22, 23 E 24
DE OUTUBRO

XXXII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA PUCPR

CIÊNCIA, ENERGIA E SOCIEDADE

SOLUÇÕES E PERSPECTIVAS INOVADORAS

INICIAÇÃO CIENTÍFICA E TECNOLÓGICA PUCPR

PROMOÇÃO E REALIZAÇÃO

PUCPR GRUPO MARISTA

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

ARAUCÁRIA

XXXII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA PUCPR

AVALIAÇÃO DO IMPACTO DE TÉCNICAS DE SELEÇÃO DE INSTÂNCIAS EM ENSEMBLES ORIENTADOS A FLUXOS DE DADOS

Autor: Vitor Rodrigues Izidoro
Orientador: Prof. Fabricio Enembreck
Curso: Ciência da Computação
Câmpus: Curitiba

CNPq
PIBIC
PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTIFÍCA

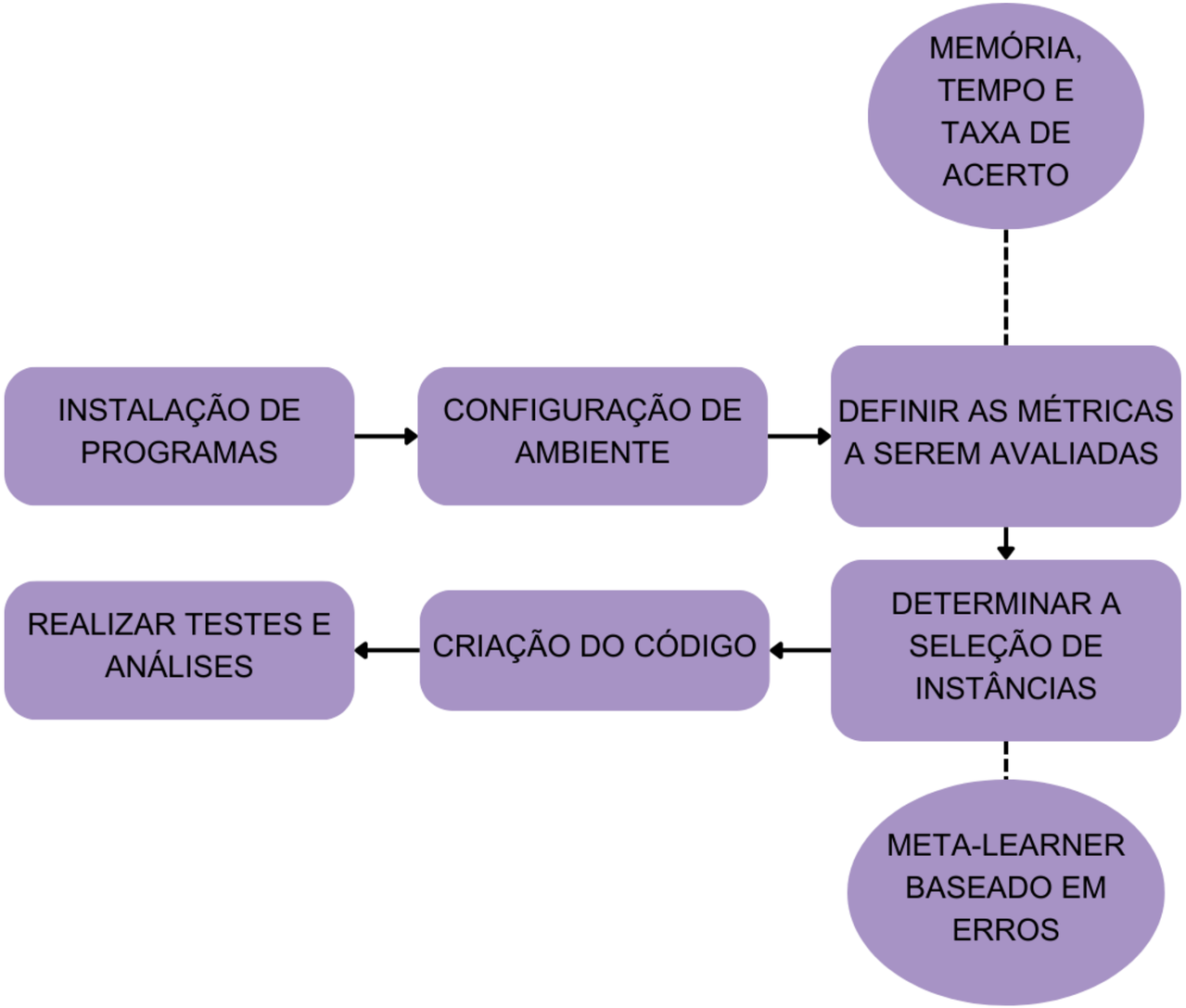
INTRODUÇÃO

- Com o aumento dos dados gerados em tempo real por dispositivos móveis e sensores, a mineração de fluxo de dados torna-se cada vez mais importante. Esta pesquisa explora como técnicas de Seleção de Instâncias podem reduzir a complexidade computacional na classificação de fluxos de dados.

