Markov blanket discovery in positive-unlabelled and semi-supervised data

Konstantinos Sechidis and Gavin Brown

School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{konstantinos.sechidis,gavin.brown}@manchester.ac.uk

Abstract. The importance of Markov blanket discovery algorithms is twofold: as the main building block in constraint-based structure learning of Bayesian network algorithms and as a technique to derive the optimal set of features in filter feature selection approaches. Equally, learning from partially labelled data is a crucial and demanding area of machine learning, and extending techniques from fully to partially supervised scenarios is a challenging problem. While there are many different algorithms to derive the Markov blanket of fully supervised nodes, the partially-labelled problem is far more challenging, and there is a lack of principled approaches in the literature. Our work derives a generalization of the conditional tests of independence for partially labelled binary target variables, which can handle the two main partially labelled scenarios: *positive-unlabelled* and *semi-supervised*. The result is a significantly deeper understanding of how to control false negative errors in Markov Blanket discovery procedures and how unlabelled data can help.

Keywords: Markov blanket discovery; Partially labelled; Positive unlabelled; Semi supervised; Mutual information

1 Introduction

Markov Blanket (MB) is an important concept that links two of the main activities of machine learning: dimensionality reduction and learning. Using Pellet & Elisseef's [15] wording "Feature selection and causal structure learning are related by a common concept: the Markov blanket."

Koller & Sahami [10] showed that the MB of a target variable is the optimal set of features for prediction. In this context discovering MB can be useful for eliminating irrelevant features or features that are redundant in the context of others, and as a result plays a fundamental role in filter *feature selection*. Furthermore, Markov blankets are important in learning Bayesian networks [14], and can also play an important role in *causal structure learning* [15].

In most real world applications, it is easier and cheaper to collect unlabelled examples than labelled ones, so transferring techniques from fully to *partial-labelled* datasets is a key challenge. Our work shows how we can recover the MB around partially labelled targets. Since the main building block of the MB discovery algorithms is the *conditional test of independence*, we will present a

The final publication is available at:
http://link.springer.com/chapter/10.1007/978-3-319-23528-8_22

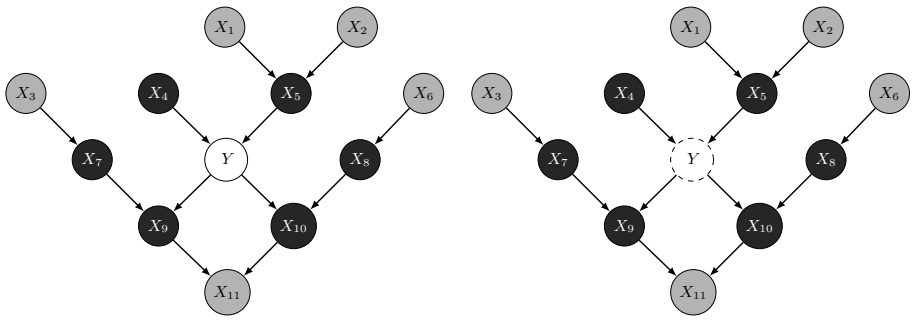
method to apply this test despite the partial labelling and how we can use the unlabelled examples in an informative way.

Section 4 explores the scenario of *positive-unlabelled* data. This is a special case of partially-labelling, where we have few labelled examples *only* from the positive class and a vast amount of unlabelled examples. Section 5 extends our work to *semi-supervised* data, where the labelled set contains examples from both classes. Finally, Section 6 presents a semi-supervised scenario that can occur in real world, the *class prior change* scenario, and shows how our approach performs better than the state of the art [1].

Before the formal presentation of the background material (Sections 2 and 3) we will motivate our work with a toy Bayesian network presented in Figure 1. The MB of the target variable Y is the feature set that contains the *parents* (X_4 and X_5), *children* (X_9 and X_{10}) and *spouses* (X_7 and X_8, which are other parents of a child of Y) of the target. There exist many techniques to derive MB by using fully-supervised datasets, Figure 1(a). But our work will focus on partially labelled scenarios where we have the values of Y only for a small subset of examples, Figure 1(b), while all the other variables are completely observed. We will suggest ways to derive the MB by controlling the two possible errors in the discovery procedure:

Falsely adding variables to the predicted Markov blanket: for example assuming that the variable X_{11} *belongs* to MB.

Falsely not adding variables to the predicted Markov blanket: for example assuming the variable X_4 *does not belong* to MB.



(a) MB in fully supervised target (b) MB in partially labelled target

Fig. 1: Toy Markov blanket example where: white nodes represent the target variable, black ones the features that *belong* to the MB of the target and grey ones the features that *do not belong* to the MB. In (a) we know the value of the target over all examples, while in (b) the target is partially observed (dashed circle) meaning that we know its value only in a small subset of the examples.

[1] Matlab code and the supplementary material with all the proofs are available in `www.cs.man.ac.uk/~gbrown/partiallylabelled/`.

2 Background: Markov blanket

In this section we will introduce the notation and the background material on Markov blanket discovery algorithms. Assuming that we have a binary classification dataset $\mathcal{D} = \{(\mathbf{x}^i, y^i)|i = 1, ..., N\}$, where the target variable Y takes the value y^+ when the example is positive, and y^- when the example is negative. The feature vector $\mathbf{x} = [x_1...x_d]$ is a realization of the d-dimensional joint random variable $\mathbf{X} = X_1...X_d$. With a slight abuse of notation, in the rest of our work, we interchange the symbol for a set of variables and for their joint random variable. Following Pearl [14] we have the following definitions.

