Identifying Important Nodes in Heterogenous Networks

Oliver Schulte, Fatemeh Riahi, Qing Li
oschulte, sriahi, liqingl@sfu.ca

School of Computing Science
Simon Fraser University
Vancouver-Burnaby, Canada

Abstract

This is a position paper that presents a new approach to identifying important nodes or entities in a complex heterogeneous network. We provide a novel definition of an importance score based on a *statistical model:* An individual is important to the extent that including an individual explicitly in the model improves the data fit of the model more than it increases the model's complexity. We apply techniques from statistical-relational learning, a recent field that combines AI and machine learning, to identify statistically important individuals in a scalable manner. We investigate empirically our approach with the OPTA soccer data set for the English premier league.

Introduction

We present a new approach, based on a statistical model, to identifying important individuals in a complex network. Many, if not most, new datasets contain information about networks whose nodes are linked entities. Identifying important individuals in a network is an important task for network analysis. Our new statistical approach is as follows. First, we learn a *baseline generic statistical model* that describes the dependencies among link types and node features in the network. The generic model refers only to classes of individuals, not to any individual in particular. While adding an individual to the model increases the expressive power of the model, it also increases the number of model parameters and hence the model complexity. A standard *statistical model selection score* quantifies the trade-off between data fit and model complexity. The importance score of an individual is the improvement in the model selection score that results from introducing the individual into the model. For typical statistical scores (e.g., BIC, AIC), the score improvement can be interpreted in minimum description length terms: whereas adding the individual to the model requires extra bits for specifying the new parameter values, it saves bits by fitting the data more closely. Our model class in this paper is Bayes nets, and the statistical score is the Bayes Information criterion (BIC).

Compared to other approaches for ranking individuals in a network, our statistical approach has several advantages. 1) If the statistical score can be evaluated quickly, as is the case with BIC, computing the score improvement associated with an individual is fast. 2) The importance metric is derived from a general metric of predictive power. Because importance is tied to correlations and probabilitic predictions, the metric provides an *explanation* of the ranking. 3) Most previous work assumes a homogeneous network with only one type of node and link (e.g., social network, Twitter, webpages) (Chen et al. 2009). We use models from statistical-relational learning that apply generally to *networks with any number of node link types.* 4) The statistical score provides a discrete decision as to whether the individual is important or not (score improvement > 0), in addition to ranking. This does not require specifying a k-value for selecting the top-k individuals.

We present a preliminary investigation of our approach on premier league soccer data. Here a player is statistically important to the extent that introducing them into a model increases the quality of predicting their team's results and other features of teams and matches.

Related Work

In a Bayesian network model, single-table features correspond to nodes (e.g., age, gender). These feature nodes should not be confused with nodes in the data network that represent individuals (e.g. $Silva$, $Chelsea$) (Neville and Jensen 2007). For single-table data, there has been much work on selecting, fusing, ranking, and scoring features. The majority of this work applies to explicitly listed features (e.g., column headers) that are shared between *independent* individuals. Single-table feature selection is different from the problem we address: 1) We describe a method for introducing *new* features that are not explicitly listed in the data. These new features are of a special type, intuitively "being related to special individual x". 2) In our definition, the importance of an individual x is based on how much being linked to x explains the features of *other* individuals. Thus our scoring method is designed to take into account the interdependence of linked individuals that is the defining aspect of relational data.

The task of identifying important individuals was studied in many contexts such as sparse data university environments (Balog et al. 2007) and for bibliographic data and digital libraries (Deng, King, and Lyu 2008)(Zhou et

al. 2007). Probability models, topic models (Griffiths and Steyvers 2004), vector space (Demartini, Gaugaz, and Nejdl 2009) and voting models have previously been used to rank individuals. Also, HITS (Kleinberg 1999) and PageRank (Page et al. 1999), algorithms were applied for scoring objects in a homogenous network (Hulgeri and Nakhe 2002; Nie et al. 2005); for an extension to heterogeneous networks see (Cao et al. 2012). Several communities have worked on sports data with the goal of predicting match results (Joseph, Fenton, and Neil 2006; Baio and Blangiardo 2010; Vaz de Melo, Almeida, and Loureiro 2008; Onody and de Castro 2004).

