Appendix to "Structural Extension to Logistic Regression"

Russell Greiner (greiner@cs.ualberta.ca)
Dept of Computing Science, University of Alberta, Edmonton, AB T6G 2H1 Canada

Xiaoyuan Su (xiaoyuan@cs.ualberta.ca)
Electrical & Computer Engineering, University of Miami, Coral Gables, FL 33124, USA

Bin Shen (bshen@cs.ualberta.ca)
Dept of Computing Science, University of Alberta, Edmonton, AB T6G 2H1 Canada

Wei Zhou (w2zhou@math.uwaterloo.ca)
Dept of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Appendix

This report contains material that complements the article (GSSZ05). Appendix A provides the proofs for Theorem 2 and Proposition 3; Appendix B provides empirical evidence supporting our use of "cross-tuning" to determine the appropriate number of iterations; Appendix C then provides additional information about the experiments we ran, including tables showing the actual datasets used (Table II), the empirical accuracy for each dataset, when given complete data (Table III) and incomplete data (Table IV). Appendix D presents empirical results (Table VI) showing how our algorithms performed over 20 UCIrvine datasets (Table V) that were missing information; Appendix E compares our ELR results with those of other algorithms, both taken from other papers (Table VII) and our results based on SVMs (Table VIII). Finally, Appendix F presents a short study that explores how the performance of ELR degrades as the model become successively less accurate; and Appendix G considers how ELR will perform when the model considered is "more complex" than the truth.

A. Proofs

Theorem 2 It is *NP*-hard to find the values for the CPtables of a fixed BN-structure that produce the largest (empirical) conditional likelihood for a given *incomplete* sample.

Proof: We reduce 3SAT to our task, using a construction similar to the one in (Coo90): Given any 3-CNF formula $\varphi \equiv \bigwedge C_i$, where each $C_i \equiv \bigvee \pm X_{ij}$, we construct the network shown in Figure 1, with one node for each variable X_i and one for each clause C_j, with an arc from X_i to C_j whenever C_j involves X_i

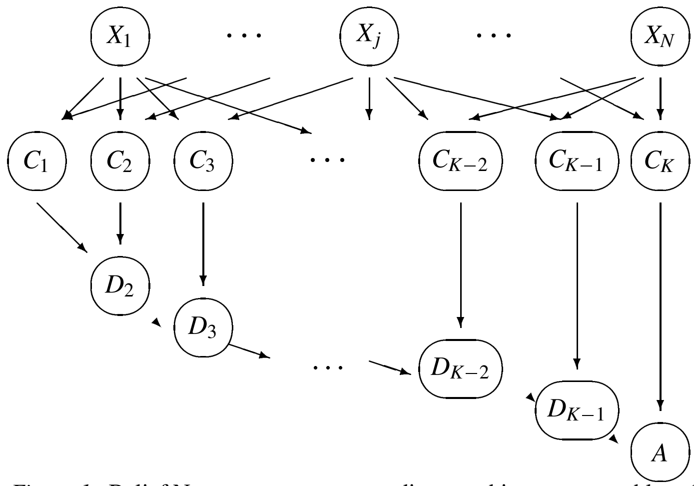


Figure 1. Belief Net structure corresponding to arbitrary SAT problem (Coo90)

— *e.g.*, if $C_1 = x_1 \vee \neg x_2 \vee x_3$ and $C_2 = \neg x_1 \vee \neg x_3 \vee x_4$, then there are links to C_1 from each of X_1, X_2 and X_3, and to C_2 from X_1, X_3 and X_4. In addition, we include $K-1$ other boolean nodes, $\{D_2, \dots, D_{K-1}, A\}$, where D_j is the child of D_{j-1} and C_j, where D_1 is identified with C_1, and A is used for D_K.

Here, we intend each C_i to be true if the assignment to the associated variables X_{i1}, X_{i2}, X_{i3} satisfies C_i; and A corresponds is the conjunction of those C_i variables. We do this using all-but-the-final instances in Table I. (Note only 3 of the X_i variables are specified in each of these instances; the other $n-3$ X_is are not, nor are any C_js nor D_ks.) There is one such instance for each clause, with exactly the assignment (of the 3 relevant variables) that falsifies this clause. Hence, the first line corresponds to $C_1 \equiv x_1 \vee \neg x_2 \vee x_3$. The final instance is just stating that the prior value for A should $P(+a) = 1.0$. The "label" of each instance always corresponds to the single variable A.

