Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

<TÍTULO DA OBRA>

Guilherme Leite Moreira de Paiva

Dissertação de Mestrado

Recife
<DATA DA DEFESA>

Universidade Federal de Pernambuco
Centro de Informática

Guilherme Leite Moreira de Paiva

<TÍTULO DA OBRA>

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: *Prof. Dr. George Darmiton*

Recife
<DATA DA DEFESA>

Dedico este trabalho à ...

Agradecimentos

<DIGITE OS AGRADECIMENTOS AQUI>

<*DIGITE AQUI A CITAÇÂO*>
—<AUTOR> (<NOTA>)

Resumo

<DIGITE O RESUMO AQUI>

Palavras-chave: <DIGITE AS PALAVRAS-CHAVE AQUI>

Abstract

Keywords: <DIGITE AS PALAVRAS-CHAVE AQUI>

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Capítulo 1 - Introdução

Neste capítulo será dada uma introdução à combinação de classificadores bem como uma breve análise dos algoritmos Bagging, Boosting e ICS-Bagging. Para que este assunto seja melhor compreendido, será apresentado o contexto histórico, a motivação do uso de combinação de classificadores e os desafios atuais. Após a seção de motivação e contextualização seguirá um detalhamento deste trabalho, abordando objetivo e estrutura do mesmo.

1.1 Motivação e Contextualização

1.1.1 Histórico

Classificar algo sempre fez parte e é algo que ocorre várias vezes com os seres humanos. Um médico ao diagnosticar um paciente está classificando o mesmo como portador de uma determinada enfermidade ou não, um sommelier ao dar sua opinião sobre um vinho pratica da mesma tarefa. Como podemos ver, a tarefa de classificar algo é de suma importância e nada mais natural que essa tarefa fosse feita por uma máquina. Inicialmente, a abordagem era criar um classificador que pudesse, dado uma instância desconhecida, dizer a qual classe esta instância pertence. No entanto, novamente fazendo analogia ao comportamento humano, ao tomar uma decisão nada mais natural que uma pessoa consulte a opinião de outras pessoas. E isto é a base para se combinar classificadores, a decisão final não é baseada na resposta de um único classificador e sim nas respostas de vários.

1.1.2 Combinação de Classificadores

Na literatura temos vários algoritmos de combinação de classificadores, os primordiais foram: Bagging e Boosting. O algoritmo Bagging introduziu o conceito de bootstrap agreggating, na qual cada classificador é treinado com uma amostra da base de dados, cada classificador é colocado em um conjunto de classificadores e quando uma instância desconhecida é colocada, pelo voto da maioria dos classificadores ou por outro tipo de votação, a instância é classificada como pertencente a uma determinada classe. Já os algoritmos da família Boosting, que tem como algoritmo mais conhecido o AdaBoost, partem de ideia de que se uma instância é classificada de forma errada, esta instância deve ter mais importância e como consequência se uma instância é classificada de forma correta, sua importância é diminuída. A motivação deste algoritmo é termos vários classificadores ditos fracos, mas que são especialistas em um subdomínio do problema. A partir desses algoritmos, que poderíamos chamar de canônicos, vários outros surgiram. Um em especial é o algoritmo ICS-Bagging. Que de forma mais ampla,

combina os classificadores amparado por uma função de *fitness*, de forma que um classificador só é adicionado na *pool* de classificadores se o fato deste classificador ser adicionado na *pool* de classificadores aumenta o *fitness* da *pool*.

1.1.3 Bases Balanceadas e Bases Desbalanceadas

A distribuição de uma classe, ou seja, a proporção de instâncias que pertence a cada classe do conjunto de dados, desempenha um importante papel na combinação de classificadores. Uma base de dados é dita desbalanceado quando o número de instâncias de uma classe é muito maior que o da outra classe. De forma intuitiva, uma base balanceada mantém um proporção mais ou menos igual entre as classes. E este ponto é de suma importância quando se trata de combinação de classificadores pois a abordagem dos algoritmos e as métricas utilizadas diferem quando uma base é desbalanceado ou balanceada.

1.2 Objetivo

O objetivo deste trabalho é analisar diversas métricas de diversidade para o algoritmo ICS-Bagging e avaliar seu desempenho. A medida de desempenho será a área sobre a curva ROC de 5-cross-fold validation de cada base de dados. As bases de dados a serem analisadas são: Glass1, Pima, Iris0, Yeast1, Yeast2, Vehicle2, Vehicle3 e Ecoli1. Todas essas bases são balanceadas e foram obtidas no repositório Keel. No final do trabalho será possível identificar qual métrica obteve o melhor desempenho e qual a melhor proporção da medida de diversidade na função de *fitness* do algoritmo ICS-Bagging.

1.3 Estrutura do Trabalho

O restante deste trabalho possui um capítulo com detalhes sobre diferentes algoritmos de combinação de classificadores bem como uma abordagem mais detalhada do algoritmo ICS-Bagging. Durante o capítulo, será abordado o conceito de métrica de diversidade, bem como várias métricas de diversidade que serão analisadas neste trabalho. Logo após, segue um capítulo mostrando os resultados das analises das diversas métricas de diversidade bem como sua proporção na função de *fitness* bem como uma análise comparativa com algoritmos clássicos. Por fim, será concluído como se dá o comportamento dessas métricas de diversidade e a importância das mesmas na função de fitness.

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).