Scoring

We use Poole's Parametrized Bayes nets that are defined as follows. The relational structure contains a list of populations $\mathcal{P}_1, \ldots, \mathcal{P}_k$, such as $player, teams, matches$. Population variables such as $Player, Team1, Team2, Match$ are associated with a unique population. A functor is a predicate or function. A **functor node** is of the form $f(\sigma_1, ..., \ldots, \sigma_a)$ where each σ_i is a constant or variable of the appropriate population. A Parametrized Bayes net is a Bayes net whose nodes are functor nodes. The state-of-the-art learn-and-join algorithm (Schulte and Khosravi 2012) takes as input (1) a relational database $\mathcal{D}$ representing a network, (2) a set of functor nodes, and produces a Bayes net for the functor nodes. The learn-and-join algorithm includes a method for extracting a default set of functor nodes from a relational database schema; they can also be chosen by the user.

The user chooses a statistical score $score(B, \mathcal{D})$ that scores a Parametrized Bayes net B for a database $\mathcal{D}$. In our experiments, we used the relational Bayes Information Criterion (BIC) (Schulte 2011; Alsanie and Cussens 2012). We evaluate the **score improvement** due to a target individual t as follows. Let t be a constant denoting an individual that instantiates population variable X, with associated population $\mathcal{P}$. Let $\mathcal{D}t$ be the database where the population of X is restricted to the single member t. Let $\mathcal{D}_t^-$ be the database where t is removed from the population of X.

1. Learn a generic model $B_\mathcal{D}$ for the entire database.
2. Apply Bayes net learning to (1) input database $\mathcal{D}_t^+$, and (2) the functor nodes that have X replaced by t. Call the result B_t.
3. The score improvement is given by

$$[\frac{1}{|\mathcal{P}|} score(B_t, \mathcal{D}_t) + \frac{|\mathcal{P}| - 1}{|\mathcal{P}|} score(B_\mathcal{D}, \mathcal{D}_t^-)] - score(B_\mathcal{D}, \mathcal{D}).$$

The score improvement formula can be interpreted as follows. Suppose that we randomly select a member x of the population X. There are two cases: 1) $x = t$ is the target individual. In that case we use the score for the target's model B_t applied to the data describing the target and its links, which is represented by $\mathcal{D}_t$. 2) $x \neq t$ is different from the target individual. In that case we use the score for the generic model $B_\mathcal{D}$ applied to the data describing the population without t, which is represented by $\mathcal{D}_t^-$. The first case occurs with probability $1/|\mathcal{P}|$ and the second with probability $(|\mathcal{P}| - 1)/|\mathcal{P}|$. Therefore the expected score using the individual as well as the generic model is

$$\frac{1}{|\mathcal{P}|} score(B_t, \mathcal{D}_t) + \frac{|\mathcal{P}| - 1}{|\mathcal{P}|} score(B_\mathcal{D}, \mathcal{D}_t^-).$$

The score improvement formula compares the score for the two models to the score for using only the generic model for all individuals, which is given by $score(B_\mathcal{D}, \mathcal{D})$.

Complexity. Typical statistical scores such as BIC can be computed in closed-form given the sufficient statistics. In the case of Bayes nets these are the counts of joint child-parent states, which can be described by conjunctive queries. The complexity of evaluating scores is therefore essentially the complexity of computing the frequency of conjunctive queries in a database. Most Bayes net learners follow a score-based approach where candidate models are repeatedly evaluated by applying the score. The cost of applying the score once to evaluate the score improvement is therefore dominated by the cost of learning the Bayes net models. Current state-of-the-art Bayes net learners for relational data scale well to databases with table sizes on the order of 10^5 (Schulte and Khosravi 2012); extending the scope of scalable relational Bayes net learning is an active research area.

Examples. To build specific models for important players, in the Bayes net of Figure 1(c), we can replace the variable $Player$ by $Player = Nasri$ (b). The database $\mathcal{D}_t$ contains only rows where team = $ManchesterCity$ (MC) and player = Nasri. The database $\mathcal{D}_t^-$ contains the rows for all the other players of MC. Figure (a) illustrates the result of the same procedure for $Player = Silva$.

To build specific models for important teams, we can replace the variable $Team$ in the generic model of Figure (left) by MC.

