode ser apenas um pequeno conjunto que irá gerar o modelo inicial de classificação ou é possível encontrar exemplos rotulados no próprio fluxo.

Considerando essas duas situações de disponibilidade de rótulos, podem ser encontradas na literatura propostas baseadas em diferentes métodos de classificação, por exemplo: redes neurais (LEITE; COSTA; GOMIDE, 2010; ASTUDILLO; OOMMEN, 2011, 2013; BOUGUELIA; BELAID; BELAID, 2013; KASABOV et al., 2013), baseada em grafos (TIWARI; KURHANEWICZ, 2010; BERTINI; LOPES; ZHAO, 2012; BERTINI; ZHAO, 2013), competição de partículas (BREVE; ZHAO, 2012, 2013), SVM (FRANDINA et al., 2013), entre outros (PAN; YANG; PAN, 2007; FDEZ-RIVEROLA et al., 2007; SILVA et al., 2011).

Uma proposta (LIANG et al., 2012) baseada em CVFDT considera que o FCD possui rótulo positivo e os dados incertos são não rotulados, realizando o aprendizado de forma semissupervisionada.

Técnicas baseadas em *ensemble* de classificadores também são populares no aprendizado



semissupervisionado em FCD. Kholghi e Keyvanpour (2011) apresentam uma proposta para um *framework* que combina semissupervisão por meio de *Active Learning* (SETTLES, 2010) e a consideração da influência de exemplos não rotulados a fim de melhorar a performance de aprendizagem. Um modelo é construído para predição de rótulos de exemplo futuros com alto valor de acurácia. Esse modelo de predição é baseado em um *ensemble* de classificadores construídos a partir de *chunks* de exemplos do FCD. Esta é uma das primeiras tentativas de incorporação de *Active Learning* semissupervisionado em FCD.

### 3.3.2 Agrupamento Semissupervisionado em FCD

No aprendizado semissupervisionado baseado em agrupamento, considera-se conhecimento prévio durante o processo de agrupamento para melhorar o aprendizado. Esta informação pode estar disponível em diferentes formas, por exemplo rótulos para parte do FCD, restrições entre pares de exemplos, informações estatísticas sobre a distribuição dos exemplos.

Quando há uma pequena quantidade de exemplos rotulados disponíveis, estes podem ser utilizados como sementes que contribuirão para guiar o algoritmo de agrupamento. O modelo de *flocking* serve de inspiração para a adaptação de um algoritmo de agrupamento para aprendizado em FCD (BRUNEAU; PICAROUGNE; GELGON, 2009), utilizando um pequeno conjunto de dados rotulados como informação para um operador de divisão de um grupo de exemplos, que permite a adaptação do agrupamento a mudanças no FCD.

Uma técnica de agrupamento semissupervisionado baseada em AP (SHI et al., 2009) utiliza informação prévia na forma de rótulos no ajuste da matriz de similaridade do modelo produzido e promove um estudo para ampliar o conjunto de dados rotulados. Baseado em *Fuzzy Pattern Matching* (CAYROL; FARRENY; PRADE, 1982; DUBOIS; PRADE; TESTEMALE, 1988), o método proposto em (MOUCHAWEH, 2010) tem o objetivo de aprender funções de pertinência com um conjunto de exemplos rotulados inicial limitado. A função de pertinência das classes é aprendida e atualizada, de acordo com a chegada de novos exemplos com rótulos.

O algoritmo *Compound Gaussian Mixture Model* (GAO; LIU; GAO, 2010), baseado no agrupamento *Gaussian Mixture Model*, utiliza amostra de dados rotulados em uma fase que aplica uma extensão do algoritmo *Expectation Maximization* (ZHOU et al., 2007) para melhorar o agrupamento. A técnica *Growing Type-2 Fuzzy Classifier* (BOUCHACHIA; VANARET, 2014) utiliza uma versão *online* do agrupamento *Gaussian Mixture Model* para gerar partições *fuzzy* tipo-2 (MENDEL; JOHN, 2002) para construir regras de classificação, empregando conjunto parcialmente rotulado no aprendizado. Esta técnica possui mecanismos para aprendizado em FCD.

Atwa e Li (2014) propõem um algoritmo de agrupamento semissupervisionado que estende o AP para lidar com FCD. Um conjunto de instâncias rotuladas é incorporado para detecção de mudança, que requer a atualização do modelo o mais rápido possível.

Em um contexto de FCD com multirrótulos, o *Hierarchical Semi-supervised Impurity based Subspace Clustering* (AHMED; KHAN; RAJESWARI, 2010) captura a correlação implícita existente entre cada par de rótulos de classe.

