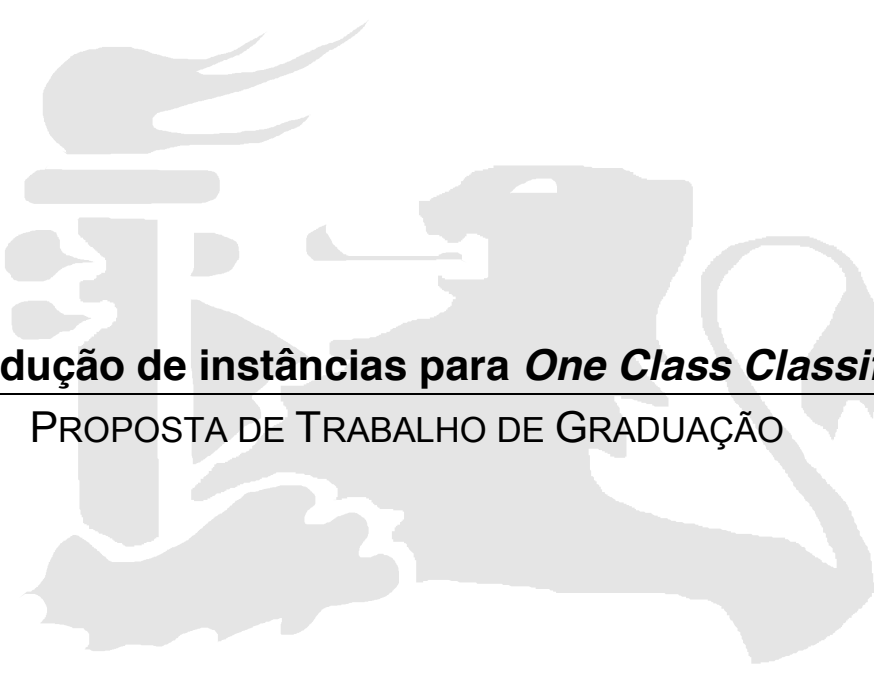




Redução de instâncias para *One Class Classification*

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Hoi Yan Mark (hym@cin.ufpe.br)

Orientador: George Darmiton da Cunha Cavalcanti (gdcc@cin.ufpe.br)

Área: Aprendizagem de Máquina

05 de Setembro de 2018

Resumo

Em certas situações, bancos de dados exigem grande espaço para armazenamento e acarretam longos períodos para treinamento de classificadores. Técnicas que realizam seleção de protótipos são uma alternativa para contornar tais problemas. Entretanto, no contexto de *One Class Classification*, há circunstâncias que impossibilitam aplicação direta dessas técnicas no *dataset*. Neste trabalho, em que o foco é este último cenário descrito, emprega-se técnica de geração de *outliers* em bancos de dados de *One Class Classification*, com o intuito de permitir atuação de técnicas que reduzam o número de instâncias, juntamente a classificadores que atuarão nos protótipos selecionados para obtenção de métricas, os quais auxiliarão na avaliação dos resultados.

Introdução

No escopo de *One Class Classification* (OCC), bancos de dados são compostos por instâncias da classe *target* e instâncias da classe *outlier*. Os objetos que apresentam características que sejam o foco do banco de dados pertencem à classe *target*, enquanto que a classe *outlier* é composta pelos objetos restantes. Durante a fase de treinamento, é suposto que há exclusivamente instâncias da classe *target*, tornando isto uma característica que distingue o OCC de problemas de classificação convencionais.

Entretanto, nessa fase de treinamento, a abundância de instâncias pode às vezes se apresentar como empecilho. Do ponto de vista dos problemas de classificação convencionais, uma alternativa para contornar esse problema seria por meio de técnicas de seleção de protótipos e técnicas de geração de protótipos, além de técnicas capazes de realizar estas juntamente, cujos objetivos são reduzir o número de instâncias presentes visando a manter a representatividade de cada classe no espaço de características sendo estudado. Estas técnicas exigem a presença de instâncias de, no mínimo, 2 classes distintas para seu funcionamento, o que prejudica sua aplicação em *datasets* de OCC, visto que problemas de OCC dispõem de apenas instâncias de uma única classe durante o processo de treino, como descrito anteriormente.

Como alternativa para este problema, *outliers* podem ser gerados para compensar a falta de instâncias de uma segunda classe. Analisando as informações dos objetos no espaço de características, é possível gerar *outliers* seguindo a hipótese [1] de que suas instâncias se apresentam no espaço obedecendo a uma distribuição normal. No entanto, teoricamente, a quantidade de objetos a ser gerada poderia crescer numa taxa absurda à medida que aumenta o número de características e haveria também o risco de *outliers* serem gerados em espaços já povoados por *targets*.

Neste trabalho, é estudado os efeitos da aplicação de técnica para geração de objetos [2], em que, considerado a distribuição das instâncias da classe *target*, utiliza a distribuição complementar a esta para geração de instâncias da classe *outlier*, prevenindo tanto uma grande quantidade de objetos a serem criados quanto a superposição de instâncias de classes distintas.

Objetivos

Este trabalho objetiva analisar o desempenho da aplicação de técnica para geração de objetos em bancos de dados. Esta análise será feita por meio das seguintes informações: *balanced accuracy metric* (BAC), taxa de redução após aplicação de algoritmos de seleção de protótipos em *datasets* com os *outliers* gerados, e tempo de classificação utilizado por classificadores de *One Class Classification*, estes são aplicados tanto sobre bancos de dados que tenham sofrido redução no seu número de instâncias quanto sobre aqueles que não passaram por este processo.

