On the statistical analysis of overlapping observations in large networks

Yuyi Wang and Jan Ramon
Department of Computer Science, KULeuven
yuyi.wang, jan.ramon@cs.kuleuven.be

An important aspect of decision making is to be able to infer unknown properties. The field of supervised learning is concerned with learning models from training data and applying such models to predict unseen properties. Many supervised statistical techniques assume that both training examples and unseen examples are drawn identically and independently (i.i.d.). Recently, there is a growing interest in network data. Unfortunately, in a network observations may not be independent as nodes are connected through edges, violating the important i.i.d. assumption of the majority of supervised learning methods. In our current work, we are addressing this problem by introducing a framework describing these dependencies and extending statistical methods to deal with them.

In a first step of our work [1], we consider pattern mining in large networks. We propose a method to measure the support of a subgraph pattern in a network. This support measure is denoted $\mathsf{s}(G, P)$ with G the network and P the pattern. The s support measure assigns to every occurrence (embedding under subgraph isomorphism) of a pattern P a contribution to its support such that for every vertex $v \in V(G)$ the sum of contributions of all occurrences of P in which v participates is bounded by 1. We show that, under certain assumptions, $\mathsf{s}(G, P)$ can be interpreted as the statistical power of the data in a network for subgraph pattern P. In particular, s gives a measure of statistical power if the occurrences of the pattern are independent of the properties of the nodes.

Consider the following example. We want to analyze the satisfaction of clients in their first lawsuit where they are assisted by a pro-deo lawyer. We construct a network with nodes for pro-deo lawyers, clients, judges and lawsuits, and connect the lawsuits to the lawyer, client and judge appearing in it. The observations of interest are then (lawsuit, lawyer, client, judge) tuples. To every observation we can assign a pair $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i$ is a feature vector containing features of the lawyer and judge, and y_i is the measurement of the satisfaction of the client about the outcome of the lawsuit. The observations are not independent since two lawsuits may share a lawyer or judge.

Above, we stated that s evaluated on a network measures the statistical power if the embeddings of the pattern (connectivity between nodes) is independent of the properties of the participating nodes. If in the above example, for reasons of fairness, judges and pro-deo lawyers are assigned randomly to lawsuits, this assumption holds.

1. Y. Wang and J. Ramon, An efficiently computable support measure for frequent subgraph pattern mining. In proceedings of ECML/PKDD 2012, pp. 362-377, 2012.

A naive method would be to ignore the problem and give all training examples the same weight, but in that case it would be hard to assess the generalization power of the models, as a larger number of observations would not necessarily mean that a more accurate model can be constructed. For example, suppose that a training network has only a limited number of judges and the outcome of a lawsuit depends entirely on the properties of the judge. Then, when the number of lawsuit observations grows above the number of judges, judges participate in several lawsuits and the improvement in model quality will be limited.

Another approach would be to select independent observations (i.e., not sharing a lawyer or judge) and feed them to learning algorithms. However, using the well-known fact that s is larger than the size of the maximum independent set, performing statistical analysis using s we can expect a greater statistical power.

To be more precise, consider the simple task of estimating the expected value $\mu = \mathbb{E}[f(\boldsymbol{x})]$ of a real-valued function $f(\boldsymbol{x})$ over properties of nodes participating in a random observation. In our lawsuit example, f could be the measurement of the satisfaction of the client about the outcome of the lawsuit. f may depend on properties of the lawyer and judge.

Suppose first we estimate μ using an average over a selection of independent observations. Then, if $\sigma^2 = \mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}) - \mu)^2]$ we obtain an estimate of μ with a standard deviation of at least $\sigma/\sqrt{|MIS|}$ where $|MIS|$ is the size of the largest independent set of observations one can construct. Note that computing the set of independent observations of maximum size is NP-hard, and hence in practice one may not be able to achieve such a good standard deviation.

s is the solution to a linear program. To estimate μ, we can make a weighted average μ_{s} over the values of $f(\boldsymbol{x}_i)$ for the several observations $\boldsymbol{x}_i$ in our sample, where we use as weights the contributions to s assigned to the observations by the solution for the linear program to calculate s. We can prove that the standard deviation of our estimate of μ is $(\mathbb{E}[(\mu - \mu_{\mathsf{s}})^2])^{1/2} = \sigma/\sqrt{\mathsf{s}}$. It is well-known that $|MIS| \leq \mathsf{s}$ always holds, so this approach is both better and faster.

In future work we want to consider settings where we can not make such strong independence assumptions that the occurrences of the pattern are independent of the properties of the nodes. Consider another example where students hire rooms and landlords rent rooms. The network contains three types of nodes : students, landlords and rooms. Students with similar properties may study the same topic and want to live close to their own campus. Landlords may specialize in certain types of rooms, apartments, furnishing style etc., so their relation with room properties is not random. As a consequence, given a sample we can not estimate from it the value of a function $f(\boldsymbol{x})$ for a randomly chosen observation with standard deviation as low as $\sigma/\sqrt{\mathsf{s}}$. One major problem is that in this setting, there is no bound on the influence of one particular node. E.g., if there is an important landlord in the training set who is not anymore present in the test set, the matching between students and rooms may shift significantly. A first step in the good direction would be to develop a measure to assess the strength of the dependency of the observations (student-room matchings) on the properties of the nodes (students and rooms), and its influence on the learning task at hand.

Acknowledgement This work was supported by ERC Starting Grant 240186 "MiGraNT : Mining Graphs and Networks : a Theory-based approach".