Ontology matching evaluation

A statistical perspective

Mohammadi, M.; Hofman, Wout; Tan, Yao Hua

Publication date
2016

Document Version
Final published version

Citation (APA)
Mohammadi, M., Hofman, W., & Tan, Y. H. (2016). *Ontology matching evaluation: A statistical perspective*. 231-232. Poster session presented at The Eleventh International Workshop on Ontology Matching, Kobe, Japan.

Ontology Matching Evaluation: A Statistical Perspective

Majid Mohammadi[1], Wout Hofman[2], Yao-hua Tan[1]

[1] Faculty of Technology, Policy and Management, Delft University of Technology, Netherlands
[2] Department of Technical Science, The Netherlands Institute of Applied Technology (TNO), Soesterberg, the Netherlands

Abstract. This paper proposes statistical approaches to test if the difference between two ontology matchers is real. Specifically, the performances of the matchers over multiple data sets are obtained and based on their performances, the conclusion can be drawn whether one method is better than one another or not. To do so, the paired t-test and Wilcoxon signed rank test are proposed and the comparisons over six recently proposed methods are reported.

Keywords: Ontology alignment, evaluation, statistical inference, paired t-test, Wilcoxon signed rank test

1 Introduction

There has been an increasing interest in ontology matching (or alignment) over the last years. As data come from various sources these days, the heterogeneity among data is inevitable. The solution to such an issue is ontology matching, which has a wide range of application from data integration and agent interoperability in computer science to matching ontologies in biomedical and geoscience. As a result, a plethora of methods have been proposed claiming that their method is better than, or competitive with, other state-ofthe-art algorithms. However, no evidence has been brought to support such a claim

2 Binary comparison of matchers

The hypothesis testing is one of the major topic in the realm of statistical inference. Here, we aim at utilizing this technique to indicate if the average difference in the performance scores of two matchers over multiple benchmarks is meaningful or not. To leverage the hypothesis testing, a null hypothesis is required. The null hypothesis (shown by H_0) states that there is no significant difference between two populations according to the available samples of the populations. The alternative hypothesis (shown by H_a), on the other hand, is the rival hypothesis and states that there is meaningful difference between two populations based on available samples. Thus, it is desirable to reject null hypothesis and accept the alternative hypothesis. In ontology matching case, the performance of various matchers over a range of data sets are available and we would like to test if the average of their performances is random. In other words, the null hypothesis and the alternative hypothesis in this case is

$$H_0 : \hat{P}^1 = \hat{P}^2$$

$$H_1 : \hat{P}^1 \neq \hat{P}^2 \tag{1}$$

where $\hat{P}^i$ is the average performances of the matcher i.
Before running any statistical test, the significant level must be determined. the α is the probability of rejecting null hypothesis when the null hypothesis is true. To the best of our knowledge, no statistical techniques have been employed to test the above-mentioned hypothesis. Firstly, the widely-used paired t-test is presented with more detail. Having hard preconditions to be satisfied, it must be warned that t-test might be inappropriate and statistically unsafe. Thus, the Wilcoxon signed-rank test is presented which is able to detect more difference even though the number of samples are not large enough.

2.1 Paired t-test

A common way to check if the difference between two matchers on different data sets is not random is to compute the paired t-test. Let $d_i = P_i^1 - P_i^2$ be the difference between the performances of two matchers over $i-th$ data set. The t statistics is computed as $t = \frac{x-\hat{x}}{\hat{\sigma}_d}$ where $\hat{x}$ and $\hat{\sigma}_d$ are sample average and standard deviation of samples, respectively. This statistics is distributed according to the Student distribution with N 1 degree of freedom. After obtaining the probability of observing the data given that H_0 being true (p-value) according to the Student distribution, the H_0 can be rejected if $p - value < \alpha$ and then Ha is accepted.

2.2 Wilcoxon Signed Rank test

The non-parametric alternative to the paired t-test is Wilcoxon singed rank test. This method ranks the absolute values of performance differences of two matchers. Then, it compares the rank of positive and negative differences. After computing the difference between two matchers over the the $i-th$ data set, d_i , the differences are ranked based on the values of d_i , disregarding its sign. if $d_i = 0$ it is ignored and the average ranks are assigned if the performances over one data set ties. Assume $W^+ = \sum_{d_i>0} rank(d_i)$ and $W^- = \sum_{d_i<0} rank(d_i)$ and $T = min(W^+, W^-)$. Then $z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$ is distributed according to the normal distribution.

3 Experimental Results

Table 1 tabulates the p-values obtained by paired t-test and Wilcoxon Signed Rank test over six recently proposed methods.

Table 1. The p-values obtained by paired t-test (above diagonal) and Wilcoxon Signed Rank test (below diagonal) over six recently proposed methods: XMAP, AML, AML2014, CroMatcher, edna and refalign.

	XMAP	AML	AML2014	CroMatcher	edna	refalign
XMAP		0.526403	0.23326767	0.00094182	0.000972	0.000939
AML	0.640625		0.05359674	0.00079181	0.113909	0.000697
AML2014	0.00647436	0.01596065		0.00026227	0.243871	0.000243
CroMatcher	0.00097656	6.10E-05	8.56E-05		2.83E-06	0.01664
edna	0.000822	0.011231	0.058088	0.000287		4.75E-06
refalign	0.000977	6.10E-05	8.50E-05	0.003906	0.000285	