Universidade do Porto

Faculdade de Engenharia

FEUP

Aplicação de C4.5 ao diagnóstico de Parkinson

Relatório Intercalar

Inteligência Artificial

3º ano do Mestrado Integrado em Engenharia Informática e Computação

Grupo E4_3:

Henrique Ferrolho – ei12079 – henriqueferrolho@gmail.com

João Pereira – ei12023 – pereiraffjoao1993@gmail.com

19 de Abril de 2015

Objectivo

O estado final da implementação do projecto deverá ser capaz de diagnosticar um indivíduo com doença de Parkinson.
Esse diagnóstico é feito tendo em conta um conjunto de exemplos (*data sets*), dos quais é possível derivar regras. Após a derivação/aprendizagem dessas regras, é possível construir uma árvore de decisão baseada nelas. Finalmente, essa árvore pode ser usada para classificar indivíduos no domínio em análise.

O objectivo deste projecto é a determinação da árvore de decisão que traduz essas regras no diagnóstico de Parkinson.

Descrição

Especificação

Um conjunto de amostras é usado na aprendizagem das regras de classificação para a população do domínio em análise. Os elementos do domínio são definidos por um conjunto de atributos/variáveis, que constituem os parâmetros das regras de classificação a derivar após o processo de aprendizagem.

O *data set* usado pelo grupo possui um *training data set* e um *test data set*.
O *training data set* contém 40 (quarenta) entradas: 20 (vinte) entradas de indivíduos diagnosticados com doença de Parkinson, e outras 20 de indivíduos saudáveis. De seguida estão descritos os métodos que foram usados para obter os *data sets*.

Training data set
Vários tipos de artefactos sonoros de todos os indivíduos (26 amostras de voz, incluindo vogais, números, palavras e frases curtas) foram gravados. Posteriormente, um grupo de 26 características baseadas em frequência linear e temporal foram extraídas de cada amostra de voz. A UPDRS (*Unified Parkinson's Disease Rating Scale*) de cada paciente (que é determinada por um médico especialista) também está disponível na base de dados.

Test data set
Depois de o conjunto de dados de treino ter sido recolhido, continuou-se a recolher dados para um novo conjunto de dados independente, a partir da análise de mais indivíduos com doença de Parkinson, avaliadas pelo mesmo especialista. Durante a recolha deste conjunto de dados, 28 pacientes com doença de Parkinson foram convidados a suster as vogais 'a' e 'o' três vezes, o que perfaz um total de 168 gravações. Os mesmos 26 recursos foram extraídos das amostras de voz desse conjunto de dados.
Este conjunto de dados pode ser usado como um conjunto independente de teste para validar os resultados obtidos no conjunto de treino.

O programa do grupo vai aplicar o algoritmo C4.5 para determinar uma árvore de decisão, com base no conjunto de dados, que traduz as regras de classificação desse conjunto de dados. Essa árvore deve poder ser utilizada na classificação de novos casos.

O grupo planeou dividir o trabalho nas seguintes fases:

- Implementação das funções que carregam e guardam os conjuntos de dados em estruturas de dados adequadas
- Aplicação do algoritmo de aprendizagem C4.5 aos conjuntos de dados carregados
- Definir um visualizador da árvore de decisão resultante, com as respectivas regras de classificação
-

  Medição detalhada de resultados nos dados de treino e de teste.

Indique as partes em que tenciona dividir o trabalho e as fases em que tenciona abordar cada parte.

Trabalho efectuado

Até à altura, o grupo já implementou as funções necessárias para carregar os dados de treino (*training data set*), bem como os dados de teste (*test data set*).
Esses dados são guardados em estruturas de dados adequadas para a sua posterior consulta e processamento.

Também foi implementada uma função para o cálculo da *entropia*. Contudo, caso o grupo opte pela utilização de uma *framework*, esta passará a não ter utilidade nenhuma.

Resultados esperados e forma de avaliação

Enumere testes a definir para validar o resultado do trabalho

Conclusões

Escreva aqui as conclusões que achar devidas.

Recursos

GitHub
https://github.com/

LibreOffice Writer
https://www.libreoffice.org/discover/writer/

UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with+
+Multiple+Types+of+Sound+Recordings