We now prove, in particular, that

There is a set of parameters for the structure in Figure 1, producing the $\widehat{\mathrm{LCL}}(\cdot)$-score, over the queries in Table I, of 0

iff

there is a satisfying assignment for the associated φ formula.

$\Leftarrow$**:** Just set the CPtable for each C_i to be the disjunction of the associated X_{i1}, X_{i2}, X_{i3} variables (its parents), with the appropriate $\pm$ parity. *E.g.*, using

Table I. Queries used in proof of Theorem 2

X_1	X_2	X_3	X_4	$\cdots$	X_n	A
0	1	0				0
0		0	1			0
	$\vdots$					$\vdots$
0		1		1		0
						1

$C_1 \equiv x_1 \vee \neg x_2 \vee x_3$, then C_1's CPtable would be

x_1	x_2	x_3	$P(+c_1 \mid x_1, x_2, x_3)$
0	0	0	1.0
0	0	1	1.0
0	1	0	0.0
0	1	1	1.0
1	0	0	1.0
1	0	1	1.0
1	1	0	1.0
1	1	1	1.0

Similarly set the CPtables for the D_j to correspond to the conjunction of its 2 parents $D_j = D_{j-1} \wedge C_j$; *e.g.*,

D_4	C_5	$P(+d_5 \mid D_4, C_5)$
0	0	0.0
0	1	0.0
1	0	0.0
1	1	1.0

.

Finally, set X_i to correspond to the satisfying assignment; *i.e.*, if $X_1 = 1$, then $\boxed{\begin{array}{c} P(+x_1) \\ \hline 1.0 \end{array}}$; and if *i.e.*, if $X_4 = 0$, then $\boxed{\begin{array}{c} P(+x_4) \\ \hline 0.0 \end{array}}$. Note that these CPtable values satify all $k+1$ of the labeled instances.

$\Rightarrow$: Here, we assume there is no satisfying assignment. Towards a contradiction, we can assume that there is a 0-LCL set of CPtable entries. This means, in particular, that $P(+a \mid x_{i1}, x_{i2}, x_{i3}) = 0$, where x_{i1}, x_{i2}, x_{i3} correspond to the assignment that violates the ith constraint. (*E.g.*, for $C_1 \equiv x_1 \vee \neg x_2 \vee x_3$, this would be $X_1 = 0$, $X_2 = 1$, $X_3 = 0$.)

Now consider the final labeled instance, $P(a)$. As there is no satisfying assignment, we know that each assignment $\mathbf{x}$ violates at least one constraint. For notation, let $\gamma^{\mathbf{x}}$ refer to one of these violations (say the one with the smallest index). So if $\mathbf{x} = \langle 0, 1, 0, \ldots \rangle$, then $\gamma^{\langle 0,1,0,\ldots \rangle} = \langle X_1 = 0,\ X_2 = 1,\ X_3 = 0 \rangle$ corresponds to the violation of the first constraint C_1. We also let $\beta^{\mathbf{x}}$ refer to the rest of the assignment.

Now observe

$$\begin{aligned} P(+a) &= \textstyle\sum_{\mathbf{x}} P(+a,\, \mathbf{x}) \\ &= \textstyle\sum_{\mathbf{x}} P(+a \,|\, \gamma^{\mathbf{x}}) \cdot P(\gamma^{\mathbf{x}}) \cdot P(\beta^{\mathbf{x}} \,|\, +a,\, \gamma^{\mathbf{x}}) \\ &= \textstyle\sum_{\mathbf{x}} \quad 0 \quad \cdot P(\gamma^{\mathbf{x}}) \cdot P(\beta^{\mathbf{x}} \,|\, +a,\, \gamma^{\mathbf{x}}) \quad = \quad 0\,, \end{aligned}$$

which shows that the final instance will be mislabeled. This proves that there can be no set of CPtable values that produce 0 LCL-score when there are no satisfying assignments. ∎

