The Chow-Liu Algorithm based on the Minimum Description Length When Discrete and Continuous Variables are Present

Joe Suzuki

Suppose, given a data set, we wish to express its correct dependency among the variables by an undirected tree. One possible way to deal with this problem is to apply the Chow-Liu algorithm [2] to the data set. Then, in our daily experience, we notice that some variables are discrete and others continuous. However, almost all the existing methods assume unrealistic cases: either all the variables are discrete or all of them are Gaussian. In this paper, we remove such restrictions and demonstrate that the proposed algorithm actually works for any data set.

Let $X^{(1)},\cdots,X^{(N)}$ be N (≥ 1) discrete random variables. Let $V:=\{1,\cdots,N\}$ and $E\subseteq\{\{i,j\}|i\neq j,i,j\in V\}$ be vertex and edge sets, and assume that the undirected graph (V,E) expresses a tree, i.e., no loop exists. We consider the associated distribution

$$Q(x^{(1)},\cdots,x^{(N)}|E) \;=\; \prod_{\{i,j\}\in E}\frac{P_{i,j}(x^{(i)},x^{(j)})}{P_i(x^{(i)})P_j(x^{(j)})}\prod_{i\in V}P_i(x^{(i)})\;, \tag{1}$$

where $\{P_i\}_{i\in V}$ and $\{P_{i,j}\}_{i\neq j}$ are obtained by marginalizing the distribution $P_{1,\cdots,N}$ of $X^{(1)},\cdots,X^{(N)}$. Then, suppose we start from $E=\{\}$ and $\mathcal{E}=\{\{i,j\}|i\neq j,i,j\in V\}$; choose $\{i,j\}\in\mathcal{E}$ that maximizes the mutual information $I(i,j)$ of $X^{(i)}$ and $X^{(j)}$, remove the pair $\{i,j\}$ from $\mathcal{E}$, and add it to E as an edge unless any loop is generated when the edge is added; and repeat this until $\mathcal{E}$ is empty to obtain a tree (V,E). Chow and Liu [2] showed that the resulting distribution expressed by (1) minimizes the Kullback divergence

$$D(P_{1,\cdots,N}||Q):=\sum_{x^{(1)}}\cdots\sum_{x^{(N)}}P_{1,\cdots,N}(x^{(1)},\cdots x^{(N)})\log\frac{P_{1,\cdots,N}(x^{(1)},\cdots x^{(N)})}{Q(x^{(1)},\cdots x^{(N)}|E)}$$

among all the trees expressed by (V,E).

In this paper, assuming that the true distribution is expressed by (1), we find (V,E) from n examples $\{X^{(1)}=x_i^{(1)},\cdots,X^{(N)}=x_i^{(N)}\}_{i=1}^n$ so that the estimation converges to the correct (V,E) with probability one as $n\to\infty$. One might want to calculate the maximum likelihood estimators based on the relative frequencies to apply the Chow-Liu algorithm. However, we would not obtain any correct (V,E) in general cases. For examples, suppose $X^{(1)},\cdots,X^{(N)}$ are independent, then such a naive Chow-Liu algorithm seeks a tree whereas no edge should be connected for the random variables. Suzuki [7] considered a similar problem to obtain a way to estimate a correct forest rather than a tree based on the MDL principle [6] but only for discrete variables.

The current paper proposes a method to estimate a forest not just for discrete but also for continuous variables by constructing a Bayesian measure. The measure is constructed by weighting histograms obtained through estimating the quantization of the original distribution rather than estimating one quantized histogram, and its universality is insured. We will find that the MDL based method [7] is contained as a special case of the proposed method. The idea is based on an extended version of the MDL principle. Although consistency is assured for the original MDL, we have not yet given proof of consistency for the extended version, where consistency is defined by the property that the selected model coincides with the true one with probability one as n grows. We will have evidence of the conjecture in our experiments.

On the other hand, very few results have been reported for the problem thus far, and the general approach has been considered to be hard. For example, recently, D. Edwards, et.al [3] considered an extended version of the Chow-Liu algorithm. However, they posed restrictive assumptions: each variable is either discrete or Gaussian; and no Gaussian variables are allowed to be in any path between two discrete variables. Also, several authors revisited a similar approach in machine learning and statistics [4, 5]. However, they are only for discrete variables and more than ten years after [7].

