CERIAS Tech Report 2011-25
Ensemble Classification for Relational Domains
by Hoda Eldardiry
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Ensemble Classification for Relational Domains

Hoda Eldardiry
Computer Science Department
Purdue Universiry
305 North University Street
West Lafayette, Indiana 47907, USA

Ensemble classification methods have been shown to produce more accurate predictions than the *base* component models (Bauer and Kohavi 1999). Due to their effectiveness, ensemble approaches have been applied in a wide range of domains to improve classification. The expected prediction error of classification models can be decomposed into bias and variance (Friedman 1997). Ensemble methods that *independently* construct component models (e.g., bagging) can improve performance by reducing the error due to *variance*, while methods that *dependently* construct component models (e.g., boosting) can improve performance by reducing the error due to *bias* and variance.

Although ensemble methods were initially developed for classification of independent and identically distributed (i.i.d.) data, they can be directly applied for relational data by using a relational classifier as the base component model. This straightforward approach can improve classification for network data, but suffers from a number of limitations. First, relational data characteristics will only be exploited by the base *relational* classifier, and not by the ensemble algorithm itself. We note that explicitly accounting for the structured nature of relational data by the ensemble mechanism can significantly improve ensemble classification. Second, ensemble learning methods that assume i.i.d. data can fail to preserve the relational structure of non-i.i.d. data, which will (1) prevent the relational base classifiers from exploiting these structures, and (2) fail to accurately capture properties of the dataset, which can lead to inaccurate models and classifications. Third, ensemble mechanisms that assume i.i.d. data are limited to reducing errors associated with i.i.d. models and fail to reduce additional sources of error associated with more powerful (e.g., collective classification (Sen et al. 2008)) models. Our key observation is that collective classification methods have error due to variance in inference (Neville and Jensen 2008). This has been overlooked by current ensemble methods that assume exact inference methods and only focus on the typical goal of reducing errors due to learning, even if the methods explicitly consider relational data (Preisach and Schmidt-Thieme 2006).

Here we study the problem of ensemble classification for relational domains by focusing on the reduction of error due to variance. We propose a relational ensemble framework that explicitly accounts for the structured nature of relational data during both learning and inference. Our proposed framework consists of two components. (1) A method for learning accurate ensembles from relational data, focusing on the reduction of error due to variance in learning, while preserving the relational characteristics in the data. (2) A method for applying ensembles in collective classification contexts, focusing on further reduction of the error due to variance in inference, which has not been considered in state of the art ensemble methods.

Our first contribution is a relational resampling method that can be used to construct ensembles from relational data to significantly improve classification accuracy of *bagging* (Breiman 1996) in relational domains (Eldardiry and Neville 2008). Our key insight is that independent sampling with replacement (typically used by bagging) can underestimate variance when applied to relational datasets. This can be explained as follows. In a relational dataset, dependencies among interrelated objects form clusters of objects with correlated attributes, which reduces the number of independent instances (e.g., from the number of objects towards the number of clusters). This reduces the effective sample size and leads to increased variance (Jensen and Neville 2002). *Independent sampling* from relational data then overestimates the sample size and thus underestimates the variance. When the bootstrap samples, and consequently the models learned on them, do not accurately capture the variance in the data, bagging fails to fully reduce error due to variance. Additionally, independent sampling from a relational dataset does not preserve the relational structure, since a node will not necessarily have all of its neighbors selected for the sample. This limits fully exploiting relational dependencies to improve classification. Empirical results show that bagging using our proposed relational resampling method significantly outperforms two baselines including bagging using i.i.d. resampling, on both synthetic and real-world datasets. In addition, bagging using our resampling mechanism can better exploit increased *autocorrelation* in the data due to preserving the relational structure during sampling. Furthermore, we explicitly show that our proposed resampling method can estimate the variance of a statistic in relational data more accurately than i.i.d. resampling.