Proposition 3 (from (GGS97; Dar00)) For the labeled training case $\langle \mathbf{e}, c\rangle$ and each "softmax" parameter $\beta_{d|\mathbf{f}}$,

$$\frac{\partial\, \widehat{\mathrm{LCL}}^{(\langle \mathbf{e},c\rangle)}(\Theta)}{\partial\, \beta_{d|\mathbf{f}}} = [P_\Theta(d,\mathbf{f}\,|\,\mathbf{e},\, c) - P_\Theta(d,\mathbf{f}\,|\,\mathbf{e})] - \theta_{d|\mathbf{f}}\,[P_\Theta(\mathbf{f}\,|\,c,\, \mathbf{e}) - P_\Theta(\mathbf{f}\,|\,\mathbf{e})]\,.$$

Proof: Below we will use $P(\chi)$ to refer to $P_\Theta(\chi)$, the value the belief net with parameters Θ will assign to the χ event. In general, for any assignment Z,

$$P(Z) \quad = \quad \sum_{\mathbf{f}'} \sum_{d'} P(Z\,|\,D{=}d', \mathbf{F}{=}\mathbf{f}')\; P(D{=}d'\,|\,\mathbf{F}{=}\mathbf{f}')\; P(\mathbf{F}{=}\mathbf{f}')\,. \qquad (1)$$

As we assume the different CPtable rows are estimated independently, and $\mathbf{F}$ is the set of parents of D, this means

$$\frac{\partial\, P(Z)}{\partial\, \beta_{d|\mathbf{f}}} \quad = \quad \sum_{d'} P(Z\,|\,d', \mathbf{f})\; \frac{\partial\, P(d'\,|\,\mathbf{f})}{\partial\, \beta_{d|\mathbf{f}}} P(\mathbf{f})\,.$$

Recalling $\theta_{d|\mathbf{f}} = P(d\,|\,\mathbf{f}) = e^{\beta_{d|\mathbf{f}}} / \sum_{d'} e^{\beta_{d'|\mathbf{f}}}$, observe that $\frac{\partial\, P(d\,|\,\mathbf{f})}{\partial\, \beta_{d|\mathbf{f}}} = \theta_{d|\mathbf{f}}(1 - \theta_{d|\mathbf{f}})$, and when $d \neq d'$, $\frac{\partial\, P(d'\,|\,\mathbf{f})}{\partial\, \beta_{d|\mathbf{f}}} = -\theta_{d|\mathbf{f}}\theta_{d'|\mathbf{f}}$. This means $\frac{\partial\, P(Z)}{\partial\, \beta_{d|\mathbf{f}}} = P(Z, d, \mathbf{f}) - \theta_{d|\mathbf{f}} P(Z, \mathbf{f})$.

Hence, as $\ln P(c\,|\,\mathbf{e}) = \ln P(c, \mathbf{e}) - \ln P(\mathbf{e})$,

$$\begin{aligned} \frac{\partial\, \ln P(c\,|\,\mathbf{e})}{\partial\, \beta_{d|\mathbf{f}}} &= \frac{\partial\, \ln P(c,\mathbf{e})}{\partial\, \beta_{d|\mathbf{f}}} - \frac{\partial\, \ln P(\mathbf{e})}{\partial\, \beta_{d|\mathbf{f}}} \\ &= \frac{1}{P(c,\mathbf{e})} \frac{\partial\, P(c,\mathbf{e})}{\partial\, \beta_{d|\mathbf{f}}} - \frac{1}{P(\mathbf{e})} \frac{\partial\, P(\mathbf{e})}{\partial\, \beta_{d|\mathbf{f}}} \\ &= \frac{1}{P(c,\mathbf{e})}[P(c,\mathbf{e},d,\mathbf{f}) - \theta_{d|\mathbf{f}} P(c,\mathbf{e},\mathbf{f})] - \frac{1}{P(\mathbf{e})}[P(\mathbf{e},d,\mathbf{f}) - \theta_{d|\mathbf{f}} P(\mathbf{e},\mathbf{f})] \\ &= [P(d,\mathbf{f}\,|\,c,\, \mathbf{e}) - P(d,\mathbf{f}\,|\,\mathbf{e})] - \theta_{d|\mathbf{f}}\,[P(\mathbf{f}\,|\,c,\, \mathbf{e}) - P(\mathbf{f}\,|\,\mathbf{e})]\,. \qquad ∎ \end{aligned}$$