Definition 1 (Markov blanket — Markov boundary).
The Markov blanket of the target Y is a set of features $\mathbf{X}_{\text{MB}}$ with the property $Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}_{\text{MB}}$ for every $\mathbf{Z} \subseteq \mathbf{X}\backslash\mathbf{X}_{\text{MB}}$. A set is called Markov boundary if it is a minimal Markov blanket, i.e. non of its subsets is a Markov blanket.

In probabilistic graphical models terminology, the target variable Y becomes conditionally independent from the rest of the graph $\mathbf{X}\backslash\mathbf{X}_{\text{MB}}$ given its Markov blanket $\mathbf{X}_{\text{MB}}$.

Learning the Markov blanket for each variable of the dataset, or in other words inferring the local structure, can naturally lead to *causal structure learning* [15]. Apart from playing a huge role in structure learning of a Bayesian network, Markov blanket is also related to another important machine learning activity: *feature selection*.

Koller & Sahami [10] published the first work about the optimality of Markov blanket in the context of feature selection. Recently, Brown et al. [5] introduced a unifying probabilistic framework and showed that many heuristically suggested feature selection criteria, including Markov blanket discovery algorithms, can be seen as iterative maximizers of a clearly specified objective function: the conditional likelihood of the training examples.

2.1 Markov blanket discovery algorithms

Margaritis & Thrun [12] introduced the first theoretically sound algorithm for Markov blanket discovery, the Grow-Shrink (GS) algorithm. This algorithm consists of two-stages: *growing* where we add features to the Candidate Markov Blanket (CMB) set until the point that the remaining features are independent with the target given the candidate blanket, and *shrinkage*, where we remove potential false positives from the CMB. Tsamardinos & Aliferis [21] suggested an improved version to this approach, the Incremental Association Markov Blanket (IAMB), which can be seen in Algorithm 1. Many measures of association have been used to decide which feature will be added in the candidate blanket during the growing phase (Alg. 1 - Line 4), with the main being the *conditional mutual information* [17]. But, Yaramakala & Margaritis [23] suggested the use of the *significance of the conditional test of independence*, which is more appropriate in statistical terms than the raw conditional mutual information value. Finally,

there is another class of algorithms that try to control the size of conditioning test in a two-phase procedure: firstly identify parent-children, then identify spouses. The most representative algorithms are the HITON [2] and the Max-Min Markov Blanket (MMMB) [22]. All of these algorithms assume faithfulness of the data distribution. As we already saw, in all Markov blanket discovery algorithms, the conditional test of independence (Alg. 1 - Line 5 and 11) plays a crucial role, and this is the focus of the next paragraph.

Algorithm 1: Incremental Association Markov Blanket (IAMB)

Input : Target Y, Features $\mathbf{X} = X_1 ... X_d$, Significance level α
Output: Markov Blanket: $\mathbf{X}_{\text{CMB}}$

1 Phase I: forward — growing
2 $\mathbf{X}_{\text{CMB}} = \emptyset$
3 **while** $\mathbf{X}_{\text{CMB}}$ has changed **do**
4 Find $X \in \mathbf{X} \backslash \mathbf{X}_{\text{CMB}}$ most strongly related with Y given $\mathbf{X}_{\text{CMB}}$
5 **if** $X \not\perp\!\!\!\perp Y | \mathbf{X}_{\text{CMB}}$ using significance level α **then**
6 Add X to $\mathbf{X}_{\text{CMB}}$
7 **end**
8 **end**
9 Phase II: backward — shrinkage
10 **foreach** $X \in \mathbf{X}_{\text{CMB}}$ **do**
11 **if** $X \perp\!\!\!\perp Y | \mathbf{X}_{\text{CMB}} \backslash X$ using significance level α **then**
12 Remove X from $\mathbf{X}_{\text{CMB}}$
13 **end**
14 **end**

2.2 Testing conditional independence in categorical data

IAMB needs to test the conditional independence of X and Y given a subset of features $\mathbf{Z}$, where in Line 5 $\mathbf{Z} = \mathbf{X}_{\text{CMB}}$ while in Line 11 $\mathbf{Z} = \mathbf{X}_{\text{CMB}} \backslash X$. In fully observed categorical data we can use the G-test, a generalised likelihood ratio test, where the test statistic can be calculated from sample data counts arranged in a contingency table [1].

***G*-statistic:** We denote by $O_{x,y,\mathbf{z}}$ the observed count of the number of times the random variable X takes on the value x from its alphabet $\mathcal{X}$, Y takes on $y \in \mathcal{Y}$ and $\mathbf{Z}$ takes on $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$, where $\mathbf{z}$ is a vector of values when we condition on more than one variable. Furthermore denote by $O_{x,.,\mathbf{z}}$, $O_{.,y,\mathbf{z}}$ and $O_{.,.,\mathbf{z}}$ the marginal counts. The estimated expected frequency of $(x, y, \mathbf{z})$, assuming X, Y are conditional independent given $\mathbf{Z}$, is given by $E_{x,y,\mathbf{z}} = \frac{O_{x,.,\mathbf{z}} O_{.,y,\mathbf{z}}}{O_{.,.,\mathbf{z}}} = \widehat{p}(x|\mathbf{z})\widehat{p}(y|\mathbf{z})O_{.,.,\mathbf{z}}$. To calculate the G-statistic we use the following formula:

$$\widehat{G}\text{-statistic} = 2 \sum_{x,y,\mathbf{z}} O_{x,y,\mathbf{z}} \ln \frac{O_{x,y,\mathbf{z}}}{E_{x,y,\mathbf{z}}} = 2 \sum_{x,y,\mathbf{z}} O_{x,y,\mathbf{z}} \ln \frac{O_{.,.,\mathbf{z}} O_{x,y,\mathbf{z}}}{O_{x,.,\mathbf{z}} O_{.,y,\mathbf{z}}} = \quad (1)$$

$$= 2N \sum_{x,y,\mathbf{z}} \widehat{p}(x,y,\mathbf{z}) \ln \frac{\widehat{p}(x,y|\mathbf{z})}{\widehat{p}(x|\mathbf{z})\widehat{p}(y|\mathbf{z})} = 2N\widehat{I}(X;Y|\mathbf{Z}), \quad (2)$$

where $\widehat{I}(X;Y|\mathbf{Z})$ is the maximum likelihood estimator of the conditional mutual information between X and Y given $\mathbf{Z}$ [8].

