UNIVERSIDADE FEDERAL DE PERNAMBUCO

Graduação em Engenharia da Computação
Centro de Informática
2015.2

Subdivisão de *clusters* aplicada ao treinamento de *ensembles* de classificadores

Proposta de Trabalho de Graduação

Aluna: Thaís Alves de Souza Melo (tasm@cin.ufpe.br)

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

INTRODUÇÃO

A identificação de *clusters* presentes em um conjunto de dados é um problema de grande importância na área de aprendizagem de máquina. É comum a obtenção de dados de treinamento não classificados que devam ser agrupados a partir de características de seus atributos, criando assim *clusters*. Essa divisão deve ocorrer de forma que os *clusters* possuam uma alta similaridade entre seus elementos, mas que sejam dissimilares de outros grupos. Atualmente, a identificação da quantidade ideal de *clusters* para determinado conjunto de dados pode ser obtida através da combinação de pelo menos 30 índices [1].

A partir da divisão da base encontrada ao se analisar as diferenças entre seus elementos, é possível realizar o treinamento supervisionado de um conjunto de classificadores, organizados em um *ensemble*. Esses classificadores, então, poderão trabalhar em conjunto para identificar a quais *clusters* novos padrões pertencem, tendo um desempenho igual ou superior a um sistema composto por apenas um classificador treinado para a base inteira [2].

Porém, mesmo que os *clusters* encontrados possuam uma alta similaridade entre seus elementos, existem conjuntos de dados que podem conter uma subdivisão interna: um conjunto de dados que busque classificar tumores em pacientes pode apresentar ligeiras diferenças entre os tumores, permitindo que tais tumores sejam classificados em subtipos, por exemplo.

Como a criação do *ensemble* de classificadores baseia-se na utilização de classificadores especializados em diferentes áreas do espaço de forma que, combinados, possuam uma melhor taxa de acerto no conjunto como um todo, a subdivisão dos *clusters* em subconjuntos pode se mostrar como uma boa maneira de separar o espaço de treinamento dos classificadores, a fim de que cada subgrupo possua um classificador especializado, otimizando a resposta do *ensemble*.

OBJETIVO

O objetivo desse trabalho de graduação é propor um algoritmo para encontrar os subgrupos de cada classe, a fim de melhorar o desempenho de sistemas de múltiplos classificadores. Nesses sistemas, os classificadores serão treinados de forma a se especializar nos subconjuntos encontrados.

METODOLOGIA

Será realizada uma revisão da literatura a fim de se conhecer mais sobre o estado da arte atual dessa área e possíveis formas de execução da subdivisão dos *clusters* e da criação de *ensembles* de classificadores.

Após essa revisão, será dado início ao desenvolvimento a uma técnica que consiga identificar a presença de subconjuntos nos *clusters* através de métricas com “ponto de joelho” [1]. Essa técnica será empregada para gerar um *ensemble* de classificadores, que será utilizado na execução de testes em bases de dados da UCI [3]. Os resultados serão analisados a partir da comparação com os resultados obtidos por *ensembles* criados a partir dos clusters, ignorando a subdivisão encontrada.

Por fim, haverá a escrita do relatório com a descrição de todo processo de pesquisa, desenvolvimento, análise e conclusões obtidas, seguido da preparação para apresentação para defesa do trabalho.

CRONOGRAMA

O cronograma previsto para a execução desse trabalho de graduação pode ser visualizado na Tabela 1.

Tabela 1. Cronograma

Atividade	Setembro		Outubro		Novembro		Dezembro		Janeiro	
Revisão da literatura										
Implementação										
Execução de testes e análises										
Escrita do relatório										
Preparação da apresentação										

POSSÍVEIS AVALIADORES

- Tsang Ing Ren

REFERÊNCIAS

[1] M. Charrad, N. Ghazzali, V. Boiteau e A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *Journal of Statistical Software,* Outubro 2014.

[2] L. Kuncheva, Combining Pattern Classifiers: Methods And Algorithms, New Jersey: John Wiley And Sons, INC., 2004.

[3] M. Lichman, "UCI Machine Learning Repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml. [Acesso em 30 Setembro 2015].

ASSINATURAS

_____

George Darmiton da Cunha Cavalcanti

(Orientador)

_____

Thaís Alves de Souza Melo

(Orientanda)