

A Brief Survey of Machine Learning Methods for Classification in Networked Data and an Application to Suspicion Scoring

Sofus Attila Macskassy¹ and Foster Provost²

¹ Fetch Technologies,
2041 Rosecrans Ave, Suite 245, El Segundo, CA 90245
sofmac@fetch.com

² New York University,
Stern School of Business, 44 W. 4th Street, New York, NY 10012
fprovost@stern.nyu.edu

Abstract. This paper surveys work from the field of machine learning on the problem of within-network learning and inference. To give motivation and context to the rest of the survey, we start by presenting some (published) applications of within-network inference. After a brief formulation of this problem and a discussion of probabilistic inference in arbitrary networks, we survey machine learning work applied to networked data, along with some important predecessors—mostly from the statistics and pattern recognition literature. We then describe an application of within-network inference in the domain of suspicion scoring in social networks. We close the paper with pointers to toolkits and benchmark data sets used in machine learning research on classification in network data. We hope that such a survey will be a useful resource to workshop participants, and perhaps will be complemented by others.

1 Introduction

This paper briefly surveys work from the field of machine learning, summarizes work in a trio of research papers [1,2,3]. This extended abstract consists of the abstracts for those papers in which we concentrate on methods published in the machine learning literature, as well as methods from other fields that have had considerable impact on the machine learning literature.

Networked data are the special case of relational data where entities are interconnected, such as web-pages or research papers (connected through citations). We focus on *within-network* inference, for which training entities are connected directly to entities whose classifications (*labels*) are to be estimated. This is in contrast to *across-network* inference: learning from one network and applying the learned models to a separate, presumably similar network [4,5]. For within-network inference, networked data have several unique characteristics that both complicate and provide leverage to learning and inference.

Although the network may contain disconnected components, generally there is not a clean separation between the entities for which class membership is known and the entities for which estimations of class membership are to be made. The data are patently not i.i.d., which introduces bias to learning and inference procedures [6]. The usual careful separation of data into training and test sets is difficult, and more importantly, thinking in terms of separating training and test sets obscures an important facet of the data. Entities with known classifications can serve two roles. They act first as training data and subsequently as background knowledge during inference. Relatedly, within-network inference allows models to use specific node identifiers to aid inference [7].

Network data generally allow *collective inference*, meaning that various inter-related values can be inferred simultaneously. For example, inference in Markov random fields [8] uses estimates of a node’s neighbor’s labels to influence the estimation of the nodes labels—and vice versa. Within-network inference complicates such procedures by pinning certain values, but again also offers opportunities such as the application of network-flow algorithms to inference. More generally, network data allow the use of the features of a node’s neighbors, although that must be done with care to avoid greatly increasing estimation variance (and thereby error) [9].

2 Network Learning

Abstract from [1]:

This paper presents NetKit, a modular toolkit for classification in networked data, and a case-study of its application to networked data used in prior machine learning research. We consider *within-network classification*: entities whose classes are to be estimated are linked to entities for which the class is known. NetKit is based on a node-centric framework in which classifiers comprise a local classifier, a relational classifier, and a collective inference procedure. Various existing node-centric relational learning algorithms can be instantiated with appropriate choices for these components, and new combinations of components realize new algorithms. The case study focuses on univariate network classification, for which the only information used is the structure of class linkage in the network (i.e., only links and some class labels). To our knowledge, no work previously has evaluated systematically the power of class-linkage alone for classification in machine learning benchmark data sets. The results demonstrate that very simple network-classification models perform quite well—well enough that they should be used regularly as baseline classifiers for studies of learning with networked data. The simplest method (which performs remarkably well) highlights the close correspondence between several existing methods introduced for different purposes—i.e., Gaussian-field classifiers, Hopfield networks, and relational-neighbor classifiers. The results also show that a small number of component combinations excel. In particular, there are two sets of techniques that are preferable in different situations, namely when few versus many labels

are known initially. We also demonstrate that link selection plays an important role similar to traditional feature selection.

3 Suspicion Scoring

Abstract from [2]:

We describe a guilt-by-association system that can be used to rank entities by their suspiciousness. We demonstrate the algorithm on a suite of data sets generated by a terrorist-world simulator developed under a DoD program. The data sets consist of thousands of people and some known links between them. We show that the system ranks truly malicious individuals highly, even if only relatively few are known to be malicious *ex ante*. When used as a tool for identifying promising data-gathering opportunities, the system focuses on gathering more information about the most suspicious people and thereby increases the density of link-age in appropriate parts of the network. We assess performance under conditions of noisy prior knowledge (score quality varies by data set under moderate noise), and whether augmenting the network with prior scores based on profiling information improves the scoring (it doesn't). Although the level of performance reported here would not support direct action on all data sets, it does recommend the consideration of network-scoring techniques as a new source of evidence in decision making. For example, the system can operate on networks far larger and more complex than could be processed by a human analyst.

Abstract from [3]:

We describe a guilt-by-association system that can be used to rank networked entities by their suspiciousness. We demonstrate the algorithm on a suite of data sets generated by a terrorist-world simulator developed to support a DoD program. Each data set consists of thousands of entities and some known links between them. The system ranks truly malicious entities highly, even if only relatively few are known to be malicious *ex ante*. When used as a tool for identifying promising data-gathering opportunities, the system focuses on gathering more information about the most suspicious entities and thereby increases the density of linkage in appropriate parts of the network. We assess performance under conditions of noisy prior knowledge of maliciousness. Although the levels of performance reported here would not support direct action on all data sets, the results do recommend the consideration of network-scoring techniques as a new source of evidence for decision making. For example, the system can operate on networks far larger and more complex than could be processed by a human analyst. This is a follow-up study to a prior paper; although there is a considerable amount of overlap, here we focus on more data sets and improve the evaluation by identifying entities with high scores simply as an artifact of the data acquisition process.

References

1. Macskassy, S.A., Provost, F.: Classification in Networked Data: A toolkit and a univariate case study. Technical Report CeDER Working Paper 04-08, Stern School of Business, New York University (2004). [June 2006 revision]
2. Macskassy, S.A., Provost, F.: Suspicion scoring based on guilt-by-association, collective inference, and focused data access. In: International Conference on Intelligence Analysis. (2005)
3. Macskassy, S.A., Provost, F.: Suspicion scoring of entities based on guilt-by-association, collective inference, and focused data access. In: Annual Conference of the North American Association for Computational Social and Organizational Science (NAACSOS). (2005)
4. Craven, M., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Quek, C.Y.: Learning to Extract Symbolic Knowledge from the World Wide Web. In: 15th Conference of the American Association for Artificial Intelligence. (1998)
5. Lu, Q., Getoor, L.: Link-Based Classification. In: Proceedings of the 20th International Conference on Machine Learning (ICML). (2003)
6. Jensen, D., Neville, J.: Linkage and Autocorrelation Cause Feature Selection Bias in Relational Learning. In: Proceedings of the 19th International Conference on Machine Learning (ICML). (2002)
7. Perlich, C., Provost, F.: Distribution-based aggregation for relational learning with identifier attributes. *Machine Learning* **62**(1/2) (2006) 65–105
8. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society* **36**(2) (1974) 192–236
9. Jensen, D., Neville, J., Gallagher, B.: Why Collective Inference Improves Relational Classification. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2004)