Hypothesis testing procedure: Under the null hypothesis that X and Y are statistically independent given $\mathbf{Z}$, the G-statistic is known to be asymptotically χ^2-distributed, with $\nu = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)|\boldsymbol{\mathcal{Z}}|$ degrees of freedom [1]. Knowing that and using (2) we can calculate the $\widehat{p}_{XY|\mathbf{Z}}$ value as $1 - F(\widehat{G})$, where F is the CDF of the χ^2-distribution and $\widehat{G}$ the observed value of the G-statistic. The p-value represents the probability of obtaining a test statistic equal or more extreme than the observed one, given that the null hypothesis holds. After calculating this value, we check to see whether it exceeds a significance level α. If $p_{XY|\mathbf{Z}} \leq \alpha$, we reject the null hypothesis, otherwise we fail to reject it. This is the procedure that we follow to take the decision in Lines 5 and 11 of the IAMB algorithm 1. Furthermore, to choose the most strongly related feature in Line 4, we evaluate the p-values and choose the feature with the smaller one.

Different types of error: Following this testing procedure, two possible types of error can occur. The significance level α defines the probability of *Type I error* or False Positive rate, that the test will reject the null hypothesis when the null hypothesis is in fact true. While the probability of *Type II error* or False Negative rate, which is denoted by β, is the probability that the test will fail to reject the null hypothesis when the alternative hypothesis is true and there is an actual effect in our data. Type II error is closely related with the concept of statistical *power* of a test, which is the probability that the test will reject the null hypothesis when the alternative hypothesis is true, i.e. $power = 1 - \beta$.

Power analysis: With such a test, it is common to perform an *a-priori power analysis* [7], where we would take a given sample size N, a required significance level α, an effect size ω, and would then compute the power of the statistical test to detect the given effect size. In order to do this we need a test statistic with a known distribution under the alternative hypothesis. Under the alternative hypothesis (i.e. when X and Y are dependent given $\mathbf{Z}$), the G-statistic has a large-sample *non-central* χ^2 distribution [1, Section 16.3.5]. The non-centrality parameter (λ) of this distribution has the same form as the G-statistic, but with sample values replaced by population values, $\lambda = 2NI(X;Y|\mathbf{Z})$. The effect size of the G-test can be naturally expressed as a function of the *conditional mutual information*, since according to Cohen [7] the effect size (ω) is the square root of the non-centrality parameter divided by the sample, thus we have $\omega = \sqrt{2I(X;Y|\mathbf{Z})}$.

Sample size determination: One important usage of a-priori power analysis is *sample size determination*. In this prospective procedure we specify the probability of Type I error (e.g. $\alpha = 0.05$), the desired probability of Type II error (e.g. $\beta = 0.01$ or $power = 0.99$) and the desired effect size that we want to observe, and we can determine the minimum number of examples (N) that we need to detect that effect.

2.3 Suggested approach for semi-supervised MB discovery

To the best of our knowledge, there is only one algorithm to derive the MB of semi-supervised targets: BASSUM (BAyesian Semi-SUpervised Method) [6]. BASSUM follows the HITON approach, finding firstly the parent-child nodes and then the spouses, and tries to take into account both labelled and unlabelled data. BASSUM makes the "traditional semi-supervised" assumption that the labelled set is an unbiased sample of the overall population, and it uses the unlabelled examples in order to improve the reliability of the conditional independence tests. For example to estimate the G-statistic, in equation (1), it uses both labelled and unlabelled data for the observed counts $O_{.,.,\mathbf{z}}$ and $O_{x,.,\mathbf{z}}$. This technique is known in statistics as *available case analysis* or *pairwise deletion*, and is affected by the ambiguity over the definition of the overall sample size, which is crucial for deriving standard errors and the sampling distributions (the reader can find more details on this issue in Allison [3, page 8]). This can lead to unpredictable results, for example there are no guarantees that the G-statistic will follow χ^2 distribution after this substitution. Another weakness of BASSUM is that it cannot be applied in partially labelled environments where we have the restriction that the labelled examples come only from one class, such as the positive-unlabelled data. In order to explore the Markov blanket of this type of data we should explore how to test conditional independence in this scenario and this is the focus of Section 4. Before that, we will formally introduce the partially-labelled data in the following section.

3 Background: Partially-labelled data

In this section we will give the background for the two partially-labelled problems on which we will focus: positive-unlabelled and semi-supervised.