Dataset and Results

The dataset in this paper is the Opta data, released by Manchester City. It is a time coded feed that lists all the ball actions within each game by each player from 2011 to 2012. Number of goals, passes, fouls, tackles, saves and blocks and also position assigned to a player in a match are examples of the information associated with each player. The information can be visualized as a heterogeneous network that links players to teams, and teams to matches.

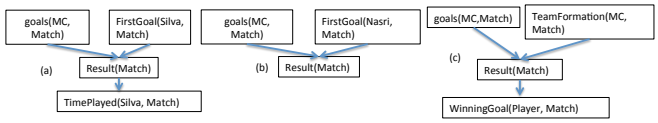

Figure 1: Generic Bayes Net for Manchester City (MC) and the special models for their players Nasri and Silva.

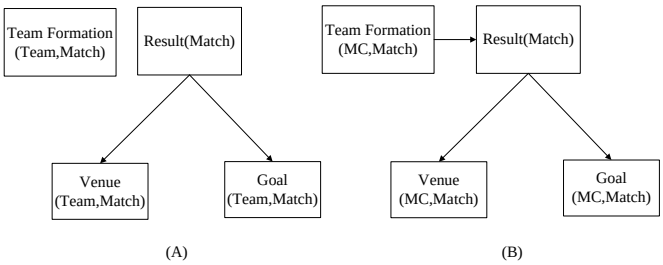

Figure 2: Generic Bayes Net for Teams and the special model for $Team = MC$ (Manchester City). In the generic model, team formation does not predict the result, but in the specific model for Manchester City it does.

Name	Position	Score Improvement	Predicted WinPercentage	Actual WinPercentage	PWP/AVT	Salary
Edin Dzeko	striker	39.58	0.766	0.76	0.0153	56750
Carlos Tevez	striker	49.227	0.663	0.76	0.0143	55800
Mario Balotelli	striker	3.523	0.737	0.75	0.0128	55650
Sergio Aguero	striker	18.712	0.807	0.82	0.0105	55550
James Milner	Mid-Fielder	68.95	0.769	0.76	0.0126	54350
David Silva	Mid-Fielder	98.324	0.7870	0.80	0.01	54400
Samir Nasri	Mid-Fielder	79.223	0.632	0.80	0.008	54150

Table 1: Data for strikers and mid-fielders of Manchester City: Model Score Improvement, Expected percentage of wins when player plays (model estimate) = PWP, Actual Percentage, PWP/Average Time played, salary.

Scoring Players

Figure 1 shows special models built for two players of Manchester City (MC). In the generic MC model, scoring the first goal does not predict the result but there is a correlation if Nasri or Silva score it. The length of time played by Silva positively correlates with higher results, but not for Nasri. This illustrates how the Bayes net analysis can find qualitative differences between individuals. We compare the player's importance score with a simple measure of their value to the team: how the MC average number of wins changes given that they play (WinPercentage). The average number of MC wins is 78%. The BN general population estimates the winning percentage at 70%. The Predicted WinPercentage column shows that this estimate is improved for each player by building a specific model (except for Nasri). The last two columns in Table 1 show that if we divide each player's WinPercentage by their average time played, there is a strong correlation with salary ($r = 0.813$). The table shows data for the strikers and midfielders for whom we could obtain salary data.

Scoring Teams

Table 2 shows that building a specific Bayes net for teams with high importance score allows the model to make more precise predictions for the teams results.

Discussion. In general, our method applies to any network that can be represented in a relational database schema. In network terms, the nodes in the Bayes net models in the soccer domain concern correlations among attributes of links and entities. For instance, Figure 1 models a relationship Player-Appears-In-Match, with attribute $FirstGoal$. The model shows how this can predict the result attribute of a Match. In information networks, we may have few or no attributes, but many links of different types. In principle, the importance score applies also to networks with link information only. Whether the importance score improves modelling of link structure (as opposed to attribute correlations) is a topic for future work.

TeamName	Score.Imp.	Exp.Res.	Exp.Res.Diff
Swansea City	6.535	0.2971	0.0899
Tottenham Hotspur	33.0191	0.3947	0.0076
Bolton Wanderers	6.483	0.2632	0.1238
Manchester City	84.128	0.7895	0.4025
Everton	85.85	0.4211	0.0341

Table 2: Data for premier league teams with significant score improvement: Score Improvement, Expected result of the team (as estimated by the specific model), Expected Result Difference from population mean over all teams (0.38).

