Name: Abhishek Sant
USC ID: 5042550863
Title: Efficient Record Linkage Using Machine Learning
Course Name: "Directed Research - Prof. Pedro Szekely"

Efficient Record Linkage Using Machine Learning

Abhishek Sant

Department of Computer Science

University of Southern California,

Los Angeles, California, USA.

asant@usc.edu

***Abstract*— This paper provides the approach of performing efficient Record Linkage using Fril and Support Vector Machines.**

***Index Terms*—Record Linkage, Fril, SVM.**

I. Introduction

Record linkage combines information from a variety of files. The basic methods compare some of the attributes across pairs of files to determine those pairs of records that are associated with the same entity. An entity might be a business, a person, or some other type of unit that is listed [1]. The goal of record linkage is to find syntactically distinct data entries that refer to the same entity in two or more input files. The process is important for both data cleaning and integration in birth defects surveillance and research.

Fril extends traditional record linkage tools with a richer set of parameters. Users may systematically and iteratively explore the optimal combination of parameter values to enhance linking performance and accuracy [2].

Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [3]. SVM is responsible for increasing the precision and Recall of the Record Linkage Data generated by Fril.

II. Record linkage approach

Initially we have 2 source files having the data which is supposed to be linked. Fril is used with a default configuration as a rest service which gives a Record Linkage output on the Jetty Server. The linkage output is shown to the user and his feedback is required for randomly chosen n number of entries where ‘n’ is less than five percentages of the records joined. Fril generates a confidence value of the rows linked. To make the feedback process simpler, the user can sort the output based on confidence and then give the feedback of first ‘n/2’ and last ‘n/2’ records.

The feedback is passed in a form of JSON to the other module which computes the feature vectors based on the algorithm provided. The algorithms include Edit Distance, Numeric Distance etc. Once the feature vectors are computed the data is provided to SVM run on WEKA. Java API’s provided by WEKA are used to get the weights.

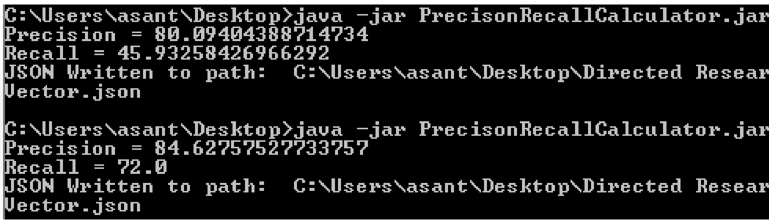
Through our experimentation we have observed that the more important an attribute is, more negative is the weight is returned by the SVM. The weights given by the SVM are scaled to the weights which Fril understands having higher weight for important attributes.

Fril is re-run with the new weights and the process is repeated till the user is satisfied with the linkage data generated.

III. Progress so far

The current system generates a default linkage output using Fril API’s for Edit Distance and Numeric Distance. The JSON injected to the phase2 computes feature vectors for these 2 algorithms and gets the weights using SVM, scales the weights and generates a JSON which acts as an input to Fril.

Following screen shot displays the improvement in precision and recall in 2 successive tests.



```
C:\Users\asant\Desktop>java -jar PrecisonRecallCalculator.jar
Precision = 80.09404388714734
Recall = 45.93258426966292
JSON Written to path:  C:\Users\asant\Desktop\Directed Resear
Vector.json

C:\Users\asant\Desktop>java -jar PrecisonRecallCalculator.jar
Precision = 84.62757527733757
Recall = 72.0
JSON Written to path:  C:\Users\asant\Desktop\Directed Resear
Vector.json
```

IV. Enhancements

Many other algorithms like Jaro-Winkler, Q Grams distance can be incorporated which would make the linkage between certain attribute more efficient.

The User Interface should be developed at Karma which would take data from the default configuration generated from Fril and create a JSON having feedback data in it.

All the implementations of newly introduced algorithms should be a part of phase2 where feature vectors are calculated to get the svm weights.

V. References

[1] ZWilliam E Winkler, “Overview of record linkage and current research directions” BUREAU OF THE CENSUS, 2006

[2] Pawel Jurczyk, James J. Lu, Li Xiong, Janet D. Cragan, Adolfo Correa, FRIL: A Tool for Comparative Record Linkage, American Medical Informatics Associations (AMIA) 2008 Annual Symposium.

[3] http://en.wikipedia.org/wiki/Support_vector_machine.