3.1 Positive-unlabelled data

Positive-Unlabelled (PU) data refers to situations where we have a small number of labelled examples from the positive class, and a large number of entirely unlabelled examples, which could be either positive *or* negative. For reasoning over PU data we will follow the formal framework of Elkan & Noto [9]. Assume that a dataset $\mathcal{D}$ is drawn i.i.d. from the joint distribution $p(\mathbf{X}, Y, S_P)$, where $\mathbf{X}$ and Y are random variables describing the feature set and the target variable, while S_P is a further random variable with possible values 's_P^+' and 's_P^-', indicating if the positive example is labelled (s_P^+) or not (s_P^-). We sample a total number of N examples out of which $N_{S_P^+}$ are labelled as positives. Thus $p(\mathbf{x}|s_P^+)$ is the probability of $\mathbf{X}$ taking the value $\mathbf{x}$ from its alphabet $\boldsymbol{\mathcal{X}}$ conditioned on the labelled set. In this context, Elkan & Noto formalise the *selected completely at random* assumption, stating that the examples for the labelled set are selected completely at random from all the positive examples:

$$p(s_P^+|\mathbf{x}, y^+) = p(s_P^+|y^+) \quad \forall \ \mathbf{x} \in \boldsymbol{\mathcal{X}}. \tag{3}$$

Building upon this assumption, Sechidis et al. [19] proved that we can test independence between a feature X and the unobservable variable Y, by simply testing the independence between X and the observable variable S_P, which can be seen as a *surrogate* version of Y. While this assumption is sufficient for testing independence and guarantees the same probability of false positives, it leads to a less powerful test, and the probability of committing a false negative error is increased by a factor which can be calculated using prior knowledge over $p(y^+)$. With our current work we extend this approach to test conditional independence.

3.2 Semi-supervised data

Semi-Supervised (SS) data refer to situations where we have a small number of labelled examples from *both* classes and a large number of unlabelled examples. For reasoning over semi-supervised data we will follow the formal framework of Smith & Elkan [20]. Assuming that the dataset is drawn i.i.d. from the joint distribution $p(\mathbf{X}, Y, S)$, where S describes whether an example is labelled (s^+) or not (s^-). We sample a total number of N examples out of which N_{S^+} are labelled as positive or negative. Smith & Elkan [20] presented the "traditional semi-supervised" scenario, where the labels are missing completely at random (MCAR), so the labelled set is an unbiased sample of the population. But apart from this traditional scenario, there are many alternative scenarios that can lead to semi-supervised data [20]. In our work, we will focus on the scenario where labelling an example is conditionally independent of the features given the class:

$$p(s^+|\mathbf{x}, y) = p(s^+|y) \quad \forall \ \ \mathbf{x} \in \boldsymbol{\mathcal{X}}, y \in \mathcal{Y}. \tag{4}$$

This assumption can be seen as a straightforward extension of the selected completely at random assumption in the semi-supervised scenario, and it is followed in numerous semi-supervised works [11, 16, 18]. A practical application where we can use this assumption is in *class-prior-change* scenario [16], which occurs when the class balance in the labelled set does not reflect the population class balance. This sampling bias is created because the labels are missing not at random (MNAR), and the missingness mechanism depends directly on the class. The "traditional semi-supervised" assumption is as a restricted version of the assumption described in equation (4), when we furthermore assume $p(s^+|y) = p(s^+) \ \ \forall \ \ y \in \mathcal{Y}$.

4 Markov blanket discovery in positive-unlabelled data

In this section we present a novel methodology for testing conditional independence in PU data. We will then see how we can use this methodology to derive Markov blanket despite the labelling restriction.

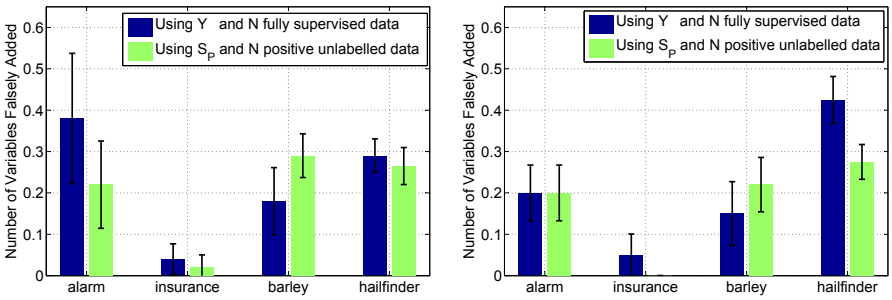
4.1 Testing conditional independence in PU data

With the following theorem we prove that a valid approach to test conditional independence is to assume all unlabelled examples to be negative and as a result use the surrogate variable S_P instead of the unobservable Y.

Theorem 1 (Testing conditional independence in PU data).
In the positive unlabelled scenario, under the selected completely at random assumption, a variable X is independent of the class label Y given a subset of features $\mathbf{Z}$ if and only if X is independent of S_P given $\mathbf{Z}$, so it holds:

$$X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp S_P|\mathbf{Z}.$$

The proof of the theorem is available in the supplementary material. Now we will verify the consequences of this theorem in the context of Markov blanket discovery. We use four widely used networks; Appendix A contains all details on data generation and on the experimental protocol. For these networks we know the true Markov blankets and we compare them with the discovered blankets through the IAMB algorithm. As we observe from Figure 2 using S_P instead of Y in the IAMB algorithm does not result to a statistical significant difference in the false positive rate, or in Markov blanket terminology the blankets derived from these two approaches are similar in terms of the variables that were *falsely added to the blanket*.



(a) $N = 2000,\ p(s_P^+) = 0.05$ (b) $N = 5000,\ p(s_P^+) = 0.05$

Fig. 2: Verification of Theorem 1. This illustrates the average number of variables falsely added in MB and the 95% confidence intervals over 10 trials when we use IAMB with Y and S_P. (a) for total sample size $N = 2000$ out of which we label only 100 positive examples and (b) for total sample size $N = 5000$ out of which we label only 250 positives.

But, while Theorem 1 tells us that the probability of committing an error is the same for the two tests $G(X; Y|\mathbf{Z})$ and $G(X; S_P|\mathbf{Z})$ when $X \perp\!\!\!\perp Y|\mathbf{Z}$, it does not say anything about the performance of these tests when the variables are conditionally dependent. In this case, we should compare the *power* of the tests, and in order to do so we should explore the non-centrality parameters of the two conditional G-tests of independence.

