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
- Visualizar a árvore de decisão resultante, com as respectivas regras de classificação
- Programar um sistema capaz de medir detalhadamente os resultados obtidos nos dados de treino e de teste.

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* para aplicar o algoritmo C4.5, em vez de implementar o algoritmo de raiz, a função do cálculo da entropia passará a não ter utilidade.

Resultados esperados e forma de avaliação

Como é referido em cima, na divisão das fases de trabalho, a fase final consiste na implementação de uma medição dos resultados obtidos. Esta medição, que acaba por ser uma validação dos resultados, pode ser obtida de duas formas diferentes que se descrevem de seguida.

A primeira forma para validar os resultados da aplicação será submeter a totalidade do conjunto de dados de treino ao algoritmo de construção da árvore de decisão. Posteriormente, cada entrada do conjunto de dados de teste será submetida a essa árvore de decisão. De acordo com o que é previsto, se tudo estiver a funcionar correctamente, a árvore irá diagnosticar todos os indivíduos do conjunto de dados de teste com a doença de Parkinson.

Uma segunda forma de validar a qualidade das árvores de decisão geradas pelo nosso programa será escolher, aleatoriamente, apenas alguns indivíduos do conjunto de dados de treino para a construção da árvore de decisão. O grupo prevê que, se tudo estiver bem implementado, essa árvore será capaz de classificar correctamente todos os casos

de treino, incluindo os indivíduos que não foram usados para a construção da árvore, bem como os indivíduos do conjunto de dados de teste.

Conclusões

O grupo conclui este relatório intercalar com um sentimento de pena por achar que o semestre em que se enquadra não reúne as condições necessárias para a realização de um projecto com este calibre. De facto, o projecto é muito interessante e teria toda a lógica, dado o curso em que estamos, implementar os algoritmos que lhe estão inerentes "desde o zero".

O grupo lamenta o facto de não haver distinção entre os grupos que optam por programar as suas próprias implementações dos algoritmos que os seus projectos envolvem, daqueles grupos que usam *frameworks* ou outro tipo de ferramentas já existentes que facilitam o processo.

Por essas razões, o grupo abandonou a ideia inicial de implementar a sua própria versão do algoritmo C4.5, para começar a pensar em formas de usar uma *framework* já existente. Desta forma, o grupo não só poupa esforço, como também poupa tempo (que tanto falta neste semestre) para dedicar às outras unidades curriculares.

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