Técnicas que utilizam informação na forma de restrições podem obtê-las a partir de exemplos rotulados, mas também pode ser um conhecimento pré-existente nesse formato. O *C-DenStream* (RUIZ; MENASALVAS; SPILIOPOULOU, 2009) é uma técnica baseada no algoritmo *DenStream* adaptado para a utilização o conceito de restrições entre pares de exemplos estendido para FCD. O *C-DenStream* foi uma das primeiras extensões do paradigma de aprendizado por agrupamento semissupervisionado estático para FCD e, embora traga ganhos nesse contexto, ainda possui as limitações do *DenStream*.

Halkidi, Spiliopoulou e Pavlou (2012) utiliza, além do FCD, um fluxo contínuo de restrições, introduzindo o conceito de multigrupos (regiões densas e sobrepostas) e implementa mecanismo para identificação de *outlier*. Sirampuj, Kangkachit e Waiyamai (2013) propõem um algoritmo para agrupamento em FCD também com uso de conhecimento prévio na forma de restrições. A técnica, que é uma extensão do *E-Stream* (UDOMMANETANAKIT; RAKTHANMANON; WAIYAMAI, 2007) possui mecanismos para lidar com restrições que mudam de acordo com o tempo (técnica de esquecimento).

Cheng et al. (2011) desenvolvem um *framework* para análise de agrupamento de textos e desenvolvimento de um novo modelo de agrupamento semissupervisionado, capaz de lidar com informação prévia na forma de restrições entre pares de exemplos e rótulos de maneira simultânea.

Uma proposta de método para agrupamento em FCD incertos (domínios onde há ruído e dados incompletos) (AGGARWAL; YU, 2008) utiliza um modelo geral de incerteza, no qual assume-se que algumas estatísticas de incerteza estão disponíveis.

### 3.3.3 Aprendizado Semissupervisionado Híbrido em FCD

É possível identificar dentro do aprendizado semissupervisionado algumas abordagens híbridas, ou seja, inspiradas em métodos de agrupamento e classificação trabalhando em conjunto para melhorar o aprendizado.

Wu, Ye e Fu (2009) apresentam um método semissupervisionado para a construção de um rastreador de tópicos (*topical crawler*), aplicando um agrupamento *k-means* com restrições entre pares para detectar novas amostras de páginas enviadas a um classificador de páginas e preditor de links para atualização de modelos aprendidos.

A proposta de Borchani, Larranaga e Bielza (2011) é a combinação do método de (DASU; KRISHNAN, 2006) adaptado em um algoritmo de agrupamento para aprendizado semissupervisionado. O algoritmo de agrupamento é utilizado para atualização do modelo quando ocorre desvio de conceito.

Técnicas baseadas em AH podem utilizar métodos de agrupamento para divisão e rotulação em suas folhas. Li, Wu e Hu (2012) estende trabalho anterior (LI; WU; HU, 2010b) e propõem um algoritmo de classificação semissupervisionada em FCD, 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 grupos de conceito e rotulação automática de dados não rotulados. Potenciais desvios de conceito são identificados e conceitos recorrentes são mantidos. Uma técnica semelhante considera informação prévia na forma de rótulos e aplica uma versão semissupervisionada do algoritmo de agrupamento *k-modes* para produzir grupos de conceito (LI; WU; HU, 2010a; WU; LI; HU, 2012).

*Clustering Feature Decision Tree* (XU; QIN; CHANG, 2011) realiza a construção de uma árvore de decisão a partir de FCD parcialmente rotulados, aplicando algoritmo de agrupamento para gerar um vetor de atributos de grupos, sumários estatísticos que serão usados para indução da árvore de decisão. Os vetores de grupos também são empregados na classificação de exemplos nas folhas da árvore.

Zhang, Zhu e Guo (2009) propõem um *framework* para construção de modelos a partir de FCD com exemplos rotulados e não rotulados. Para a construção do modelo, os dados do FCD são associados a quatro categorias distintas, cada qual correspondendo à situação de desvio de conceito, podendo existir ou não nos exemplos rotulados e não rotulados. Em seguida, é aplicado método de aprendizado SVM semissupervisionado baseado no *k-means*.

A técnica *Concurrent Semi-supervised Learning of Data Streams* (NGUYEN et al., 2011; NGUYEN; NG; WOON, 2013) aplica o potencial de aprendizado semissupervisionado concorrente, onde um modelo de agrupamento e um de classificação são construídos de forma simultânea e colaborativa, fazendo uso de um pequeno conjunto de exemplos rotulados encontrados em um FCD.

