Improving Bayesian Learning Using Public Knowledge

Farid Seifi[1], Chris Drummond[2], Nathalie Japkowicz[1] and Stan Matwin[1,3]

[1] School of Information Technology and Engineering
University of Ottawa
Ottawa Ontario Canada, K1N 6N5
fseif050@uottawa.ca,nat@site.uottawa.ca,stan@site.uottawa.ca

[2] Institute for Information Technology, National Research Council Canada, Ottawa, Ontario, Canada, K1A 0R6, Chris.Drummond@nrc-cnrc.gc.ca

[3] Institute for Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract. Both intensional and extensional background knowledge have previously been used in inductive problems to complement the training set used for a task. In this research, we propose to explore the usefulness, for inductive learning, of a new kind of intensional background knowledge: the inter-relationships or conditional probability distributions between subsets of attributes. Such information could be mined from publicly available knowledge sources but including only some of the attributes involved in the inductive task at hand. The purpose of our work is to show how this information can be useful in inductive tasks, and under what circumstances. We will consider injection of background knowledge into Bayesian Networks and explore its effectiveness on training sets of different sizes. We show that this additional knowledge not only improves the estimate of classification accuracy it also reduces the variance in the accuracy of the model.

Key words: Bayesian Networks, Public Knowledge, Classification.

1 Introduction

While standard machine learning acquires knowledge from instances of the learning problem, there has always been interest in a more cognitively plausible scenario in which learning - besides the training instances - utilizes also background knowledge. In many inductive problems, the training set, which is a set of labeled samples, could be complemented using intensional or extensional background knowledge in order to improve the learning performance [5, 9]. In Inductive Logic Programming, intensional background knowledge is provided in the form of a theory expressed in logical form. In Semi-Supervised Learning, the extensional background knowledge is provided in the form of unlabeled data.

In this research, we propose to explore a different type of intensional background knowledge. In many domains, there exist publicly available very large, and related, data sets, for example from demographics and statistical surveys.

This sort of information is ubiquitous: it is published by many national governments, international organizations, and private companies. Such data sets may not have exactly the same attributes as the data set we are studying. However, using an intensionalising process [6], we can derive intensional background knowledge, in the form of distributions, from this extensional background knowledge, given as collections of instances. A question that we consider here is whether it is possible to use such information to improve machine learning methods.

Let us consider a simple medical example. Suppose we are learning from data a model for the prediction of heart attacks in patients. The data used in the inductive learning of this model may include attributes describing sleep disturbance, as a disease outcome, and stress, as a disease, but does not include enough instances to relate these attributes in a statistically significant way. There exists, independently of the data used in model building, a large medical survey that describes quantitatively sleep disturbance in patients who experience cardiac problems or stress. This set could be used in learning a better predictive model, capturing the important relationship between sleep disturbance, stress, and a heart attack, if we can integrate the data from the medical study with the data we are using in learning the predictive model.

The big challenge in this research is how such background knowledge can be integrated with the existing data sets. Bayesian learning is a natural candidate as it draws on distributional data for its assessment of the probabilities of an instance belonging to different classes of the concept. In Bayesian Networks the attribute inter-relationships are encoded into a network structure. We propose here to replace parts of this structure, some of the conditional probability distributions, with more accurate alternatives, which are available as background knowledge contained in large public data sets, e.g. statistical surveys.

The paper is structured as follows: Section 2 discusses how background knowledge is added to the network. In Section 3 experiments and discussions are provided. Section 4 contains conclusion. A more detailed version of our work is availavle at [1].

2 Injecting Public Knowledge into a Network

In a Bayesian network [8, 7], there is a structure which encodes a set of conditional independence assumptions between attributes; a node is conditionally independent of its non-descendants given its parents. Also, there are conditional probability distributions capturing each attribute's dependency on others, typically represented by multi-dimensional tables. Together, these define the joint probability distribution of the attributes and class. With such a distribution, we can use Bayes rule to do inference. There exist many different ways of building Bayesian networks from training data. We used the software package *BN predictor* [3, 4] to build the network and used a maximum likelihood estimator (frequency counts) to construct the tables.

Normally, we obtain the conditional probability distributions which we use in Bayesian Network inference from the training set. If we do not have enough training data samples, our estimates of the true distribution will be poor and the result will not be an accurate classifier. These distributions are independent, so it could be possible to improve the performance even by replacement of a few of them with accurate alternatives, obtained from statistical surveys.