OBJETIVOS

- Avaliar e desenvolver técnicas de seleção de instâncias para ensembles orientados a fluxos de dados, viabilizando a aplicação desses algoritmos em cenários de larga escala.

METODOLOGIA



METODOLOGIA

O código principal para este projeto foi desenvolvido utilizando como inspiração o código do DriftDetectionMethodClassifier, .java pois a estrutura de seleção do learner, escolha de arff's externos já havia sido desenvolvida. A ideia base para a criação do código era que tal seleção de instâncias teria como objetivo reduzir a quantidade de instâncias que seriam utilizadas no treinamento e aumentar a taxa de acerto dos Algoritmos.

Classe InstanceSelectedClassifier

variáveis:

classifier: classificador atual;

Função resetLearningImpl():

Inicializa 'classifiers' como novo classificador;

Reseta aprendizagem de 'classifier';

Função trainOnInstanceImpl(instância):

Se classe verdadeira da instância igual classe prevista pelo classificador:

Não faça nada;

Senão:

Treine o classificador na instância;

Função getVotesForInstance(instância):

Retorna os votos do classificador para a instância;

DISCUSSÃO DE RESULTADOS

Tabela 01 – Tabela de ranking da taxa de acerto

Dataset	Levering	IS-levering	ARF	IS-ARF	OzaBoost	IS-OzaBoost	SRP	IS-SRP
Airlines	4	8	2	6	3	7	**1**	5
Kdd99	7	6	4	3	8	5	2	**1**
Kddcup	5	4	3	8	7	6	**1**	2
Keystroke	5	8	2	4	6	7	**1**	3
Luxembourg	4	6	**1**	2	7	8	3	5
NOAA	3	5	**1**	6	4	8	2	7
Nomao	6	5	2	3	8	7	**1**	4
Outdoor	7	5	4	3	8	6	2	**1**
Ozone	2	5	**1**	7	4	8	3	6
Poker	**1**	4	6	3	7	8	5	2
Rialto	6	5	4	3	8	7	**1**	2
Média:	4,545	5,54	2,72	4,36	6,36	7,00	2,00	3,45

- Como podemos analisar na tabela ao lado, a média geral de acerto de algoritmos com a seleção de instâncias é pior que seus algoritmos originais sem a seleção e tendo apenas excessões a este padrão em poucos datasets testados (para análise, quanto mais próximo de 1 melhor).

DISCUSSÃO DE RESULTADOS

Tabela 02 – Tabela de ranking de tempo

Dataset	levering	IS-levering	ARF	IS-ARF	OzaBoost	IS-OzaBoost	SRP	IS-SRP
Airlines	7	5	6	3	2	1	8	4
Kdd99	2	3	5	7	1	4	8	6
Kddcup	6	7	4	3	1	2	8	5
Keystroke	6	3	7	4	2	1	8	5
luxembourg	7	3	6	4	2	1	8	5
NOAA	6	3	7	4	2	1	8	5
nomao	6	3	7	4	2	1	8	5
outdoor	7	3	6	4	1	2	8	5
ozone	7	3	6	4	2	1	8	5
poker	6	3	7	4	2	1	8	5
rialto	6	3	7	4	2	1	8	5
Média:	6	3,54	6,18	4,09	1,72	1,45	8	5

- Agora, analisando a tabela de tempo médio, podemos concluir que em todas as médias gerais os algoritmos, a seleção de instâncias teve um tempo melhor que todos os algoritmos originais (para análise, quanto mais próximo de 1 melhor).

DISCUSSÃO DE RESULTADOS

- Nos casos testados, a seleção de instâncias sacrifica o percentual de acerto em troca de uma melhoria no custo computacional e na velocidade de treinamento. A escolha de realizar a seleção de instâncias depende muito da necessidade de cada projeto: se for necessário um processamento mais rápido e com menor custo computacional, a seleção de instâncias seria uma solução viável.

CONSIDERAÇÕES FINAIS

- Os ensembles, apesar de serem precisos, são lentos e pesados para aplicações em larga escala. A seleção de instâncias, por sua vez, reduz o consumo de recursos computacionais, mas sacrifica a taxa de acerto. É necessário mais desenvolvimento em técnicas de seleção que minimizem essa perda de precisão, vendo que a simplicidade da técnica avaliada potencializou as limitações.

XXXII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA PUCPR

AVALIAÇÃO DO IMPACTO DE TÉCNICAS DE SELEÇÃO DE INSTÂNCIAS EM ENSEMBLES ORIENTADOS A FLUXOS DE DADOS

Pesquisa apresentada por Vitor Rodrigues Izidoro com orientação de Fabricio Enembreck

Curitiba 2024

CNPq PIBIC PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTIFÍCA