Statistical Challenges to Inductive Inference in Linked Data

David Jensen
Experimental Knowledge Systems Laboratory
Computer Science Department, University of Massachusetts, Amherst MA 01003-4610
jensen@cs.umass.edu

Introduction

Many data sets can be represented naturally as collections of linked objects. For example, document collections can be represented as documents (nodes) connected by citations and hypertext references (links). Similarly, organizations can be represented as people (nodes) connected by reporting relationships, social relationships, and communication patterns (links). This type of data representation is facilitated by the growing availability of object-oriented databases and hypertext systems. The analysis of such data, often called *link analysis*, is becoming increasingly important in such diverse fields as law enforcement, fraud detection, epidemiology, and information retrieval.[1]

One important class of techniques for link analysis involves *inductive inference* — the construction of predictive relationships from linked data. For example, researchers in information retrieval might construct models to predict whether a particular WWW document is a homepage based on features of other documents to which it is connected (e.g., pages listing publications or family photos). Similarly, law enforcement agents might construct models to predict whether a person is a potential money-launderer based on bank deposits, international travel, business connections, and the arrest records of known associates.[2] Constructing such predictive relationships is an essential part of link analysis.

Despite growing data analysis opportunities, relatively little work examines the unique statistical challenges of inductive inference in linked data. This paper examines three such challenges: 1) statistical dependence caused by linked instances; 2) bias introduced by sampling density; and 3) multiple comparisons intensified by feature combinatorics. In general, current systems for link analysis ignore these challenges. Some systems are almost entirely graphical display tools, and do not explicitly support inductive inference. Other systems, such as algorithms for inductive logic programming (ILP), either ignore the peculiar statistical challenges of inductive inference in linked data or only consider special cases of linked data for which these problems are less severe. However, increasing numbers of systems do attempt inductive inferences about relationships in large, interconnected sets of objects (e.g., Slattery and Craven 1998), and addressing the challenges below will be essential to valid inferences in these systems.

Instance Linkage

One important characteristic of linked data is that features of some instances may be dependent on the features of other instances. To understand how this affects inductive inferences, first consider the case of independent instances, shown schematically in Figure 1a. Each node $\{A_1...A_n\}$ represents an instance in a training set. Each node could, for example, represent a web document.

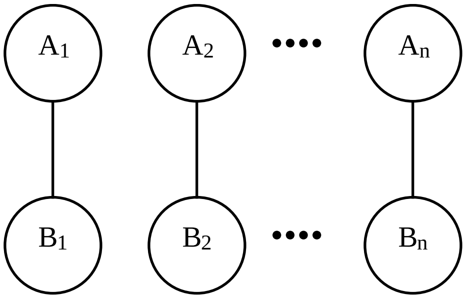

Figure 1a: Independent Instances

To construct a rule for identifying homepages, we might consider features of the documents themselves (e.g., page name, page size, number of images). We might also consider features of other web documents linked to those pages (e.g., the name, size, or number of images in documents that point to the prospective homepages).

[1] See the web pages of the 1998 AAAI Fall Symposium on AI and Link Analysis for more information on link analysis and the relevance of various AI techniques <http://eksl-www.cs.umass.edu/aila/>

[2] For an additional discussion of the use of link analysis to detect money laundering, see Jensen (1997) and Goldberg and Senator (1995).

Accepted to Uncertainty99: The Workshop on AI and Statistics

That is, we might consider features of documents represented by the nodes $\{B_1...B_n\}$ in Figure 1a.

By extending the feature set of the nodes $\{A_1...A_n\}$ to include features of the nodes $\{B_1...B_n\}$, the instances in our training set now consist of the subgraphs $\{\{A_1,B_1\}...\{A_n,B_n\}\}$ rather than only the nodes $\{A_1...A_n\}$. However, this does not change the relative independence of the instances in our training set because the subgraphs $\{\{A_1,B_1\}...\{A_n,B_n\}\}$ are still independent.

However, now consider the case shown in Figure 1b. Here there are still *n* prospective homepages $\{A_1...A_n\}$, but all of these nodes share a common referring page B_1. Now the instances in our training set are the overlapping subgraphs $\{\{A_1,B_1\}...\{A_n,B_1\}\}$. Thus, the extended features of the nodes $\{A_1...A_n\}$ are not independent. For example, if we know the value of the feature `size-of-referring-page` for node A_1, we know it for all other nodes $\{A_2...A_n\}$. Non-independent instances such as these are likely to arise in any highly interconnected training set.

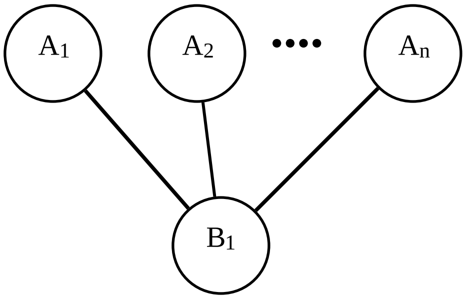

Figure 1b: Dependent Instances

The dependence of instance subgraphs can have a strong effect on the statistical significance of relationships among features in linked data. For example, suppose we wish to evaluate a rule that uses a binary feature of the nodes $\{B_1...B_n\}$ to predict the binary class of nodes $\{A_1...A_n\}$. One such rule would be:

```
  if  name(B,"faculty.html") and
      points-to(B,A)
then  homepage(A)
else  not homepage(A).
```

This particular rule has *a priori* plausibility — a page named "faculty.html" is likely to point to the personal homepages of professors — but we wish to determine the support provided by the training set. We calculate a score *x* for the rule and then determine the statistical significance of that score — $p(X \geq x | H_0)$, the probability of obtaining a score at least as large as *x* under the null hypothesis H_0. In this case, the null hypothesis is that the name of a referring page ("faculty.html") is independent of whether a web document is a homepage.

Suppose the training set contains the instances shown in Figure 1a.[3] Further, suppose the rule above is true for each instance in the training set. What is the probability of this occurring for the instances in Figure 1a under the null hypothesis? For each subgraph, the probability is the sum of two joint events: the probability of jointly matching the "if" and "then" clauses of the rule, and the probability of jointly failing on the "if" clause of the rule and matching the "else" clause. Thus, for all *n* subgraphs in Figure 1a, the probability is:

$$p = (\ p(A)p(B) + p(\neg A)p(\neg B)\ )^n \qquad (1)$$

Where, for notational convenience:

$$\begin{aligned} p &= p(X \geq x | H_0) \\ p(A) &= p(homepage(A)) \\ p(B) &= p(name(B,\text{"faculty.html"}) \text{ and} \\ &\quad\ points\text{-}to(B,A)) \\ p(\neg A) &= 1\text{-}p(A). \end{aligned}$$

In contrast, suppose that the training set contained the instances in Figure 1b, instead of the instances in Figure 1a. Again, suppose the entire training set is consistent with the rule above. The probability of this occurring for the instances in Figure 1b under the null hypothesis is:

$$p = p(A)^n\, p(B) + p(\neg A)^n\, p(\neg B) \qquad (2)$$

The difference between equations 1 and 2 is subtle, but in practice it can result in large disparities in *p*. For example, if $p(A) = 0.1$, $p(B) = 0.4$, and $n = 8$, then $p = 0.013$ for Figure 1a and $p = 0.26$ for Figure 1b.[4] That is, statistical significance calculations that incorrectly assume independent subgraphs can underestimate *p* by more than an order of magnitude.

[3] While the training set may contain many instances in addition to those subgraphs in Figures 1a and 1b, we will focus only on these two potential portions of the training set.

[4] It is not necessarily obvious what feature values make Figures 1a and 1b "the same." For example, for the homepage rule to be true for all instances in Figure 1b, the nodes $\{A_1...A_n\}$ must either all be homepages or not, a condition not required for the rule to be true for all instances in Figure 1a. However, imposing this additional condition on Figure 1a reduces *p* to 0.007, thereby increasing the disparity between the estimates.

General solutions to the statistical challenges introduced by instance linkage include specialized sampling and randomization tests. Sampling techniques for linked data could discard instances that share nodes with other instances, producing data samples similar to Figure 1a. This reduces sample size, but maintains the assumption of instance independence implicit in most induction algorithms. Also, randomization tests (Edgington 1995; Cohen 1995; Jensen 1992) could be used to construct sampling distributions appropriate to particular linked training sets. Randomization tests increase the computational complexity of some portions of induction algorithms by a constant factor (typically 100-1000), but require a minimum of other assumptions.

Sampling Density

In addition to complicating calculations of statistical significance, instance linkage also complicates sampling. The probability distributions of features in linked data depend on *sampling density* — the proportion of the entire population present in a particular training set. Sampling density does not affect probability distributions in traditional training sets because each instance is independent. In linked data, however, fractional sampling densities bias estimates of features such as the "order" of a node (the number of other nodes directly linked to a given node) and the probability of a node having one or more relatives possessing a particular feature value.

For example, consider the nodes and links shown in Figure 2. In the population of all nodes, the central node A_0 has order six (it directly links to six other nodes). However, a sampling density of 0.5 might result in a training set containing only the nodes {A_0, A_1, A_3, A_4, A_6}, shown in Figure 2 in bold. Based on the training set, a naive algorithm would severely underestimate the order of A_0 — in this case, an estimate of 2. Similarly, a naive algorithm would overestimate the number of link traversals required to reach node A_1 from A_0.

Such under- and overestimates can reduce the effectiveness of induction algorithms, which often implicitly assume that probability distributions of features remain constant between training and test sets. This assumption will hold only when sampling densities remain constant between the sets. For example, consider a training set with a sampling density of 0.5 that is used to derive rules for identifying homepages. If the order of a web document was used as one feature indicative of a homepage, then applying such a rule to data with higher sampling densities would bias the rule toward (or away from) predicting homepages, almost certainly reducing overall accuracy.