Outras propostas utilizam aprendizado semissupervisionado com objetivo de apenas esten-

der o conjunto de exemplos rotulados e, então, aplicar método de aprendizado supervisionado com mecanismos disponíveis para lidar com as particularidades de FCD. Wu, Yang e Zhou (2006) coloca a proposta de um algoritmo de aprendizado semissupervisionado baseado em treinamento por agrupamento, para seleção de exemplos confiáveis a serem rotulados e utilizados no retreinamento de um classificador. A técnica proposta por (YU et al., 2009) aplica um algoritmo de agrupamento semissupervisionado aos exemplos parcialmente rotulados do FCD na tentativa de estender o conjunto de exemplos rotulados, utilizando-os para atualização de um modelo supervisionado que conta com mecanismos de esquecimento.

O *framework COMPOSE* (DYER; CAPO; POLIKAR, 2014) aprende desvios de conceitos em ambiente de FCD onde há apenas um conjunto inicial de dados rotulados e, após a inicialização, apenas dados não rotulados. O *COMPOSE* segue três 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 englobam os exemplos, 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.

### 3.3.4 Ensemble de Modelos para Aprendizado Semissupervisionado em FCD

Algumas propostas para aprendizado semissupervisionado em FCD tem a intenção de aproveitar da construção de diversos modelos trabalhando em conjunto para melhorar a representação dos exemplos do FCD.

O trabalho apresentado em (ZHANG et al., 2012) é uma adaptação do trabalho (ZHANG; ZHU; GUO, 2009), onde para cada categoria de exemplo de treinamento é construído um modelo distinto para classificação, baseado em SVM. Em (ZHANG et al., 2014) os modelos base para o *ensemble* são construídos por método de aprendizado semissupervisionado, utilizando conjunto de exemplos rotulados e não rotulados. Informação histórica é mantida como parte de peso no fator de decisão para classificação de novos exemplos.

O trabalho de Nahar et al. (2014b) propõe um *framework* para detecção de *cyberbullying* utilizando um classificador *ensemble* semissupervisionado. Em outra proposta (NAHAR et al., 2014a), a técnica utiliza inclui a extensão do conjunto de dados rotulados por meio de um classificador *ensemble*, com aplicação de um algoritmo *fuzzy SVM* para ponderar o espaço de atributos do domínio.

Outras técnicas têm a extensão do conjunto de exemplos rotulados como parte do processo de aprendizado semissupervisionado. Cao e He (2008) apresenta um algoritmo iterativo que recupera rótulos de acordo com níveis de confiança para melhorar o sistema aprendido pela geração de vários modelos de classificação. A técnica utilizada por Ahmadi e Beigy (2012) treina classificadores usando os exemplos rotulados e tenta classificar os exemplos não rotulados por meio do *ensemble* para estender o conjunto de treinamento e adaptar os modelos de classificação.

Os trabalhos (MASUD et al., 2008a) e (MASUD et al., 2008b) descrevem uma proposta baseada na construção de microgrupos pela aplicação de método de agrupamento semissupervisionado e construção de classificadores pelo algoritmo *K-Nearest Neighbors* para cada *chunk* de exemplos do FCD. Os  $L$  melhores modelos (de acordo com acurácia individual) são utilizados em um *ensemble*.

A proposta de Ditzler e Polikar (2011) apresenta um *ensemble* onde classificadores são treinados a partir dos exemplos rotulados do FCD. O modelo de classificação utiliza pesos para determinar a influência de cada classificador na decisão final e esses pesos são determinados pela distância entre componentes de um *Gaussian Mixture Model* treinado com o conjunto completo de exemplos.

Masud et al. (2012) propõem um *ensemble* onde cada modelo de classificação é construindo como uma coleção de microgrupos, usando agrupamento semissupervisionado, e exemplos não rotulados são classificados de acordo com o conjunto de modelos.

Uma proposta (LIU et al., 2013) mantém um *ensemble* de modelos mistos, baseados em métodos de classificação e agrupamento. Os exemplos rotulados são utilizados para treinamento de classificador supervisionado e novos exemplos rotulados são empregados na atualização desse classificador. Os exemplos não rotulados são utilizados na construção de modelos não supervisionado. O *ensemble* segue um modelo semissupervisionado de classificação de forma a maximizar o consenso entre os diferentes modelos.

### 3.4 Considerações Finais

Neste capítulo foram colocadas algumas técnicas de aprendizado em FCD. A maioria das propostas sugere adaptações para métodos de aprendizado em conjuntos estáticos, a fim de incluir mecanismos que possam lidar com as limitações de aprendizado em FCD.

Recentemente, percebe-se uma preocupação em elaborar técnicas que possam realizar o

aprendizado de forma semissupervisionada, uma vez que a maior parte dos domínios não possuem uma quantidade grande de rótulos. O modelo supervisionado aprendido apenas pelo pequeno conjunto de exemplos rotulados pode ser ineficiente, enquanto modelos aprendidos de forma não supervisionada perdem a chance de melhorar o resultado pela consideração de informação prévia sobre o FCD.