Metodologia

A metodologia adotada para este trabalho consiste nos seguintes processos: geração de *outliers* para 10 *datasets* do Repositório de Aprendizagem de Máquina da UCI, redução de número de instâncias por 2 métodos de seleção de protótipos e aplicação de 3 classificadores de *One Class Classification* tanto nos bancos de dados resultantes da geração de *outliers* quanto nestes após passarem pelos processos de redução de número de instâncias. Pode-se visualizar o fluxo adotado pelos experimentos na Figura 1.

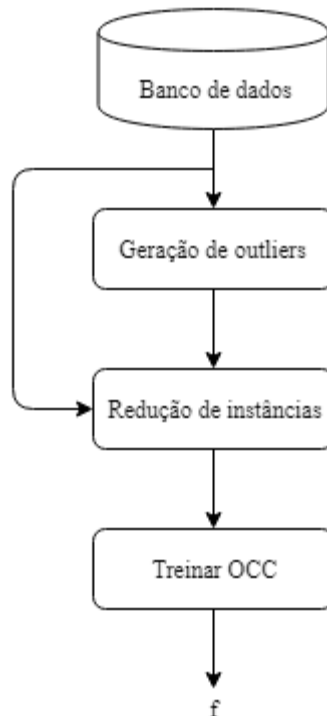


Figura 1: Descrição do fluxo do projeto em que ‘f’ representa o modelo treinado pelo experimento.

Na primeira etapa, referente à geração de *outliers*, a distribuição dos objetos pertencentes à classe *target* do banco de dados escolhido é extraída. Em seguida, um número fixo de *outliers* é gerado, utilizando uma distribuição complementar a que foi extraída, de acordo com um parâmetro predefinido pelo usuário, neste projeto foi adotado uma quantidade equivalente a 10 vezes o número de instâncias que são da classe *target*.

O banco de dados inicial junto aos seus *outliers* gerados pelo processo anterior podem então ter ou não seu número de instâncias reduzido. Em caso de redução, esta poderá ser feita pelos algoritmos de seleção de protótipos *Edited Nearest Neighbor* (ENN) [3] ou *Decremental Reduction Optimization Procedure 3* (DROP3) [4].

Na última etapa do fluxo, os dados passarão por um dos 3 classificadores de *One Class Classification*. Estes são os mesmos adotados em [1], isto é, *One Class Nearest Neighbor* (OCNN), *Minimum Spanning Tree* (MST) e *Support Vector Data Description* (SVDD) [5].

Os resultados dos experimentos são apresentados em tabelas. Estas contêm informações descritas na seção de Objetivos e, exceto a tabela com as taxas de redução, as restantes são geradas para cada um dos fluxos possíveis no projeto partindo da geração dos *outliers* até chegar aos classificadores.

Os parâmetros adotados, baseados naqueles escolhidos em [1], para cada um dos algoritmos utilizados neste projeto podem ser visualizados na Tabela 1. Enquanto que na Tabela 2, são descritos os 10 *datasets* utilizados para este projeto.

Algoritmo	Parâmetros
Geração de <i>outliers</i>	alpha = 1,2 beta = 10
ENN	Número de vizinhos = 3 Distância Euclidiana
DROP3	Número de vizinhos = 3 Distância Euclidiana
OCNN	Fração de rejeição = 0,5 Distância Euclidiana
MST	Fração de rejeição = 0,5 Max. Path = 20
SVDD	Tipo de <i>kernel</i> = RBF C = 5,0 Fração de rejeição = 0,5 Otimização com programação quadrática

Tabela 1: Algoritmos utilizados e seus respectivos parâmetros

Nome	Nº de objetos	Nº de instâncias <i>target</i>	Nº de atributos
Climate model, simulation crashes	540	494	18
Contraceptive method choice	1473	629	9
Credit Approval	690	307	15
Haberman's survival	306	225	3
Hill-Valley	606	305	101
Parkinsons	195	147	23
Pima Indian Diabetes	768	500	8
Sonar, mines versus rocks	208	111	60
Statlog (Heart)	270	150	13
Wisconsin Breast Cancer (Original)	699	458	10

Tabela 2: Bancos de dados utilizados nos experimentos

Cronograma

	Período																	
Atividade	Junho		Julho				Agosto				Setembro				Outubro			
Revisão bibliográfica	X	X	X	X														
Implementação				X	X	X	X	X	X	X								
Estudo do algoritmo							X	X	X	X								
Experimentos							X	X	X	X								
Avaliação dos resultados									X	X	X							
Escrita do TG											X	X	X	X	X	X	X	
Preparação da apresentação														X	X	X	X	

Referências

- [1] Krawczyk, Bartosz & Triguero, Isaac & García, Salvador & Wozniak, Michal & Herrera, Francisco. (2018). Instance reduction for one-class classification. Knowledge and Information Systems.
- [2] Chesner Désir, Simon Bernard, Caroline Petitjean, Laurent Heutte, One class random forests, Pattern Recognition, Volume 46, Issue 12, 2013, Pages 3490-3506, ISSN 0031-3203.
- [3] D.L. Wilson. Asymptotic Properties Of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man and Cybernetics 2:3 (1972) 408-421.
- [4] D.R. Wilson, T.R. Martinez. Reduction Techniques For Instance-Based Learning Algorithms. Machine Learning 38:3 (2000) 257-286.
- [5] Tax DMJ, Duin RPW (2004) Support vector data description. Mach Learn 54(1):45–66

Possíveis Avaliadores

Prof. Paulo Salgado Gomes de Mattos Neto

Assinaturas

Recife, 05 de SETEMBRO de 2018

A handwritten signature in black ink, appearing to read 'Hoi Yan Mark', written over a horizontal line.

Hoi Yan Mark

(Aluno)

A handwritten signature in black ink, appearing to read 'George Darmiton da Cunha Cavalcanti', written over a horizontal line.

George Darmiton da Cunha Cavalcanti

(Orientador)