We propose improving Bayesian networks by replacing some of the conditional probability distributions - represented in the form of tables and corresponding to the edges of the network - with their accurate alternatives which are available as background knowledge.

3 Experiments and Discussions

For concluding wheather the replacement of a selected set of distributions makes a better classifier or not we run a set of experiments. In each experiment a large data set as well as a training and testing sets are sampled from the huge data set which represents the universe. Then the classifier is trained using the small training set. More specifically, potentially inaccurate conditional probability distributions are built from the training set. Instead of using statistical surveys to extract accurate distributions, we use the distributions which were obtained from the large data set. Then we replace the selected set of distributions with accurate alternatives and compute the performance of the new modified classifier. We run several experiments with the same replacements and then we use paired t-test to see whether these sets of replacements make a significantly better classifier or not. Our experiments show that replacing more distributions results in a more accurate classifier unless a distribution is not extracted based on correct attribute dependencies. The Letter data set from the UCI machine learning repository [2] is used as the real data set. In addition, an artificial data set from the heart attack domain is used in a second experiment.

Our experiments on letter data set show that replacements of the conditional probability distributions with accurate alternatives make significantly better classifiers. We run several experiments, as explained above, all with the same replacements. The accuracies of these experiments on the modified models are compared with unmodified ones, using paired t-test, to show that those specific replacements made a significantly better classifier. In all cases, the variance of the accuracies of the modified model is smaller than that of the unmodified model. This means that when we replace a conditional probability distribution in a Bayesian Network with an accurate alternative, the new model tends to be more robust when sampling new data sets for training and testing.

We have tested the effect of replacement of different permutations of conditional probability distributions in the letter data set. For all, except one, replacements we experienced that modified model is significantly better. The results are obtained using the paired t-test with 95% confidence interval.

Replacing the conditional probability distribution of one table, which is related to attribute *xegvy*, with the accurate alternative we obtain a less accurate classifier. One reason is, according to attribute evaluators (such as information gain and chi square), this attribute has less effect on the results of classification than others. The effectiveness of replacement of the conditional probability distribution of an attribute is directly related to the correctness of all its conditional dependencies. Therefore, another reason for this negative result is that the conditional dependencies of attribute *xegvy* is not extracted correctly.

Again, our experiments on artificial heart attack data set show that the replacement of conditional probability distributions, which are in the form of tables, with accurate alternatives, makes significantly better classifiers. For this purpose we run several experiments each with an individual replacement. The modified models for some cases are statistically significant and for some extremely statistically significant. The results of these experiments again show that the variance in the accuracy of the modified model is smaller than the variance in the accuracy of the unmodified model. It is also experienced that using incomplete or wrong dependencies for an attribute may led to a not statistically significant classifier.

4 Conclusion

In this study we propose a practical method for improving Bayesian classifiers by using background knowledge from large, publicly available datasets existing independently of the training data set. We present a method which manipulates the Bayesian Network's conditional probability distributions, given in the form of tables, based on background knowledge. The idea is tested on a real and an artificial data set. The results show that such changes produce significantly better classifiers than normal Bayesian Network classifiers.

References

1. Detailed Long Paper. `http://www.archive.org/details/ImprovingBayesianLearningUsingPublicKnowledge`.
2. C. Blake and C. Merz. UCI repository of machine learning databases. Univ. of California at Irvine. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
3. J. Cheng. BN powerpredictor. `http://webdocs.cs.ualberta.ca/~jcheng/bnsoft.htm`.
4. J. Cheng and R. Greiner. Learning bayesian belief network classifiers: algorithms and system. *LNCS*, 2056:141–151, 2001.
5. P. Clark and S. Matwin. Learning domain theories using abstract beckground knowledge. In P. Brazdil, editor, *ECML*, volume 667 of *Lecture Notes in Computer Science*, pages 360–365. Springer, 1993.
6. P. A. Flach. From extensional to intensional knowledge: Inductive logic programming techniques and their application to deductive databases. *Transactions and Change in Logic Databases, LNCS*, 1472:356–387, 1998.
7. D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.
8. T. M. Mitchell. *Machine learning.* McGraw Hill, 1997.
9. P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In C. E. Brodley, editor, *ICML*, volume 69 of *ACM International Conference Proceeding Series.* ACM, 2004.