As técnicas apresentadas neste capítulo utilizam métodos variados de aprendizado de máquina e implementam diferentes mecanismos para atacar os problemas característicos de aprendizado em FCD. Alguns trabalhos sugerem adaptações e extensões para melhorar um ou outro aspecto de um método. De forma geral, a forma de como lidar com o tempo (mecanismos de esquecimento) e detecção de desvio de conceitos são duas tarefas relevantes que não são consideradas por todas as técnicas de aprendizado em FCD. De qualquer forma, a revisão colocada no capítulo contribui para uma visão geral do estado da arte quanto ao aprendizado em FCD.

O Capítulo 4 discute a proposta de trabalho para elaboração da tese de doutorado dentro do tema de aprendizado de máquina semissupervisionado em FCD.

# Capítulo 4

## PROPOSTA DE TRABALHO

---

Tendo em conta a revisão bibliográfica apresentada neste documento, este capítulo apresenta a proposta de trabalho para desenvolvimento de tese abordando o tema de aprendizado em FCD. Além de descrever os temas mais relevantes que serão tratados na proposta, este capítulo também contém a descrição das atividades principais que serão realizadas e o cronograma geral para sua execução. Para finalizar, são apontadas as contribuições que se almeja alcançar.

O domínio de FCD possui características particulares que tornam inviável a aplicação de métodos clássicos de aprendizado. Entre essas características, pode-se citar as restrições de tempo e espaço, a chegada de dados em fluxo contínuo potencialmente infinito e a possibilidade de mudança na distribuição dos dados. O estudo e utilização de mecanismos que possam atacar os desafios inerentes a esses domínios é de grande importância para que a técnica de aprendizado aplicada seja bem sucedida.

O aprendizado em FCD tem sido abordado por meio de mecanismos supervisionados e não supervisionados. Métodos de aprendizado semissupervisionado tem sido aplicados a domínios em que estão disponíveis uma grande quantidade de dados não rotulados e poucos dados rotulados. Uma vez que muitos problemas reais de FCD possuem essa característica, o aprendizado semissupervisionado apresenta-se como um mecanismo adequado e abre um novo leque de possibilidades para tratar dos problemas da área.

A proposta geral deste trabalho é o estudo, a análise e a geração de modelo de aprendizado por meio de técnica semissupervisionada tendo como classe de problema abordado o domínio de fluxos contínuos de dados. De forma geral, o objetivo dessa proposta pode ser apresentado em duas vertentes complementares:

- Investigação de mecanismos específicos para lidar com aspectos característicos de FCD, incluindo mas não limitados a

- Restrições de tempo e espaço: em um domínio FCD os dados chegam de forma contínua e constante, o armazenamento de todos os exemplos para posterior processamento é inviável e, se não processados logo, os exemplos são perdidos;
  - Esquecimento e histórico: em aplicações FCD os exemplos mais recentes tem maior importância, mas os exemplos antigos não devem ser totalmente descartados à chegada de novos;
  - Detecção de desvios de conceito e adaptação de modelo: aplicações FCD podem se estender por um longo período de tempo e, durante esse período, podem ocorrer mudanças na distribuição dos exemplos. Essas mudanças devem ser detectadas para acionar a adaptação do modelo do FCD;
  - Conhecimento sob demanda: um algoritmo de aprendizado em FCD deve ser capaz de produzir um modelo representativo do conjunto de exemplos periodicamente, seja por requisição do usuário ou em intervalos de tempos pré-definidos;
- Geração de modelo de distribuição de dados de um FCD por meio de agrupamento semi-supervisionado, considerando os mecanismos específicos para lidar com aspectos característicos de FCD.

As próximas seções apresentam com maiores detalhes os tópicos principais que serão tratados no trabalho proposto e o cronograma para sua execução.

## 4.1 Tópicos Principais

A proposta de trabalho apresentada aqui envolve o estudo de dois temas relevantes em aprendizado de máquina: o aprendizado em fluxo contínuo de dados e o aprendizado semi-supervisionado. A seguir, são apresentados resumidamente os principais tópicos relacionados em esses temas, que serão focalizados durante o desenvolvimento da pesquisa.

### 4.1.1 Aprendizado em Fluxos Contínuos de Dados

O aprendizado em FCD possui características bem definidas, que o diferenciam do aprendizado tradicional. Assim, o estudo de técnicas voltadas para aprendizado em FCD deve incluir o tratamento das questões típicas desse campo de pesquisa.



#### 4.1.1.1 Limitações de Tempo e Espaço

Dados em fluxos contínuos são gerados em volume elevado e de forma contínua, potencialmente infinita. Assim, torna-se inviável armazenar os dados em estruturas de dados em memória, para que fiquem disponíveis para acesso pelo algoritmo de aprendizado, como ocorre no aprendizado tradicional. A medida que se tornam disponíveis para o algoritmo, os dados podem ser armazenados por algum tempo e devem, posteriormente, ser descartados. No aprendizado em fluxo contínuo é necessário tratar a questão de representação dos dados por meio de estruturas alternativas, que não necessitem do armazenamento do dado completo.

