Two-Phase Feature Recommendation in Entity Resolution

Introduction

Entity Resolution (ER) refers to identifying **records** that **correspond to** the **same real-world entity** from one or more databases. There are normally two main stages in ER: blocking and classification. Blocking uses one or several features to divide data into different blocks. Classification also uses some features to find the records that refer to the same entity in each block.

Database

Records	Surname	Given name	Postcode	Address
1	Green	Jim	2612	Masson St.
2	Green	James	2612	Masson St.
3	Green	Jim	2612	Masson Street
4	White	Jim	2612	Watson Street

Do they refer to the same person

Problem Statement

The three main challenges of feature recommendation are:

- **Features** selected to compare record pairs
- **Methods** by which to evaluate the features (e.g. attributes)
- **Training sets** that used for feature selection.

Existing solutions consider the selected features in blocking and classification separately, ignore the correlation between features applied in blocking and classification.

We aim for a Two-Phase **Feature Recommendation System** that

- selects optimal features for blocking
- and uses **correlation based feature selection** technique to select features for classification

Feature Recommendation Approach

A Weak Training Set

A weak training set is used to generate a set of positive pairs (i.e. positive set) and a set of negative pairs (i.e. negative set) based on feature value similarity.
Here, we use log TFIDF measure as similarity measure:

$$sim(r_1, r_2) = \sum_{i=1}^{|F|} w(r_1, f_i) w(r_2, f_i)$$

Where $w'(r,f) = log(tf_{r,f} + 1) log(\frac{|F|}{df_f} + 1)$, and $w(r,f) = \frac{w'(r,f)}{\sqrt{\sum_{i=1}^{|F|} w'(r,f_i)^2}}$

Here, $tf_{r,f}$ is the frequency of the feature value in record r; df_f is the number of records in which feature value f appears; $|F|$ is the number of features each record contains; $w(r,f)$ is the normalized TFIDF weight of a feature in a record.

First Phase

In the first phase, we select optimal features for blocking. After positive and negative pairs are generated, we can figure out the most relevant features:

- Find out distinguished features among the database which (1) must cover a minimum number of negative pairs; (2) have feature values in the negative set that are different from the mode value.
- Calculate the optimal score of each feature (f_i) by function:

$$\theta(f_i) = \sum_{d \in f_i} \frac{freq(d)}{freq(m) * n}$$

Here, m is the mode of feature (f_i), $freq(d)$ is the number of feature value equal to d in the negative set, n is the number of different feature values of feature f_i.

- For the optimal score of a feature set ($f_{i,j,\dots}$), we have:

$$\theta(f_{i,j,\dots}) = \sum_{d_i \in f_i, d_j \in f_j} \frac{freq(d_{i,j})}{freq(m) * n}$$

Second Phase

In the second phase, we aim to recommend features for classification, which considers the correlation between features.

Feature correlation score, which measures the relevance of feature f_j to feature f_i, is defined as:

$$\delta(f_i, f_j) = \theta(f_i) * \theta(f_j) * \sum_{d_i \in f_i, d_j \in f_j} conf(d_j \rightarrow d_i)$$

Here, $conf(d_j \rightarrow d_i) = \frac{freq(d_i, d_j)}{freq(d_j)}$ is the confidence to calculate the probability that d_i will appear when d_j appears.

We may generate multiple correlated features in the second phase separately using the above function.

Example

Take the above dataset as an example, we may have below results from the weak training set:

Positive	Records	Negative
	1	
	2	
	3	
	4	

Let the minimum number be 2. We may have two features as candidates that cover at least 2 pairs of records.

Feature	Surname	Address
Optimal Score	0.5	0.333

Hence, we select “Surname” for blocking

Then we can have a table of feature correlation scores based on “Surname” in Phase One.

Feature	Given Name	Post code	Address
Correlation Score	0.05	0	0.333

Hence, we select “Address” as a correlated feature for Classification

Jingyu Shao and Qing Wang

Research School of Computer Science, College of Engineering and Computer Science,
The Australian National University, Canberra ACT 0200, Australia
{Jingyu.shao ,qing.wang}@anu.edu.au

Australian National University

ANU College of Engineering & Computer Science