Let $P^n(i) := P^n(\{x_k^{(i)}\}_{k=1}^n) = \prod_{k=1}^n P(X^{(i)} = x_k^{(i)})$ and $P^n(i,j) := P^n(\{x_k^{(i)}, x_k^{(j)}\}_{k=1}^n) = \prod_{k=1}^n P(X^{(i)} = x_k^{(i)}, X^{(j)} = x_k^{(j)})$. Then, we construct measures $R^n(i) := R^n(\{x_k^{(i)}\}_{k=1}^n)$ and $R^n(i,j) := R^n(\{x_k^{(i)}, x_k^{(j)}\}_{k=1}^n)$ such that

$$\sum R^n(i) = 1 \ , \ R^n(i) \geq 0 \ , \ \frac{1}{n} \log \frac{P^n(i)}{R^n(i)} \to 0 \ , \ \sum R^n(i,j) = 1 \ , \ R^n(i,j) \geq 0 \ , \ \frac{1}{n} \log \frac{P^n(i,j)}{R^n(i,j)} \to 0$$

for any $P^n(i)$ and $P^n(i,j)$, respectively. Notice that R^n follows P^n for any P^n as $n \to \infty$ (universal Bayesian measures). We define $R^n(x^n|E)$ and $Q^n(x^n|E)$ by $\prod_{\{i,j\} \in E} \frac{R^n(i,j)}{R^n(i)R^n(j)} \prod_{i \in V} R^n(i)$ and $\prod_{k=1}^n Q(x_k^{(1)}, \cdots, x_k^{(N)}|E) = \prod_{\{i,j\} \in E} \frac{P^n(i,j)}{P^n(i)P^n(j)} \prod_{i \in V} P^n(i)$, respectively. Then, for discrete variables, we obtain Theorems 1 and 2:

Theorem 1 For any probability distribution Q expressed by (1), with probability one as $n \to \infty$,

$$\frac{1}{n} \log \frac{Q^n(x^n|E)}{R^n(x^n|E)} \to 0 \ . \tag{2}$$

Theorem 2 The variant of the Chow-Liu algorithm using

$$J(i,j) := \frac{1}{n} \log \frac{1 - p_{ij}}{p_{ij}} + \frac{1}{n} \log \frac{R^n(i,j)}{R^n(i)R^n(j)}$$

instead of $I(i,j)$ obtains the true forest with probability one as $n \to \infty$ if the true distribution is expressed by (1) and the the prior probability p_{ij} of $X^{(i)}$ and $X^{(j)}$ being independent is given for $i \neq j$.

The idea is to maximize the posterior probability that is proportional to $\prod_{\{i,j\} \in E} \frac{(1 - p_{i,j}) R^n(i,j)}{p_{i,j} R^n(i) R^n(j)}$ for forest (V, E) given n examples x^n.

The main result of this paper is to extend Theorems 1 and 2 to the case that each of $X^{(1)}, \cdots, X^{(N)}$ is either discrete or continuous. The idea is to extend the universal Bayesian measures $R^n(i)$ and $R^n(i,j)$ to the ones not assuming either discrete or continuous. The quantities are obtained by estimating and weighting histograms [8], and the derivation is based on the Radon-Nikodym theorem [1].

References

[1] P. Billingsley.*Probability & Measure* (1995): (3rd ed.). New York : Wiley.

[2] C. K. Chow and C. N. Liu. "Approximating discrete probability distributions with dependence trees". *IEEE Transactions on Information Theory*, IT-14(3):462–467, May (1968).

[3] , David Edwards, Gabriel CG de Abreu, Rodrigo Labouriau, "Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests", *MBC Bioinformatics* Vol. 11, No. 18 (2010).

[4] Liang P, Srebro N, "Methods and experiments with bounded tree-width Markov networks". *Tech rep MIT* 2004.

[5] Panayidou K, "Estimation of Tree Structure for Variable Selection". *PhD thesis University of Oxford* (2010).

[6] J.Rissanen, "Modeling by shortest data description". *Automatica* 14: 465-471 (1978).

[7] J. Suzuki "A Construction of Bayesian Networks from Databases on the MDL principle", *Uncertainty in Artificial Intelligence*, Washington DC, July 1993.

[8] J. Suzuki, "MDL/Bayesian Criteria based on Universal Coding/Measure",Solomonoff 85th Memorial Conference, Victoria, Australia, Nov. 2011 (in press for Springer LNAI).