#### 4.1.1.2 Esquecimento

Devido às características do FCD, os métodos de aprendizado não têm acesso ao conjunto de dados completo. Dessa forma, é necessário definir mecanismos para delimitar a porção dos dados que será considerada a cada passo e a maneira como essa porção deve ser atualizada. Geralmente são utilizadas técnicas que descartam os dados mais antigos para liberar espaço para dados novos. Esse tipo de mecanismo é chamado de mecanismo de esquecimento.

#### 4.1.1.3 Detecção de Desvio de Conceito e Adaptação de Modelo

A possibilidade de ocorrência de mudança na distribuição dos dados de um fluxo contínuo é um dos maiores desafios encontrados pelos pesquisadores que investigam esse tipo de aprendizado. Diferentemente do que ocorre no aprendizado clássico, em que o conjunto de dados é finito e assume-se que tem uma distribuição definida e fixa, no aprendizado em FCD é necessário identificar e tratar as eventuais mudanças no comportamento do conjunto de dados. A mudança na distribuição dos dados de um FCD é conhecida como desvio de conceito. A capacidade das técnicas de aprendizado de detectar tais mudanças e prover mecanismos para adaptação do modelo sendo gerado é fundamental para o sucesso dessas técnicas.

### 4.1.2 Aprendizado Semissupervisionado

É comum encontrar no mundo real conjuntos de dados que contam com apenas uma pequena parcela de dados rotulados, ou algum outro tipo de informação sobre a distribuição dos dados no conjunto. Nesses casos, a aplicação de aprendizado de semissupervisionado pode ser destacada pela capacidade de utilizar de informação prévia e um conjunto de dados não rotulados de maneira conjunta para criar um modelo. Trabalhos anteriores (LOPES; CAMARGO,

2011, 2012) realizados dentro do tema mostram a capacidade desses métodos.

Abordagens semissupervisionadas vem se firmando dentro da área de aprendizado de máquina. Inicialmente, os métodos que consideram semissupervisão estavam associados a um conjunto de dados estático e total. Nesse contexto, o processo de aprendizagem ocorre de forma similar ao de métodos tradicionais de aprendizado.

A falta de rótulos pode ser uma característica comum quando fala-se de dados reais de fluxo contínuo. O crescimento do conjunto de dados, muitas vezes em velocidade alta, impede a rotulação total e manual dos dados para serem aplicados a métodos de aprendizado supervisionado. Por outro lado, pode ser interessante incorporar alguns dados rotulados ou informação extra sobre o conjunto e dados não rotulados para obter melhores resultados para o modelo aprendido e, em se tratando de FCD, a adaptação desse modelo.

É necessário a definição de uma abordagem para o aprendizado semissupervisionado e de como será considerada a informação prévia.

Uma vez que pretende-se aplicar o método a dados de fluxo contínuo, além das questões particulares de aprendizado semissupervisionado, deve-se atentar para a inclusão de mecanismos que lidem com as dificuldades inerentes ao domínio de FCD, em especial os descritos na seção anterior.

## 4.2 Cronograma de Atividades

Considerando a proposta apresentada, os problemas a serem tratados e exigências do Programa de Pós-Graduação em Ciência da Computação, são especificadas as atividades a fim de elaborar um cronograma para a execução, avaliação e finalização da tese de doutorado:

1. Revisão Bibliográfica: revisão bibliográfica sobre o tema do projeto, buscando reunir, estudar e organizar as diversas publicações a respeito do assunto;
2. Desenvolvimento da abordagem - Investigação de mecanismos para tratamento de:
  - (a) Restrições de tempo e espaço;
  - (b) Esquecimento dos dados mais antigos e atribuição de relevância maior aos dados mais recentes;
  - (c) Detecção de desvio de conceito e adaptação do modelo ao novo conceito;

3. Desenvolvimento da abordagem - Proposta de procedimento para geração de modelo de distribuição de dados de um FCD usando aprendizado semissupervisionado, considerando os mecanismos investigados no passo anterior;
4. Implementação e análise dos resultados:
  - (a) Geração de bases de dados artificiais e seleção de bases de dados de repositórios para usar nos experimentos;
  - (b) Implementação ou adaptação de implementação do método proposto, seguido da execução com os conjuntos de dados definidos no item anterior;
  - (c) Análise dos resultados;
5. Redação da tese: escrita da tese de doutorado e defesa perante banca examinadora.
6. Publicações: Escrita e submissão de artigos para publicação dos resultados obtidos.