Theorem 2 (Power of PU conditional test of independence).
In the positive unlabelled scenario, under the selected completely at random assumption, when a variable X is dependent on the class label Y given a subset of features $\mathbf{Z}$, $X \not\perp\!\!\!\perp Y|\mathbf{Z}$, we have: $I(X; Y|\mathbf{Z}) > I(X; S_P|\mathbf{Z})$.

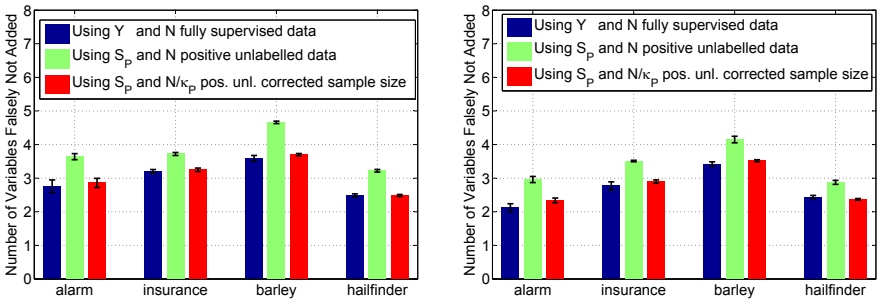
While with the following theorem we will quantify the amount of power that we are loosing with the naive assumption of all unlabelled examples being negative.

Theorem 3 (Correction factor for PU test).
The non-centrality parameter of the conditional G-test between X and S_P given a subset of features $\mathbf{Z}$ takes the form:

$$\lambda_{G(X;S_P|\mathbf{Z})} = \kappa_P \lambda_{G(X;Y|\mathbf{Z})} = \kappa_P 2NI(X;Y|\mathbf{Z}),$$

where $\kappa_P = \frac{1-p(y^+)}{p(y^+)} \frac{p(s_P^+)}{1-p(s_P^+)} = \frac{1-p(y^+)}{p(y^+)} \frac{N_{S_P^+}}{N-N_{S_P^+}}$.

The proofs of the last two theorems are also available in the supplementary material. So, by using prior knowledge over the $p(y^+)$ we can use the naive test for sample size determination, and decide the amount of data that we need in order to have similar performance with the unobservable fully supervised test. Now we will illustrate the last theorems again in the context of MB discovery. A direct consequence of Theorem 2 is that using S_P instead of Y results in a higher number of *false negative* errors. In the MB discovery context this will result in a larger number of variables falsely not added to the predicted blanket, since we assumed that the variables were independent when in fact they were dependent. In order to verify experimentally this conclusion we will compare again the discovered blankets by using S_P instead of Y. As we see in Figure 3, the number of variables that were falsely not added is higher when we are using S_P. This Figure also verifies Theorem 3, where we see that the number of variables falsely removed when using the naive test $G(X;S_P|\mathbf{Z})$ with increased sample size N/κ_P is the same as when using the unobservable test $G(X;Y|\mathbf{Z})$ with N data.

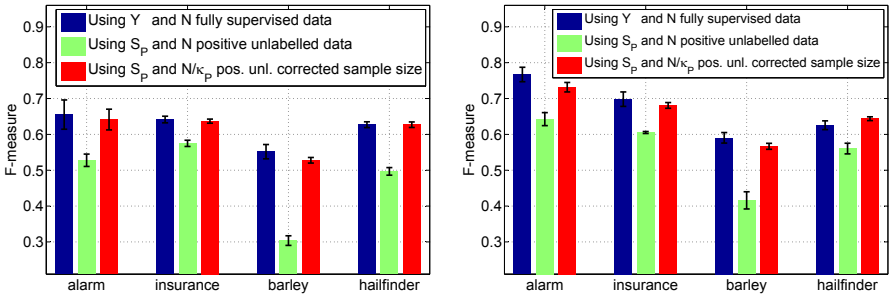


(a) $N = 2000,\ p(s_P^+) = 0.05$ (b) $N = 5000,\ p(s_P^+) = 0.05$

Fig. 3: Verification of Theorems 2 and 3. This illustrates the average number of variables falsely not added to the MB and the 95% confidence intervals over 10 trials when we use IAMB with Y and S_P. (a) for total sample size $N = 2000$ and (b) for total sample size $N = 5000$. In all the scenarios we label 5% of the total examples as positives.

4.2 Evaluation of Markov blanket discovery in PU data

For an overall evaluation of the derived blankets using S_P instead of Y we will use the F-measure, which is the harmonic mean of precision and recall, against the ground truth [17]. In Figure 4, we observe that the assumption of all unlabelled examples to be negative gives worse results than the fully-supervised scenario, and that the difference between the two approaches gets smaller as we increase sample size. Furthermore, using the correction factor κ_P to increase the sample size of the naive approach makes the two techniques perform similar.



(a) $N = 2000,\ p(s_P^+) = 0.05$ (b) $N = 5000,\ p(s_P^+) = 0.05$

Fig. 4: Comparing the performance in terms of F-measure when we use IAMB with Y and S_P. (a) for total sample size $N = 2000$ and (b) for total sample size $N = 5000$. In all the scenarios we label 5% of the total examples as positives.

5 Markov blanket discovery in semi-supervised data

In this section we will present two informative ways, in terms of power, to test conditional independence in semi-supervised data. Then we will suggest an algorithm for Markov blanket discovery where we will incorporate prior knowledge to choose the optimal way for testing conditional independence.

