Privacy-Preserving Approach to Bayesian Network Structure Learning from Distributed Data

Olivier Regnier-Coudert
IDEAS Research Institute
Robert Gordon University
Aberdeen, UK
o.regnier-coudert@rgu.ac.uk

John McCall
IDEAS Research Institute
Robert Gordon University
Aberdeen, UK
j.mccall@rgu.ac.uk

ABSTRACT

In many situations, data is scattered across different sites, making the modeling process difficult or sometimes impossible. Some applications could benefit from collaborations between organisations but data security or privacy policies often act as a barrier to data mining on such contexts.

In this paper, we present a novel approach to learning Bayesian Networks (BN) structures from multiple datasets, based on the use of Ensembles and an Island Model Genetic Algorithm (IMGA). The proposed design ensures no data is shared during the process and can fit many applications.

Categories and Subject Descriptors

I.6.5 [**Simulation and Modeling**]: Model Development —*Modeling methodologies*; I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search—*Graph and tree search strategies, Heuristic methods*

General Terms

Algorithms

Keywords

Bayesian Networks, Genetic Algorithm, Island Model, Distributed Data Mining

1. INTRODUCTION

Distributed data mining (DDM) has become an important area of research. Growing data storage capacities and the increase of data collections in organizations have lead to having data often spread across many sites.

Most of the traditional data modeling tools are developed to be trained from a centralized dataset. In order to take advantage of this amount of data, data modeling tools need to be adapted. The second hurdle that may prevent models from being built is linked with data restrictions that can exist between sites. These can be related to commercial sensitivity of the data such as the information that can be shared between industrial operators and maintenance contractors; or related to individual privacy, such as patient confidentiality. In this paper, we explain how the representation domain of Bayesian Networks (BN) [3] can be used to address these issues.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

2. DISTRIBUTED DATA AND BAYESIAN NETWORKS

In most Evolutionary Algorithms (EA) applications that aim to find the best BN from data, solutions are represented in a consistent manner. A BN is composed of two main components. Its structure reflects the different conditional dependencies that may exist between the variables while parameters (or conditional probability tables) quantify these dependencies. In EAs, a solution is represented as an ordering of variables (nodes), where a node cannot be assigned as a parent a node located at a previous position [3]. This ensures solutions produced will not have cycles, one of the requirements needed with BN. From a given ordering, greedy heuristics such as the K2 algorithm [2] can be used to obtain the final structure that fits best the data. The scoring function used in K2 reflects how likely a network is to generate the same dataset.

The challenge of learning BN from distributed data has been approached in several ways. In [1], local BNs are constructed at each site before being recombined to produce a global model. Despite retrieving similar structure as centralized methods, such approach presents flaws with respect to data privacy, as some selected instances are shared to compute the final parameters. Focus on privacy preservation is higher in [6] where a method to compute K2 score and the parameters from distant sites without data sharing is presented. Other approaches including cryptography or noise addition have also been proposed for different kinds of data mining techniques but no method has yet been implemented which focuses on the exchange of BN structures [6].

3. PROPOSED APPROACH

The proposed approach to distributed BN learning focuses on taking advantage of the domain representation by mean of ensembles and IMGA as illustrated in Figure 1.

3.1 Island Model Genetic Algorithm

In a typical GA, solutions are evolved based on operators applied after some selection on the population. A GA ideally stops when convergence is reached. With the introduction of IMGA [5], evolution is performed in parallel on different populations. Selected solutions are introduced in neighboring populations when migration occurs at some set

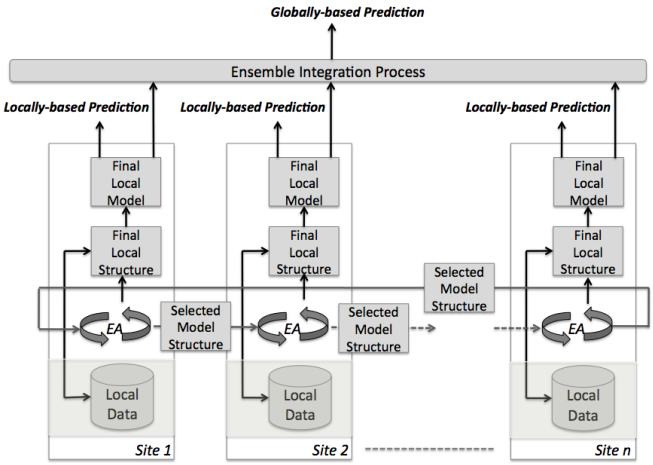

Figure 1: Ensemble of IMGA-based BNs

intervals. This process increases the chance for evolution to differ between islands, exploring a larger portion of the search space, but also reduces the likelihood to have early convergence in each population, or at each site.