A Tabela 4.1 traz a alocação de tempo para a execução das atividades listadas nesta seção, em um cronograma a partir do mês de março de 2015 até a data prevista para defesa da tese, dezembro de 2016. Cada marca (X) corresponde a um mês..

**Tabela 4.1: Cronograma de atividades a partir de março de 2015**

	2015										2016											
Atividade	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
2		X	X	X	X	X	X	X	X	X												
3								X	X	X	X	X	X	X								
4										X	X	X	X	X	X	X	X	X				
5																		X	X	X	X	
6										X	X				X	X					X	X

### 4.3 Contribuições Esperadas

Com este trabalho espera-se obter resultados científicos que contribuirão para o desenvolvimento da área de Aprendizado Semissupervisionado em domínios de Fluxos Contínuos de Dados e para o crescimento e amadurecimento do grupos de pesquisa, como:

- Proposta de técnica para aprendizado a partir de exemplos em FCD baseada em aprendizado semissupervisionado, incluindo mecanismos específicos para lidar com algumas

das limitações e particularidades deste tipo de domínio, tais como: esquecimento, desvio de conceito, restrições de tempo e espaço;

- Documentação e ambiente de experimentação para aplicação da técnica proposta a diferentes problemas a fim de:
  - Complementar a formação de outros membros do grupo de pesquisa no tema;
  - Possibilitar a reprodução de experimentos por outros pesquisadores da área de forma simplificada.
- Publicação da caracterização da técnica proposta e resultados de experimentos em veículos de reconhecida relevância;
- Aplicação potencial da técnica proposta a domínios para resolução de problemas reais, promovendo convênios com outros pesquisadores de áreas diversificadas;

## 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. A Framework for Clustering Evolving Data Streams. In: *Proceedings of the 29th International Conference on Very Large Data bases*. [S.l.: s.n.], 2003. v. 29, p. 81–92.
- 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.
- AHMADI, Z.; BEIGY, H. Semi-supervised Ensemble Learning of Data Streams in the Presence of Concept Drift. In: *Hybrid Artificial Intelligent Systems*. [S.l.]: Springer Berlin Heidelberg, 2012. p. 526537.
- 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. DENGRIIS-Stream: A Density-Grid based Clustering Algorithm for Evolving Data Streams over Sliding Window. In: *International Conference on Data Mining and Computer Engineering*. [S.l.: s.n.], 2012. p. 206–210.
- 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.; OOMMEN, B. J. 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.
- ATASHPAZ-GARGARI, E.; LUCAS, C. Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. In: *2007 IEEE Congress on Evolutionary Computation, CEC 2007*. [S.l.: s.n.], 2007. p. 4661–4667.
- 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.
- BASU, S.; BANERJEE, A.; MOONEY, R. Semi-supervised Clustering by Seeding. In: *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*. [S.l.: s.n.], 2002. p. 27–34.
- BASU, S.; MOONEY, R. J. Active Semi-Supervision for Pairwise Constrained Clustering. In: *Proceedings of the SIAM International Conference on Data Mining (SDM-2004)*. [S.l.: s.n.], 2004. p. 333–344.
- BENSAID, A. M.; BEZDEK, J. C. Semi-Supervised Point Prototype Clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 12, n. 5, p. 625–643, 1998.
- BENSAID, A. M. et al. Partially supervised clustering for image segmentation. *Pattern Recognition*, v. 29, p. 859–871, 1996.
- 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.
- BIFET, A. et al. MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering. In: *Workshop on Applications of Pattern Analysis*. [S.l.: s.n.], 2010. v. 11, p. 44–50.
- 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.; LARRANAGA, P.; 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.
- CAYROL, M.; FARRENY, H.; PRADE, H. Fuzzy Pattern Matching. *Kybernetes*, v. 11, n. 2, p. 103–116, 1982.
- CHAIRUKWATTANA, R. et al. SE-Stream: Dimension Projection for Evolution-Based Clustering of High Dimensional Data Streams. In: *Knowledge and Systems Engineering*. [S.l.]: Springer International Publishing, 2014. p. 365–376.
- 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.