B. Empirical Evidence Justifying Cross-Tuning

Gradient based learners have to determine when to stop climbing. A naive implementation would climb for a fixed pre-set number of iterations, or would continue climbing as long as the empirical accuracy is increasing. Our empirical studies (on both ELR and APN) show that these approaches are problematic, as these systems will typically overfit or underfit. To demonstrate this, we present 5-fold cross validation learning curves from TAN+ELR training results on the CLEVE dataset. For each cross validation run, we performed 20 iterations over the training data, and plotted the "Resubstitution Error" and "Generalization Error" after each gradient descent iteration; see Figure 2 below. The "Generalization Error" is the testing error of the resulting system on the hold-out fold after each training iteration. (*I.e.*, we divided CLEVE data into 5 folds: {F1, F2, F3, F4, F5}; in each iteration of the first cross validation run, we used F1+F2+F3+F4 for training, then evaluated the resulting system against the F5 hold-out testing data to produce the "Generalization Error"). Many of the plots show that ELR's gradient ascent starts overfitting significantly only after a few training iterations.

Based on the generalization error plots, we see that ELR should stop after {2, 1, 1, 4, 5} iterations, for these 5 cross validation runs. Of course, ELR will not know these "optimal iteration numbers" as they are based on the hold-out data, which is *not* available at training time.

Fortunately, ELR estimates these numbers from the available training data, using a standard method we call "cross-tuning", described in Section 4 of the manuscript, to try to identify the number of climbs (iterations) that is appropriate for each specific dataset. Cross-tuning first splits the training set into n parts (folds), then successively trains on $n-1$ folds and evaluates on the remaining one. In particular, for each instance, it runs the ELR algorithm on $n-1$ folds for a large number of iterations, and measures the quality of the resulting classifier on the other fold. For each run, it determines which iteration produces the smallest generalization error. Cross-tuning then picks the median value m over these runs. Later, when running on the full dataset (all n folds), it will run for m iterations before stopping.

The paired t-tests of ELR results on the UCI benchmark datasets shows that cross-tuning is essential in ELR learning: NB+ELR(+xt) $\leftarrow_{(p<0.03)}$ NB+ELR(-xt) and TAN+ELR(+xt) $\Leftarrow_{(p<0.05)}$ TAN+ELR(-xt). Here NB+ELR(-xt) is comparable to TAN+ELR(-xt), whose performance was significantly degraded by overfitting. This shows cross-tuning can be effective to prevent overfitting especially when learning parameters of complex BN structures.

The obvious downside of cross-tuning, of course, is computation expense; see timing information in Table IX.

To demonstrate how cross-tuning works to help avoid overfitting, we revisit the experiments on the CLEVE dataset. For the first cross validation run,