Our second contribution consists of an ensemble *inference* approach for relational domains. We make the key

observation that *collective classification* models in statistical relational learning suffer from two sources of variance error (Neville and Jensen 2008). Collective classification methods (Sen et al. 2008) learn a model of the dependencies in a relational graph (e.g., social network) and then apply the learned model to *collectively* (i.e., jointly) infer the unknown class labels in the graph. The first source of error for these models is the typical variance due to *learning*—as variation in the data used for estimation causes variation in the learned models. The second source of error is due to variance in *inference*—since predictions are propagated throughout the network during inference, variation due to approximate inference and variation in the test data can both increase prediction variance and thus increase error. We develop a relational ensemble framework that uses a novel form of *across-model* collective inference for collective classification (Eldardiry and Neville 2011). Since collective inference models exploit the structure in relational data by propagating inferences across links–when a network consists of multiple relationship graphs, an ensemble of collective inference models can be learned from the available link structures. We explore this idea on network datasets with multiple link graphs, by learning a separate model from each link graph and applying them for inference using a novel across-model collective inference method. Our method propagates inference information across simultaneous collective inference processes running on the base models of the ensemble to reduce *inference variance*. Then the algorithm combines the final model predictions to reduce *learning variance*. This is the first ensemble technique that aims to reduce error due to inference variance. Empirical results show that our ensemble inference method significantly outperforms all baselines under a variety of conditions, using synthetic and real-world datasets. We explicitly show that the increased accuracy is due to the inference variance reduction achieved by our proposed method. We also show that our algorithm best exploits autocorrelation, linkage, and additional base models. Additionally, our method is the most robust to missing labels due to its ability to best exploit the available label information.

In future work, we plan to develop a unified ensemble classification framework for relational domains which combines the two proposed methods to reduce errors due to variance in both learning and inference. This can be achieved by using the proposed relational resampling method to construct the ensembles, and the proposed across-model collective inference method for ensemble inference. Using relational resampling instead of separate link structures for learning will make our proposed framework applicable in single-source network settings besides multiple-source ones. In addition, we will conduct a bias/variance analysis of the proposed ensemble framework, with the goal of theoretically confirming the following conjectures, which are supported by initial empirical evidence: (1) bagging using relational resampling significantly outperforms bagging using i.i.d. resampling due to more accurate capturing, and therefore reduction, of learning variance, and (2) across-model collective ensemble classification significantly outperforms the baseline relational ensemble approach due to reduction of variance in both learning and inference. Finally, we plan to define and prove theoretical properties of relational ensemble classification.

There are several lines of works related to our proposed research. Previous work on sampling from structured data has investigated graph-based *reuse sampling* techniques for lattice graphs, which use small, overlapping subgraphs as pseudosamples (Hall and Jing 1996). This work is related to our work on resampling. However, this work was developed for spatial and temporal datasets, where the link structure is generally homogeneous, and will not work for relational data since it will be difficult to determine the effective sample size of a relational dataset analytically due to the heterogeneous link structure. Related to our ensemble algorithm, Preisach and Schmidt-Thieme (2006) learn a separate logistic regression classifier from each relational source then combine the classifiers using voting and stacking. This is similar to our method since it uses an ensemble approach to combine models learned from different link types. However, their method was not designed for collective classification models, thus the approach is intended to reduce learning error, not inference error. We compare to a baseline method that uses this straightforward relational ensemble approach and our method does significantly better since it reduces both types of error as opposed to just learning error. The work of Gao et al. (2009) presents a method to maximize consensus among the decisions of multiple supervised and unsupervised models. This method is similar to our approach since it combines predictions from multiple models and uses label propagation for prediction. However, we note that the label propagation mechanism used is designed to maximize consensus among the model outputs after inference, rather than during a collective inference process over a relational network. In addition, the method is designed primarily for i.i.d. learners where again, there will be no inference error.

References

Bauer, E., and Kohavi, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning* 36.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Eldardiry, H., and Neville, J. 2008. A resampling technique for relational data graphs. In *Proc. of SNA-SIGKDD'08.*

Eldardiry, H., and Neville, J. 2011. Across-model collective ensemble classification. In *Proc. of AAAI'11*, to appear.

Friedman, J. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* 1(1).

Gao, J.; Liang, F.; Fan, W.; Sun, Y.; and Han, J. 2009. Graph-based consensus maximization among multiple supervised and unsupervised models. In *Proc. of NIPS'09.*

Hall, P., and Jing, B. 1996. On sample reuse methods for dependent data. *Journal of the Royal Statistical Society, Series B* 58:727–737.

Jensen, D., and Neville, J. 2002. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proc. of ICML'02.*

Neville, J., and Jensen, D. 2008. A bias/variance decomposition for models using collective inference. *Machine Learning Journal.*

Preisach, C., and Schmidt-Thieme, L. 2006. Relational ensemble classification. In *Proc. of ICDM'06.*

Sen, P.; Namata, G. M.; Bilgic, M.; Getoor, L.; Gallagher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI Magazine* 29(3):93–106.