- CHEN, Y.; TU, L. Density-based clustering for real-time stream data. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 07*. [S.l.: s.n.], 2007. p. 133–142.
- 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 Research*, v. 7, p. 551–585, 2006.
- DASU, T.; KRISHNAN, S. An information-theoretic approach to detecting changes in multi-dimensional data streams. In: *Proceedings of the Symposium on the Interface of Statistics, Computing Science, and Applications*. [S.l.: s.n.], 2006. p. 1–24.
- 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.
- DUBOIS, D.; PRADE, H.; TESTEMALE, C. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, v. 28, n. 3, p. 313–331, 1988.
- 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.
- FORESTIERO, A.; PIZZUTI, C.; SPEZZANO, G. A single pass algorithm for clustering evolving data streams based on swarm intelligence. *Data Mining and Knowledge Discovery*, v. 26, p. 1–26, 2013. ISSN 13845810.
- 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.
- GAMA, J.; GABER, M. M. (Ed.). *Learning from Data Streams: Processing Techniques in Sensor Networks*. [S.l.]: Springer, 2007.
- GAMA, J. a. *Knowledge Discovery from Data Streams*. [S.l.]: Chapman and Hall, 2010. 255 p.
- GAMA, J. a. et al. Learning with drift detection. In: *Advances in Artificial Intelligence (SBIA2004)*. [S.l.: s.n.], 2004. p. 286–295.
- GAMA, J. a.; RODRIGUES, P. P.; LOPES, L. Clustering Distributed Sensor Data Streams Using Local Processing and Reduced Communication. *Intelligent Data Analysis*, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 15, n. 1, p. 3–28, 2011.
- 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.
- GENG, C. et al. An algorithm of semi-supervised web-page classification based on fuzzy clustering. In: *Proceedings - 2009 International Forum on Information Technology and Applications, IFITA 2009*. [S.l.: s.n.], 2009. v. 1, p. 3–7.
- 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.
- GRIRA, N.; CRUCIANU, M.; BOUJEMAA, N. Active semi-supervised fuzzy clustering for image database categorization. In: *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval - MIR '05*. [S.l.: s.n.], 2005. p. 9–16.

- GRIRA, N.; CRUCIANU, M.; BOUJEMAA, N. Fuzzy clustering with pairwise constraints for knowledge-driven image categorisation. In: *IEE Proceedings - Vision, Image, and Signal Processing*. [S.l.: s.n.], 2006. v. 153, p. 299–304.
- GRIRA, N.; CRUCIANU, M.; BOUJEMAA, N. Active semi-supervised fuzzy clustering. *Pattern Recognition*, v. 41, n. 5, p. 1851–1861, 2008.
- GUHA, S. et al. Clustering data streams. In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. [S.l.: s.n.], 2000. p. 359–366.
- 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).
- HASSANI, M. et al. Density-Based Projected Clustering of Data Streams. In: *Scalable Uncertainty Management*. [S.l.: s.n.], 2012. p. 311–324.
- 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, 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.; DOMINGOS, P. *VFML - A toolkit for mining high-speed time-changing data streams*. 2003. Disponível em: <<http://www.cs.washington.edu/dm/vfml/>>.
- 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.

- ISAKSSON, C.; DUNHAM, M. H.; HAHSLER, M. SOStream: Self Organizing Density-Based Clustering over Data Stream. In: *Machine Learning and Data Mining in Pattern Recognition*. [S.l.: s.n.], 2012. v. 7376, p. 264–278.
- 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, WGECC 2008*. [S.l.: s.n.], 2008. p. 517–521.
- JIANG, N.; GRUENWALD, L. Research issues in data stream association rule mining. *ACM SIGMOD Record*, v. 35, p. 14–19, 2006.
- JINXIAN, L.; HUI, L. A density-based clustering over evolving heterogeneous data stream. In: *2009 Second ISECS International Colloquium on Computing, Communication, Control, and Management, CCCM 2009*. [S.l.: s.n.], 2009. v. 4, p. 275–277.
- 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.
- KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, v. 43, p. 59–69, 1982.
- KRANEN, P. et al. The ClusTree: indexing micro-clusters for anytime stream mining. *Knowledge and Information Systems*, Springer-Verlag, v. 29, n. 2, p. 249–272, 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.
- LABZOUR, N.; BENSAID, A.; BEZDEK, J. Improved semi-supervised point-prototype clustering algorithms. *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence*, v. 2, p. 1383–1387, 1998.
- 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.
- LEMONS, 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.
- 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.
- LI-XIONG, L. L.-x. L. et al. A three-step clustering algorithm over an evolving data stream. In: *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*. [S.l.: s.n.], 2009. v. 1, p. 160–164.
- 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, H.; HUANG, S.-t. A Genetic Semi-supervised Fuzzy Clustering Approach to Text Classification. In: *Advances in Web-Age Information Management*. [S.l.: s.n.], 2003. p. 173–180.
- 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.
- LOPES, P.; CAMARGO, H. Semi-supervised Clustering in Fuzzy Rule Generation. In: *CSBC 2011*. [S.l.: s.n.], 2011.
- LOPES, P. D. A.; CAMARGO, H. D. A. Automatic labeling by means of semi-supervised fuzzy clustering as a boosting mechanism in the generation of fuzzy rules. In: *Information Reuse and . . .* [S.l.: s.n.], 2012. p. 279–286. ISBN 9781467322843.
- 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.
- 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.
- MEESUKSABAI, W.; KANGKACHIT, T.; WAIYAMAI, K. HUE-Stream: Evolution-Based Clustering Technique for Heterogeneous Data Streams with Uncertainty. In: *Advanced Data Mining and Applications*. [S.l.]: Springer Berlin Heidelberg, 2011. p. 27–40.
- MENDEL, J. M.; JOHN, R. I. B. Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, v. 10, p. 117–127, 2002. ISSN 10636706.
- 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.
- NAHAR, V. et al. Detecting cyberbullying in social networks using multi-agent system. *Web Intelligence and Agent Systems*, v. 12, n. 4, p. 375–388, 2014.
- 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.-I. et al. Concurrent Semi-supervised Learning of. In: *Data Warehousing and Knowledge Discovery*. [S.l.]: Springer Berlin Heidelberg, 2011. p. 445–459.
- NTOUTSI, I. et al. Density-based Projected Clustering over High Dimensional Data Streams. In: *Proceedings of the third SIAM International Conference on Data Mining*. [S.l.: s.n.], 2012. p. 987–998.
- O'CALLAGHAN, L. et al. Streaming-data algorithms for high-quality clustering. In: *Proceedings 18th International Conference on Data Engineering*. [S.l.: s.n.], 2002. p. 685–694.
- 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. et al. Fuzzy clustering with partial supervision in organization and classification of digital images. *IEEE Transactions on Fuzzy Systems*, v. 16, n. 4, p. 1008–1026, 2008. ISSN 10636706.
- PEDRYCZ, W.; GOMIDE, F. *An Introduction to Fuzzy Sets: Analysis and Design*. [S.l.]: MIT Press, 1998. (A Bradford book).
- PEDRYCZ, W.; WALETZKY, J. Fuzzy clustering with partial supervision. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, v. 27, n. 5, p. 787–795, 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.
- 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.
- REN, J.; MA, R. Density-based data streams clustering over sliding windows. In: *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009*. [S.l.: s.n.], 2009. v. 5, p. 248–252.
- 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.
- SETTLES, B. *Active Learning Literature Survey*. [S.l.], 2010. 65 p. Disponível em: <<http://burrsettles.com/pub/settles.activelearning.pdf>>.

