IO2006, ISEG Lisboa, 8-11 Outubro

EXTRAIR CONHECIMENTO DE PROVÉRBIOS

Armando B. Mendes (amendes@notes.uac.pt)
CEEAplA e Universidade dos Açores, Departamento de Matemática

Günther Matthias A. Funk (mfunk@notes.uac.pt)
Universidade dos Açores, Departamento de Matemática

Mª. Gabriela C.B. Funk (funk@notes.uac.pt)
Universidade dos Açores, Departamento de Línguas e Literaturas Modernas

PALAVRAS-CHAVE:
Geração de conhecimento; *Data mining*, Provérbios; Desnormalização; Modelos discriminantes lógicos; Árvores de regressão.

RESUMO:

Com base no "Adagiário Popular Açoriano" de Armando Côrtes-Rodrigues e muitas outras colectâneas portuguesas deste século foram recolhidos cerca de 25.000 frases idiomáticas. Em pré-testes para filtragem de frases idiomáticas menos usadas, foram inquiridos indivíduos com idades superiores a 40 anos, residentes em diferentes localidades da ilha de São Miguel, tendo sido a recolha efectuada em lares de idosos e centros paroquiais de convívio. Estes inquéritos de reconhecimento passivo (apenas indicar os que conhece) foram posteriormente comparados com alguns resultados de reconhecimento activo (completar um provérbio iniciado), tendo-se concluído da proximidade de resultados entre reconhecido activo e passivo. Observou-se igualmente o reconhecimento activo e passivo de inquiridos sem saber ler ou escrever, concluindo-se novamente não existir diferenças significativas para diferentes extractos sociais, ainda que existissem expectativas de que os iletrados tivessem mais apetência por provérbios.

Após esta primeira fase verificou-se que cerca de dois quintos dos exemplares testados não tinham sido reconhecidos por nenhum dos inquiridos, tendo sido excluídos de inquéritos subsequentes. No projecto de maior envergadura que englobou todas as ilhas dos Açores e algumas localidades dos EUA com forte componente de imigração açoriana, utilizou-se uma base de dados com cerca de 5.000 provérbios conhecidos em diferentes ilhas dos Açores e com índice de reconhecimento superior a 10% no pré-teste. Utilizou-se um processo de amostragem por quotas, controlando os factores: sexo, três classes de idade e 2 classes do grau de habilitações. Verificando-se posteriormente que a amostra era representativa da população para as faixas de idade consideradas, género e profissões.

O cruzamento entre os inquiridos e os provérbios resultou numa tabela com cerca de 250.000 registos. Sobre estes dados foram já realizados diferentes trabalhos de estatística descritiva resultando na publicação de três livros pelos dois últimos autores deste artigo: "Pérolas da Sabedoria Popular Portuguesa: Provérbios de São Miguel", "Provérbios das Ilhas do Grupo Central dos Açores" e "Provérbios Açorianos nos EUA".

Com o objectivo de extracção de conhecimento por análise de dados é, agora, necessário construir uma tabela de dados a analisar. No caso presente, entre outros objectivos, pretende-se identificar grupos de indivíduos com maior capacidade de reconhecer provérbios e, em simultâneo, caracterizar esses grupos. Partindo de bases de dados normalizadas a construção da tabela de dados de indivíduos com o máximo de atributos passa por um processo de desnormalização, por fazes de agregação de valores de atributos discretos e por um rigoroso controlo da qualidade dos dados obtidos. Nesta comunicação apresentam-se alguns exemplos de problemas que podem surgir em cada uma das fazes referidas.

Utilizam-se algoritmos de *machine learning* como árvores de regressão e modelos discriminantes lógicos como CHAID e CART. Verificou-se nomeadamente uma redução significativa do conhecimento dos provérbios por parte de inquiridos com menos de 40 anos. Tal facto constitui um padrão normal em qualquer cultura se considerarmos que o processo de familiarização com os textos proverbiais necessita de um mínimo de experiência de vida. No entanto, verificou-se igualmente uma diminuição na percentagem de provérbios reconhecidos em inquiridos com idades superiores a 65 anos. De referir que a densidade proverbial respeitante ao conjunto dos informantes luso-americanos cabe perfeitamente dentro do espectro das taxas homólogas registadas no arquipélago, pelo que se conclui que a cultura açoriana se mantém ainda viva nessas zonas de emigração. No entanto, identificaram-se importantes diferenças entre os EUA e o Canadá, observando-se um grau de reconhecimento de provérbios acima da média no primeiro e abaixo no segundo.