Segundo trabalho prático de inteligência computacional

Nome

Departamento de Ciências da Computação
Universidade Estadual de Montes Claros
Herberth Amaral
15 de julho de 2015

1 Introdução

O presente trabalho analisa o uso de redes neuro-fuzzy para fazer classificação de dados de comparação de registros, a fim de categoriza-los como registros correspondentes ou não-correspondentes, em um processo conhecido como record linkage.

A base de dados contém os resultados de comparação de dados demográficos provenientes do registro epidemiológico de câncer do estado alemão de Rhine-Westphalia [1].

Uma versão semelhante deste trabalho utilizando o mesmo dataset em uma rede neural do tipo perceptron de múltiplas camadas já foi apresentado para esta disciplina.

2 Record linkage

Esta seção contém uma breve introdução do processo de record linkage e foi inteiramente inspirada em um survey sobre o assunto [2].

Record linkage é o processo de identificar registros diferentes ou múltiplos que correspondem a uma entidade real ou a um objeto. Segundo o processo é conhecido como outros nomes, como merge-purge, data deduplication, instance identification, coreference resolution, identity uncertainty e duplicate detection.

O processo de record linkage é composto majoritariamente de 3 fases: preparação de dados, comparação de campos e, finalmente, detecção de duplicidade.

A fase de preparação de dados conta com os passos de análise, transformação de dados e normalização de dados. Esta fase tem como objetivo remover heterogenidades dos dados com o fim de facilitar (ou mesmo permitir) a comparação. Um exemplo seria um processo de record linkage para achar duplicações em uma base de dados de pacientes: os acentos dos nomes seriam removidos, todas as letras seriam colocadas em minúsculo e contrações, como "Ma"seriam convertidos para "Maria".

A fase de comparação de campos tem o objetivo de medir o nível de semelhança entre os campos de dois registros previamente preparados. Essa métrica é obtida através de comparação feita com algoritmos específicos que levam em conta a distância de edição (distância Levenshtein, distância Jaro-Winkler, Smith-Waterman, dentre outras), de token (strings atômicas, WHIRL) e semelhança fonética (soundex, NYSSIIS, ONCA, Metaphone).

Finalmente, há a fase de detecção de registros duplicados. Nesta fase, os dados provenientes da etapa anterior são analisados e classificados entre correspondentes ou não-correspondentes. Esta classifcação pode ser probabilística ou determinística.

3 Características da base de dados

A base de dados é composta de registros provenientes da fase de comparação de dados e cada linha contém as seguintes informações:

1. *id*_1: identificador do primeiro registro.
2. *id*_2: identificador do segundo registro.
3. cmp_fname_c1: nível de concordância do primeiro nome, primeiro componente
4. cmp_fname_c2: nível de concordância do primeiro nome, segundo componente
5. cmp_lname_c1: nível de concordância do primeiro sobrnome, primeiro componente
6. cmp_lname_c2: nível de concordância do primeiro sobrnome, segundo componente
7. *cmp_sex*: nível de concordância do sexo
8. *cmp_bd*: nível de concordância agreement da data de nascimento, componente do dia
9. *cmp_bm*: nível de concordância agreement da data de nascimento, componente do mês
10. *cmp_by*: nível de concordância agreement da data de nascimento, componente do ano
11. *cmp_plz*: nível de concordância de código postal
12. *is_match*: indica se os dois registros são iguais (TRUE para iguais, FALSE para não-iguais)

Os níveis de concordância ficam entre 0 e 1, sendo que 0 indica nenhuma semelhança/concordância e 1 indica total semelhança/concordância.

4 Metodologia

A base de dados contém 5749132 registros, divididos em dez blocos de igual tamanho. O primeiro bloco foi usado para treinar a rede neural e os outros nove foram usados para validar o treinamento.

Apesar da riqueza de detalhes de dados, nem todos os campos do dataset foram utilizados neste trabalho. Os campos usados, nomeadamente, foram: *cmp_fname_c*1, *cmp_lname_c*1 e uma combinação dos campos *cmp_bd*, *cmp_bm* e *cmp_by* como forma de representar a equivalência de data de nascimento.

O código-fonte utilizado para a realização deste trabalho pode ser encontrado em https://github.com/herberthamaral/mestrado/tree/master/IC/trabalho-fuzzy (acesso em 15/05/2015).

5 Resultados

Os testes demonstraram um resultado melhor do que a versão anterior deste projeto: apenas 38 erros de classificação em um total de 574913 registros de teste (0.0056%) contra 108 da abordagem anterior.

O bom resultado pode ser explicado pela uniformidade da base de dados.

Apesar de conseguir um resultado um pouco melhor, o tempo de treinamento subiu de aproximadamente 5 minutos na versão MLP para 32 minutos na versão ANFIS.

Referências

Referências

[1] Irene Schmidtmann, Gael Hammer, Murat Sariyar, Aslihan Gerhold-Ay; Evaluation des Krebsregisters NRW Schwerpunkt Record Linkage. Technical Report, *IMBEI 2009*. Disponível em `https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns`. Acessado em 15/07/2015

[2] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios; Duplicate Record Detection: A Survey. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 19, NO. 1, Janeiro de 2007.