- 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.
- SHEIKHOLESAMI, 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.
- TASOULIS, K. D.; ROSS, G.; ADAMS, N. M. Visualising the Cluster Structure of Data Streams. In: *Advances in Intelligent Data Analysis*. [S.l.: s.n.], 2007. p. 81–92.
- 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.
- UDOMMANETANAKIT, K.; RAKTHANMANON, T.; WAIYAMAI, K. E-Stream: Evolution-Based Technique for Stream Clustering. In: *Advanced Data Mining and Applications*. [S.l.]: Springer Berlin Heidelberg, 2007. p. 605–615.
- 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.

- WAGSTAFF, K. et al. Constrained K-means Clustering with Background Knowledge. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. [S.l.: s.n.], 2001. p. 577–584.
- 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.
- YAN, Y.; CHEN, L. Label-based semi-supervised fuzzy co-clustering for document categorization. In: *2011 8th International Conference on Information, Communications & Signal Processing*. [S.l.]: IEEE, 2011. p. 1–5.
- YANG, C. Y. C.; ZHOU, J. Z. J. HClustream: A Novel Approach for Clustering Evolving Heterogeneous Data Stream. In: *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*. [S.l.: s.n.], 2006. p. 682–688.
- YANG, Y. et al. Dynamic density-based clustering algorithm over uncertain data streams. In: *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*. [S.l.: s.n.], 2012. p. 2664–2670.
- 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, B. et al. A New Semi-supervised Learning Based Ensemble Classifier for Recurring Data Stream. In: *Pervasive Computing and the Networked World*. [S.l.]: Springer International Publishing, 2014. v. 8351, p. 759–765.
- 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 Discovery Database*. [S.l.]: Springer Berlin Heidelberg, 2008. p. 628–643.
- ZHENPENG, L. et al. An Improved Semi-supervised K-Means Algorithm Based on Information Gain. In: *2014 IEEE 17th International Conference on Computational Science and Engineering*. [S.l.: s.n.], 2014. p. 1960–1963. ISBN 978-1-4799-7981-3.
- ZHOU, A. et al. Tracking clusters in evolving data streams over sliding windows. *Knowledge and Information Systems*, v. 15, p. 181–214, 2008.
- ZHOU, A. et al. Distributed data stream clustering: A fast EM-based approach. In: *Proceedings - International Conference on Data Engineering*. [S.l.: s.n.], 2007. p. 736–745.
- ZHU, X.; GOLDBERG, A. B. *Introduction to Semi-Supervised Learning*. [S.l.: s.n.], 2009. 130 p.

---

ZHU, Y.; SHASHA, D. StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. *Proceedings of the 28th international conference on Very Large Data Bases*, v. 54, p. 358–369, 2002.