5.1 Testing conditional independence in semi-supervised data

We will introduce two variables in the semi-supervised scenario, which can be seen as noisy versions of the unobservable random variable Y. The first one is S_P, which we already used in the PU scenario, and is a binary random variable that takes the value s_P^+ when a positive example is labelled, and s_P^- in any other case. The second variable is S_N, which is also a binary random variable that takes the value s_N^+ when a negative example is labelled and s_N^- otherwise. Using these two variables, the selected completely at random assumptions described in equation (4) can be written as:

$$p(s_P^+|\mathbf{x}, y^+) = p(s_P^+|y^+) \ \text{ and } \ p(s_N^+|\mathbf{x}, y^-) = p(s_N^+|y^-) \quad \forall \ \mathbf{x} \in \boldsymbol{\mathcal{X}}.$$

So, using S_P instead of Y is equivalent to making the assumption that all unlabelled examples are negative, as we did in the positive-unlabelled scenario, while using S_N instead of Y is equivalent to assuming all unlabelled examples being positive. In this section we will prove the versions of the three theorems we presented earlier for both variables S_P and S_N in the semi-supervised scenario.

Firstly we will show that testing conditional independence by assuming the unlabelled examples to be either positive or negative is a valid approach.

Theorem 4 (Testing conditional independence in SS data).
In the semi-supervised scenario, under the selected completely at random assumption, a variable X is independent of the class label Y given a subset of features $\mathbf{Z}$ if and only if X is independent of S_P given $\mathbf{Z}$ and the same result holds for S_N: $X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp S_P|\mathbf{Z}$ and $X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp S_N|\mathbf{Z}$.

Proof. Since the selected completely at random assumption holds for both classes, this theorem is a direct consequence of Theorem 1.

The consequence of this assumption is that the derived conditional tests of independence are less powerful that the unobservable fully supervised test, as we prove with the following theorem.

Theorem 5 (Power of the SS conditional tests of independence).
In the semi-supervised scenario, under the selected completely at random assumption, when a variable X is dependent of the class label Y given a subset of features $\mathbf{Z}$, $X \not\perp\!\!\!\perp Y|\mathbf{Z}$, we have: $I(X;Y|\mathbf{Z}) > I(X;S_P|\mathbf{Z})$ and $I(X;Y|\mathbf{Z}) > I(X;S_N|\mathbf{Z})$.

Proof. Since the selected completely at random assumption holds for both classes, this theorem is a direct consequence of Theorem 2.

Finally, with the following theorem we can quantify the amount of power that we are loosing by assuming that the unlabelled examples are negative (i.e. using S_P) or positive (i.e. using S_N).

Theorem 6 (Correction factors for SS tests).
The non-centrality parameter of the conditional G-test can take the form:

$$\lambda_{G(X;S_P|\mathbf{Z})} = \kappa_P \lambda_{G(X;Y|\mathbf{Z})} = \kappa_P 2NI(X;Y|\mathbf{Z}) \quad \textit{and}$$
$$\lambda_{G(X;S_N|\mathbf{Z})} = \kappa_N \lambda_{G(X;Y|\mathbf{Z})} = \kappa_N 2NI(X;Y|\mathbf{Z}),$$

where $\kappa_P = \frac{1-p(y^+)}{p(y^+)} \frac{p(s_P^+)}{1-p(s_P^+)}$ *and* $\kappa_N = \frac{p(y^+)}{1-p(y^+)} \frac{p(s_N^+)}{1-p(s_N^+)}$.

Proof. Since the selected completely at random assumption holds for both classes, this theorem is a direct consequence of Theorem 3.

5.2 Incorporating prior knowledge on Markov blanket discovery

Since using S_P or S_N are both valid approaches it is preferable to use the most powerful test. In order to do so, we can use some "soft" prior knowledge over

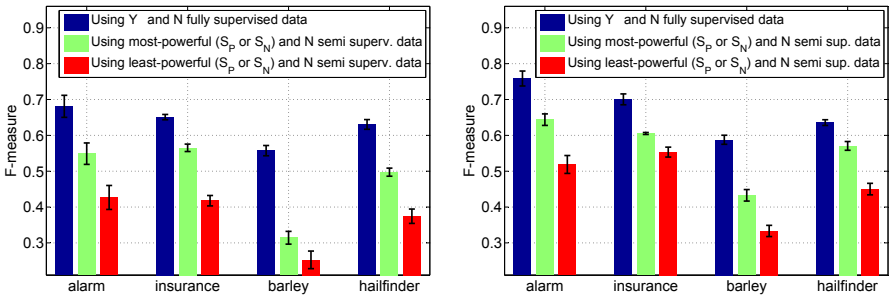
the probability $p(y^+)$[2]. We call it "soft" because there is no need to know the exact value, but we only need to know if it is greater or smaller than a quantity calculated from the observed dataset. The following corollary gives more details.

Corollary 1 (Incorporating prior knowledge).
In order to have the smallest number of falsely missing variables from the Markov Blanket we should use S_P instead of S_N, when the following inequality holds

$$\kappa_P > \kappa_N \Leftrightarrow p(y^+) < \frac{1}{1+\sqrt{\frac{(1-p(s_P^+))p(s_N^+)}{p(s_P^+)(1-p(s_N^+))}}}.$$

When the opposing inequality holds the most powerful choice is S_N. When equality holds, both approaches are equivalent.

We can estimate $p(s_P^+)$ and $p(s_N^+)$ from the observed data, and, using some prior knowledge over $p(y^+)$, we can decide the most powerful option. In Figure 5 we compare in terms of F-measure the derived Markov blankets when we use the most powerful and the least powerful choice. As we observe by incorporating prior knowledge as Corollary 1 describes, choosing to test with the most powerful option, results in remarkably better performance than with the least powerful option.



