IDENTIFICAÇÃO DE CONTATOS DUPLICADOS COM O CLASSIFICADOR C4.5

MACHADO, Rafael; LINDENAU, Guilherme; PINHEIRO, Rafael; NUNES, Eliza
BORGES, Eduardo N.
rafaelmachado@furg.br

Evento: Congresso de Iniciação Científica
Área do conhecimento: Ciências Exatas e da Terra, Ciência da Computação

Palavras-chave: mineração de dados; classificação; deduplicação.

1 INTRODUÇÃO

Com a explosão do número de aplicações Web disponíveis, os usuários tendem a acumular diversas contas em diferentes serviços. Gerenciar estas informações é uma tarefa complexa para o usuário. Funções básicas dos dispositivos móveis podem ser prejudicadas pela redundância da informação coletada automaticamente por diferentes aplicações. Este trabalho apresenta uma avaliação experimental de um classificador utilizado para deduplicação (BORGES et al., 2011) de contatos.

2 MATERIAIS E MÉTODOS

O experimento realizado utilizou uma base de dados real com 1962 contatos extraída do GMail. Foram selecionados todos os registros, contendo somente os atributos nome, e-mail e telefone.

O estudo é baseado na Descoberta de Conhecimento em Bases de Dados, definida por TAN, STEINBACH e KUMAR (2005) como uma metodologia não trivial de identificar padrões potencialmente úteis e compreensíveis em meio às observações presentes em uma base de dados. As *strings* foram pré-processadas com o objetivo de normalizar os nomes, e-mails e telefones. Na transformação dos dados, foi gerada uma lista de pares de contatos. Cada par foi avaliado por um usuário especialista que os rotulou como contatos pares replicados (1) ou distintos (2). Para cada par, também foram calculadas as funções de similaridade Jaccard, JaroWinkler e MongeElkan (COHEN, RAVIKUMAR & FIENBERG, 2003) entre os nomes dos contatos. Com estes valores foi gerado o conjunto de dados final utilizado na mineração de dados, como mostra a figura 1.

Figura 1 – Exemplo arquivo csv.

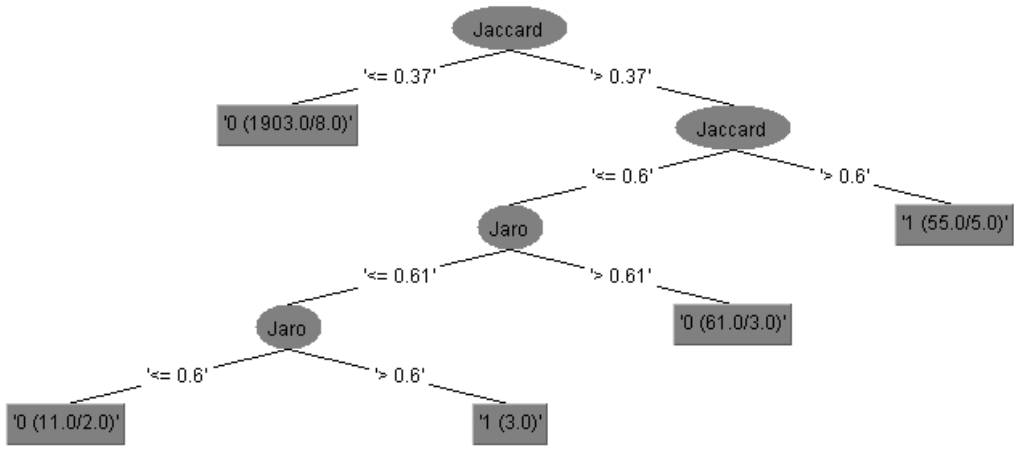
DadosRéplica - 2.csv

```
1  id1,id2,Jaccard,Jaro,MongeElkan,replica
2  3,4,0.33,0.91,0.64,0
3  3,5,0.33,0.9,0.64,0
4  3,776,0.25,0.64,0.64,0
5  4,5,0.33,0.93,0.71,0
6  6,1071,0.33,0.68,0.63,0
7  8,336,0.25,0.76,0.72,0
```

3 RESULTADOS

Através do classificador C4.5, implementado na ferramenta Weka (HALL et al., 2009) como J48, foram obtidos os resultados apresentados na figura 2.

Figura 2 – Árvore gerada pelo classificador



A função mais importante foi Jaccard. Localizada na raiz da árvore, ela é quem melhor distribui os registros nas classes réplica (1) ou pares distintos (0). Pares cujos valores retornados pela função são menores ou iguais a 37% foram classificados como pares distintos enquanto maiores que 60% como pares replicados. Os demais pares foi avaliado pela função JaroWinkler de acordo com os limiares indicados nas arestas. Nota-se que a função MongeElkan não aparece nos resultados. Isso ocorre devido ao classificador não ter encontrado um padrão nos valores obtidos por esta função de similaridade. Em suma, a medida de acurácia do modelo foi igual a 98,82%.

4 CONSIDERAÇÕES FINAIS

Como principal conclusão destaca-se a ótima qualidade do algoritmo C4.5 e combinado com as funções Jaccard e JaroWinkler na verificação de dados de contatos replicados em a partir de múltiplas fontes de dados.

REFERÊNCIAS

BORGES E.; BECKER, K.; HEUSER, C.; GALANTE, R. A classification-based approach for bibliographic metadata deduplication. In: Proceedings of the IADIS International Conference WWW/Internet, p. 221-228, Rio de Janeiro, 2011.

COHEN, W.; RAVIKUMAR, P.; FIENBERG, S. A comparison of string metrics for matching names and records. In: KDD Workshop on Data Cleaning and Object Consolidation. p. 73-78. 2003.

HALL, M. et al. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, v. 11, n. 1, p. 10-18, 2009.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining. Addison-Wesley, 2005.