With respect to privacy preservation, it is a complex task to compute the K2 score from different locations. However, BN structures can easily be learnt locally before being transmitted between sites during IMGA migration. Resulting BN structures at each site will be based on local data but will also consider neighboring structures which bring information on data distribution from other sites. Parameter learning can then be done applying the local data on the learnt structure. Alternatively, more advanced techniques such as in [6] can be chosen.

3.2 Ensembles for Decision Making

In many applications, local models are not of major interest. Ensembles can be implemented to integrate a pool of local models to produce a global prediction. In this prospect, they have shown that their use can lead to prediction of higher quality than centralized models [4]. Such results are obtained when a high coverage is reached among local models, that is when each of them wrongly classify different instances. Influencing what structures to introduce in a population and when to do so can have a direct impact on the errors that will be made by the resulting model. With this in mind, optimizing IMGA migration is likely to influence the ensemble coverage and in the mean time the quality of the global model.

In Figure 1, we have illustrated how global prediction can be achieved based on the proposed IMGA approach. Integration of the local BNs will represent another important aspect of the implementation. It can be done using the local classification through selection, such as voting, or through combination. In the latter, global classification is made based upon the local outcome prediction probabilities.

3.3 Evaluation

The proposed design raises challenges in many area of data mining. It is important to evaluate it on a wide range of problems. Well known benchmarks for BNs such as ASIA and ALARM will be used to assess the quality of the models obtained using different settings. Evaluation needs to focus on the quality of the structure found by comparing learned networks using the new approach with the ones obtained from centralized data. Classification performance will be assessed using common methods such as cross-validation and ROC analysis. In addition, Kullback-Leibler distance represents a good way to quantify the difference of distribution between the probabilities of the distributed and centralized implementations.

Our current research activity also focuses on medical treatment optimization and more precisely prostate cancer stage prediction. Retrospective data on patients treated in the UK with prostate cancer has been collected by the British Association of Urological Surgeons and is available for the project. Due to its confidential aspect and to the presence of information concerning treatment centers, this dataset is particularly well adapted to the problem. By distributing data according to the center numbers, it represents a real-world application to the proposed design where both local and global predictions are considered as important.

4. CONCLUSIONS

We propose a novel approach to handle distributed data while maintaining requirements on data privacy. It contributes to the research on DDM in two main aspects. First, regardless of classification performances, we proposed a way to learn BN structures without sharing data or adding noise which can prove useful for applications where data relationship understanding is the priority. Second, and because of the use of ensembles, such design is promising with respect to the improvement of the quality of the classification. Setting algorithms and evaluations will be an essential part of the implementation to reach satisfactory results.

5. REFERENCES

[1] R. Chen, K. Sivakumar, and H. Kargupta. Collective mining of Bayesian networks from distributed heterogeneous data. *Knowledge and Information Systems*, 6(2):164–187, 2004.

[2] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[3] R. Kabli, F. Herrmann, and J. McCall. A chain-model genetic algorithm for bayesian network structure learning. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1271–1278. ACM, 2007.

[4] A. Tsymbal, S. Puuronen, and D. Patterson. Ensemble feature selection with the simple Bayesian classification. *Information Fusion*, 4(2):87–100, 2003.

[5] D. Whitley, S. Rana, and R. Heckendorn. The island model genetic algorithm: On separability, population size and convergence. *Journal of Computing and Information Technology*, 7:33–48, 1999.

[6] Z. Yang and R. Wright. Privacy-preserving computation of Bayesian networks on vertically partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, pages 1253–1264, 2006.