(a) $N = 2000,\ N_{S_P^+} = 100,\ N_{S_N^+} = 100$ (b) $N = 5000,\ N_{S_P^+} = 250,\ N_{S_N^+} = 250$

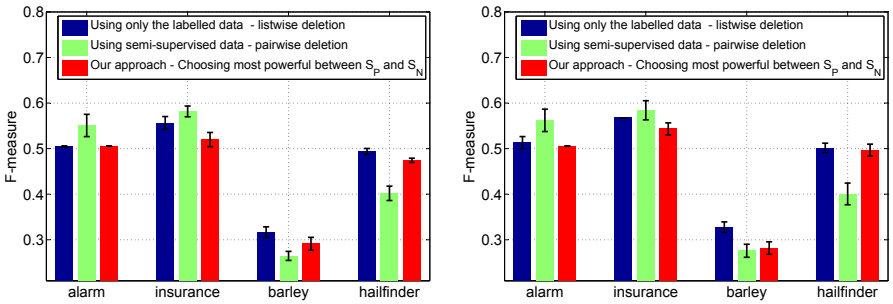
Fig. 5: Comparing the performance in terms of F-measure when we use the unobservable variable Y and the most and least powerful choice between S_P and S_N. (a) for sample size $N = 2000$ out of which we label only 100 positive and 100 negative examples and (b) for sample size $N = 5000$ out of which we label only 250 positive and 250 negative examples.

[2] When the labelling depends directly in the class, eq. (4), we cannot have an unbiased estimator for this probability without further assumptions, more details in [16].

6 Exploring our framework under class prior change — When and how the unlabelled data help

In this section, we will present how our approach performs in a real world problem where the class balance in the labelled set does not reflect the balance over the overall population; such situation is known as *class-prior-change* [16]. We compare our framework with the following two approaches: ignoring the unlabelled examples, a procedure known in statistic as *listwise deletion* [3], or using the unlabelled data to have more reliable estimates for the marginal counts of the features, a procedure known in statistics as *available case analysis* or *pairwise deletion* [3]. The latter is followed in BASSUM [6]; Section 2.3 gives more details about this approach and its limitations.

Firstly, let's assume that the semi-supervised data are generated under the "traditional semi-supervised" scenario, where the labelled set is an unbiased sample from the overall population. As a result, the class-ratio in the labelled set is the same to the population class-ratio. In mathematical notation it holds $\frac{p(y^+|s^+)}{p(y^-|s^+)} = \frac{p(y^+)}{p(y^-)}$, where the *lhs* is the class-ratio in the labelled set and in *rhs* the population class-ratio. As we observe in Figure 6, choosing the most powerful option between S_P and S_N performs similarly with ignoring completely the unlabelled examples. As it was expected, using the semi-supervised data with pairwise deletion has unpredictable performance and often performs much worse than using only the labelled examples.

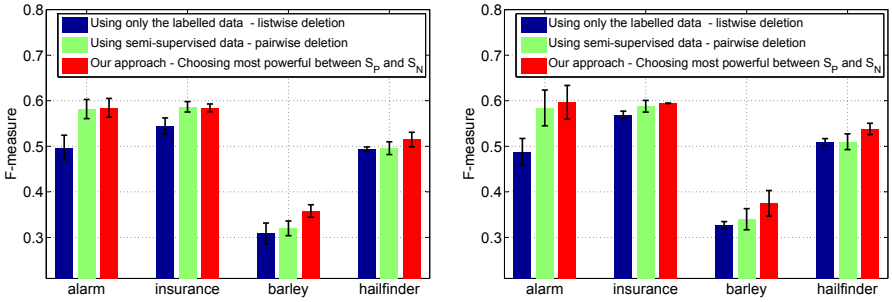


(a) $N = 2000,\ N_{S^+} = 200$ (b) $N = 5000,\ N_{S^+} = 250$

Fig. 6: Traditional semi-supervised scenario. Comparing the performance in terms of F-measure when we have the same class-ratio in the labelled-set as in the overall population. (a) for sample size $N = 2000$ out of which we label only 200 examples and (b) $N = 5000$ out of which we label only 250 examples.

Now, let's assume we have semi-supervised data under the class-prior-change scenario (for more details see Section 3.2). In our simulation we sample the labelled data in order to have a class ratio in the labelled set inverse than the population ratio. In mathematical notation it holds $\frac{p(y^+|s^+)}{p(y^-|s^+)} = \left(\frac{p(y^+)}{p(y^-)}\right)^{-1}$, where the *lhs* is the class-ratio in the labelled set and in *rhs* the inverse of the population

class-ratio. As we observe in Figure 7, choosing the most powerful option between S_P and S_N performs statistically better than ignoring the unlabelled examples. Our approach performs better on average than the pairwise deletion, while the latter one performs comparably to the listwise deletion in many settings.



(a) $N = 2000,\ N_{S^+} = 200$ (b) $N = 5000,\ N_{S^+} = 250$

Fig. 7: Class prior change semi-supervised scenario. Comparing the performance in terms of F-measure when we have inverse class-ratio in the labelled-set than in the overall population. (a) for sample size $N = 2000$ out of which we label only 200 examples and (b) $N = 5000$ out of which we label only 250 examples.

Furthermore, our approach can be applied in scenarios where we have labelled examples only from one class, which cannot be handled with the other two approaches. Also, with our approach, we can control the power of our tests, which is not the case in pairwise deletion procedure. To sum up, in class-prior-change scenarios we can use Corollary 1 and some "soft" prior knowledge over $p(y^+)$ in order to decide which of the following two assumptions is better: to assume all unlabelled examples are negative (i.e. use S_P) or to assume all unlabelled examples are positive (i.e. use S_N).

7 Conclusions and future work

With our work we derive a generalization of conditional tests of independence for partially labelled data and we present a framework on how we can use unlabelled data for discovering Markov blankets around partially labelled target nodes.