Conclusion

We introduced a new statistical approach to identifying important individuals in a heterogenous network. The importance scores are fast to compute. The score results point to qualitative differences between individuals, and improve estimates of a player's contribution. These estimates correlate strongly with contribution metrics that are independent of the score (e.g., player salary). Questions for future work include defining a discriminative version of our importance score, how to identify clusters of statistically similar players, and how to combine the generic Bayes net and the individual Bayes nets into a single Bayesian hierarchical model (Spiegelhalter et al. 1996)(Gyftodimos and Flach).

References

Alsanie, W., and Cussens, J. 2012. Learning recursive prism programs with observed outcomes. In *Proceedings of the ICML-2012 Workshop on Statistical-Relational Learning*.

Baio, G., and Blangiardo, M. 2010. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics* 37(2):253–264.

Balog, K.; Bogers, T.; Azzopardi, L.; de Rijke, M.; and van den Bosch, A. 2007. Broad expertise retrieval in sparse data environments. SIGIR '07, 551–558. New York, NY, USA: ACM.

Cao, L.; Jin, X.; Yin, Z.; Pozo, A. D.; Luo, J.; Han, J.; and Huang, T. S. 2012. Rankcompete: Simultaneous ranking and clustering of information networks. *Neurocomputing* 95:98–104.

Chen, H.; Liu, H.; Han, J.; and Yin, X. 2009. Exploring optimization of semantic relationship graph for multi-relational Bayesian classification. *Decision Support Systems* 48(1):112–121.

Demartini, G.; Gaugaz, J.; and Nejdl, W. 2009. A vector space model for ranking entities and its application to expert search. In *31st European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, 189–201. Berlin, Heidelberg: Springer-Verlag.

Deng, H.; King, I.; and Lyu, M. R. 2008. Formal models for expert finding on dblp bibliography data. In *In ICDM*, 163–172.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101:5228–5235.

Gyftodimos, E., and Flach, P. Hierarchical bayesian networks: A probabilistic reasoning model for structured domains.

Hulgeri, A., and Nakhe, C. 2002. Keyword searching and browsing in databases using banks. In *Proceedings of the 18th International Conference on Data Engineering*, ICDE, 431–. Washington, DC, USA: IEEE Computer Society.

Joseph, A.; Fenton, N. E.; and Neil, M. 2006. Predicting football results using bayesian nets and other machine learning techniques. *Know.-Based Syst.* 19(7):544–553.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46(5):604–632.

ManchesterCity. 2012. Url:http://www.mcfc.com/.

Neville, J., and Jensen, D. 2007. Relational dependency networks. In *Introduction to Statistical Relational Learning*. MIT Press. chapter 8, 239–268.

Nie, Z.; Zhang, Y.; Wen, J.-R.; and Ma, W.-Y. 2005. Object-level ranking: bringing order to web objects. In *Proceedings of the 14th international conference on World Wide Web*, 567–574. New York, NY, USA: ACM.

Onody, R. N., and de Castro, P. A. 2004. Complex network study of brazilian soccer players.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66. Previous number = SIDL-WP-1999-0120.

Schulte, O., and Khosravi, H. 2012. Learning graphical models for relational data via lattice search. *Machine Learning* 88(3):331–368.

Schulte, O.; Khosravi, H.; Kirkpatrick, A.; Gao, T.; and Zhu, Y. 2012. Modelling relational statistics with bayes nets. In *Inductive Logic Programming (ILP)*.

Schulte, O. 2011. A tractable pseudo-likelihood function for Bayes nets applied to relational data. In *SIAM SDM*, 462–473.

Spiegelhalter, D.; Thomas, A.; Best, N.; and Gilks, W. 1996. Bugs: Bayesian inference using gibbs sampling. Technical report, Institute of Public Health, Cambridge, UK.

Vaz de Melo, P. O.; Almeida, V. A.; and Loureiro, A. A. 2008. Can complex network metrics predict the behavior of nba teams? KDD '08, 695–703. ACM.

Zhou, D.; Orshanskiy, S. A.; Zha, H.; and Giles, C. L. 2007. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, 739–744. IEEE Computer Society.