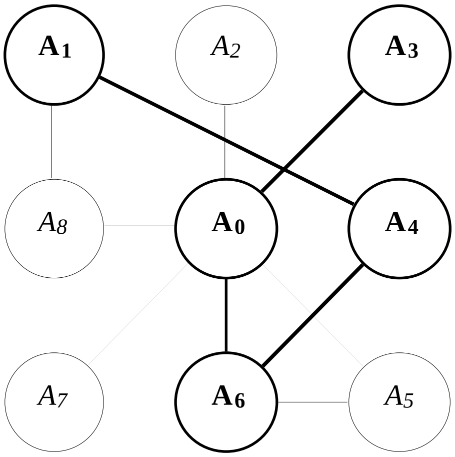

Figure 2: Partial Sampling

This challenge is particularly problematic for some applications of link analysis. For example, law enforcement data about criminal activities is nearly always fragmentary and incomplete. The sampling density can vary greatly from case to case, based on the level of effort expended to gather evidence. Finally, sampling densities are almost never known with certainty.

The challenges associated with sampling density can be partially addressed for some problems by gathering exhaustive samples (sampling density = 1) surrounding independent instances. For example, a set of 100 individual web pages could be sparsely sampled from a very large population of pages, then subnetworks of pages surrounding those 100 could be exhaustively sampled. The surrounding pages would only be used to infer characteristics of the original 100 sparsely-sampled pages. If future test cases are also surrounded by exhaustive samples, problems of under- and overestimation can be avoided. This approach is implicitly used in some inductive logic programming tasks that employ complete molecules or family trees as instances.

Feature Combinatorics

Linked data intensify a challenge already faced by nearly all induction algorithms — adjusting for multiple comparisons. Many induction algorithms use a common procedure: 1) generate n items (e.g., models or model components); 2) calculate a score for each item based on the training set; and 3) select the item with the maximum score (x_{max}). The sampling distribution of x_{max} depends on n, the number items compared, and improperly adjusting for the effects of multiple comparisons can lead to three common pathologies of induction algorithms — overfitting, oversearching, and attribute selection errors (Jensen and Cohen 1998).

Linked data intensify these challenges by increasing the number of possible features. For example, in a standard training set, each instance might have k features. However, if each instance in the training set is linked to only five other instances, there are more than $100k$ features less than four links away. Even relatively modest linking can increase the feature set by several orders of magnitude.

Much higher branching factors can occur in practice. For example, social science research in the 1960s found that any two randomly-selected individuals in the United States could be linked by a chain of six or fewer first-name acquaintances (Milgram 1967). Thus, a link analysis using first-name acquaintance as a linking criterion would have access to over $200x10^6k$ features less than seven links away. In addition, linked data allow construction of entirely new features (e.g., order). Clearly, multiple comparisons can pose a serious challenge to inductive inference in linked data.

Several techniques exist to adjust for multiple comparisons, including new data samples, cross-validation, randomization tests, and Bonferroni adjustment. See Jensen and Cohen (1998) for details and references.

Contributing issues

Several other issues can contribute to the basic challenges outlined above. For example, many link analysis data sets suffer from problems of ambiguous coreference — it is not always apparent when two links reference the same node. For example, this problem can occur when the same web pages can be reached using different URLs. Similarly, in link analysis of financial transactions to detect money laundering, it is not always clear when two financial records reference the same individual (Goldberg and Senator 1995).

Ambiguous coreference can hide occurrences of instance linkage. Two instance subgraphs may appear to be independent, but actually overlap, because two apparently independent nodes actually represent the same object. Such cases make it difficult to properly adjust for the statistical consequences of instance linkage. In addition, ambiguous coreference can mimic the challenges associated with sampling density. If two nodes represent the same object, but each contains only some of the links associated with the complete object, then many of the problems associated with partial sampling densities will be replicated.

Full Paper

The full paper will provide more theoretical analysis, greater detail, and additional examples of three challenges outlined above. We will also explore ambiguous coreference and other issues that intensify the challenges outlined above. Finally, we will discuss implemented solutions to these challenges in the context of a WWW analysis tool we are currently constructing.

References

Cohen, P.R. (1995). *Empirical Methods for Artificial Intelligence*. MIT Press.

Edgington, E.S. (1995). *Randomization Tests*. Marcel Dekker.

Goldberg, H., and T. Senator (1995). Restructuring databases for knowledge discovery by consolidation and link formation. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. AAAI Press.

Jensen, D. (1997). Prospective assessment of AI technologies for fraud detection: A case study. *AI Approaches to Fraud Detection and Risk Management, Collected Papers from the 1997 Workshop*. Technical Report WS-97-07. AAAI Press.

Jensen, D. (1992). *Induction with Randomization Testing*. Dissertation. Washington University.

Jensen, D. and P.R. Cohen (1998). Multiple comparisons in induction algorithms. *Machine Learning* (to appear). Earlier versions of this work were presented at the 1997 Workshop on AI and Statistics.

Milgram, S. (1967). The small-world problem. *Psychology Today* 1(1): 60-76.

Slattery, S. and M. Craven (1998). Learning to exploit document relationships and structure: The case for relational learning on the web. *Proceedings of the Conference on Automated Learning and Discovery*.