In positive-unlabelled data, we proved that assuming all unlabelled examples are negative is sufficient for testing conditional independence but it will increase the number of the variables that are falsely missing from the predicted blanket. Furthermore, with a correction factor, we quantified the amount of power we are losing by this assumption, and we present how we can take this into account for adjusting the sample size in order to perform the same as in fully-supervised scenarios.

Then, we extended our methodology to semi-supervised data, where we can make two valid assumptions over the unlabelled examples: assume them either positive or negative. We explored the consequences of these two assumptions

again in terms of possible errors in Markov blanket discovery procedures, and we suggested a way to use some "soft" prior knowledge to take the optimal decision. Finally, we presented a practical semi-supervised scenario in which the usage of unlabelled examples under our framework proved to be more beneficial compared to other suggested approaches.

A future research direction could be to explore how we can use our methodology for structure learning of Bayesian networks. Since our techniques are informative in terms of power, they can be used in structure learning approaches that have control over the false negative rate to prevent over constraint structures; for example, our framework generalises the work presented by Bacciu et al. [4] for partially-labelled data. Furthermore, our work for structure learning in partially labelled data can be used in combination with recently suggested methods for parameter learning from incomplete data by Mohan et al. [13].

A Generation of network data and experimental protocol

The networks used are standard benchmarks for Markov blanket discovery taken from the Bayesian network repository[3]. For target variables we used nodes that have at least one child, one parent and one spouse in their Markov blanket. Furthermore we chose as positive examples (y^+) those examples with class value 1, while the rest of the examples formed the negative set. We also focused in nodes that had prior probability $p(y^+)$ between 0.15 and 0.50, which is an area of interest for PU data. For the supervised scenarios (i.e. when we used the variable Y) we perform 10 trials of size $N = 2000$ and 5000. For each trial we sample 30 different partially labelled datasets, and the outcome was the most frequently derived Markov blanket. For all experiments we fixed the significance of the tests to be $\alpha = 0.10$. Table 1 presents the summary of the Networks used in the current work.

Table 1: A summary of the networks used in the experimental studies

Network	Number of target nodes	Total number of nodes	Average MB size of target nodes	Average prior prob. $p(y^+)$ of target nodes
alarm	5	37	5.6	0.21
insurance	10	27	6.2	0.32
barley	10	48	5.6	0.31
hailfinder	20	56	4.9	0.31

Acknowledgments. The research leading to these results has received funding from EPSRC Anyscale project EP/L000725/1 and the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 318633. This work was supported by EPSRC grant [EP/I028099/1]. Sechidis gratefully acknowledges the support of the Propondis Foundation.

[3] Downloaded from `http://www.bnlearn.com/bnrepository/`

References

1. Agresti, A.: Categorical Data Analysis. Wiley Series in Probability and Statistics, Wiley-Interscience, 3rd edn. (2013)
2. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blan. induction for causal discovery and feat. selection for classification part I: Algor. and empirical eval. JMLR 11, 171–234 (2010)
3. Allison, P.: Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136 (2001)
4. Bacciu, D., Etchells, T., Lisboa, P., Whittaker, J.: Efficient identification of independence networks using mutual information. Comp. Stats 28(2), 621–646 (2013)
5. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. The Journal of Machine Learning Research (JMLR) 13(1), 27–66 (2012)
6. Cai, R., Zhang, Z., Hao, Z.: BASSUM: A bayesian semi-supervised method for classification feature selection. Pattern Recognition 44(4), 811–820 (2011)
7. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Routledge Academic (1988)
8. Cover, T.M., Thomas, J.A.: Elements of information theory. J. Wiley & Sons (2006)
9. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (2008)
10. Koller, D., Sahami, M.: Toward optimal feature selection. In: International Conference of Machine Learning (ICML). pp. 284–292 (1996)
11. Lawrence, N.D., Jordan, M.I.: Gaussian processes and the null-category noise model. In: Semi-Supervised Learning, chap. 8, pp. 137–150. MIT Press (2006)
12. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: NIPS, pp. 505–511. MIT Press (1999)
13. Mohan, K., Van den Broeck, G., Choi, A., Pearl, J.: Efficient algorithms for bayesian network parameter learning from incomplete data. In: Conference on Uncertainty in Artificial Intelligence (UAI) (2015)
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1988)
15. Pellet, J.P., Elisseeff, A.: Using markov blankets for causal structure learning. The Journal of Machine Learning Research (JMLR) 9, 1295–1342 (2008)
16. Plessis, M.C.d., Sugiyama, M.: Semi-supervised learning of class balance under class-prior change by distribution matching. In: 29th ICML (2012)
17. Pocock, A., Luján, M., Brown, G.: Informative priors for markov blanket discovery. In: 15th AISTATS (2012)
18. Rosset, S., Zhu, J., Zou, H., Hastie, T.J.: A method for inferring label sampling mechanisms in semi-supervised learning. In: NIPS (2004)
19. Sechidis, K., Calvo, B., Brown, G.: Statistical hypothesis testing in positive unlabelled data. In: Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pp. 66–81. Springer Berlin Heidelberg (2014)
20. Smith, A.T., Elkan, C.: Making generative classifiers robust to selection bias. In: 13th ACM SIGKDD Inter. conf. on Knwl. Disc. and Data min. pp. 657–666 (2007)
21. Tsamardinos, I., Aliferis, C.F.: Towards principled feature selection: Relevancy, filters and wrappers. In: AISTATS (2003)
22. Tsamardinos, I., Aliferis, C.F., Statnikov, A.: Time and sample efficient discovery of markov blankets and direct causal relations. In: ACM SIGKDD (2003)
23. Yaramakala, S., Margaritis, D.: Speculative markov blanket discovery for optimal feature selection. In: 5th ICDM. IEEE (2005)