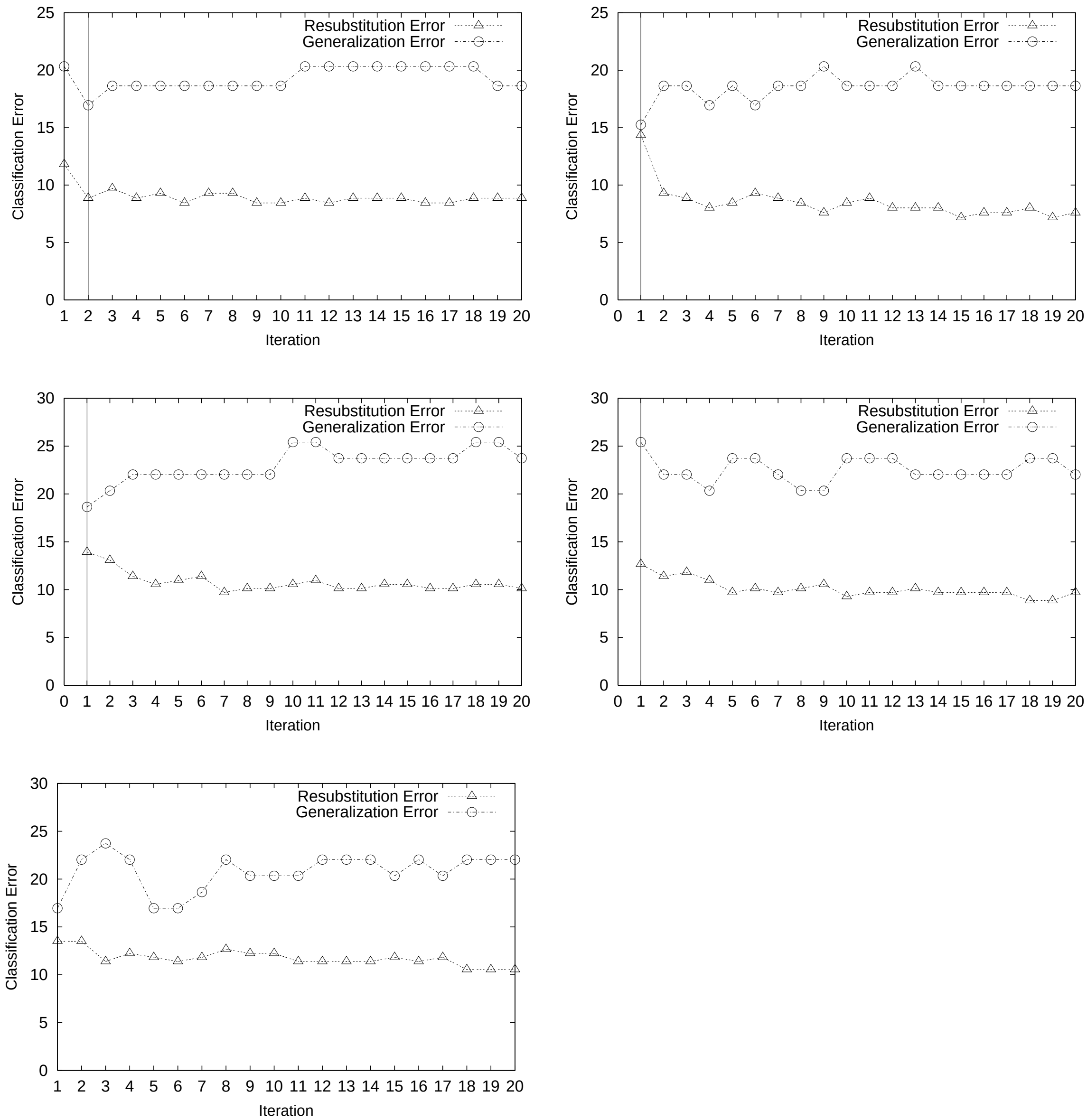


Figure 2. Cross-Tuning Experiments: Resubstitution vs Generalization Error, as function of Number of Iterations (CLEVE dataset), for 5 different CV folds

we split the training data from folds {F1,F2, F3, F4} into another 5 folds for cross-tuning; call them 1CT = {1CT1, 1CT2, ..., 1CT5}. (Note: F1$\cup$ F2 $\cup$ F3 $\cup$ F4 = 1CT1 $\cup$ 1CT2 $\cup \ldots \cup$ 1CT5.) We then ran 5-fold cross-tuning on 1CT, here by using 4 folds of 1CT for training and the remaining 1CT fold for testing, over 20 iterations. Each cross-tuning run determined an iteration number that produced the smallest testing error on the hold-out 1CT fold. After 5-fold cross-tuning runs, we took the median value of the 5 estimates and used it as the iteration number in the training on the full 1CT set.

For this first cross-validation run, this produced an estimate of 2, which we see (from top left graph in Figure 2) is correct. We similarly computed this quantity for the other four cross-validation scenarios, producing {2, 1, 1, 3, 5} respectively for the 5 cross validation runs. Notice cross-tuning identified the correct stopping number in 4 of the 5 cross validation run. The only exception is the fourth one, where it returned 3, not 4.

C. Data for Experiments

We compared the relative effectiveness of ELR with various other classifiers, over the same 25 datasets that (FGG97) used for their comparisons: 23 from UCIrvine repository (BM00), plus "MOFN-3-7-10" and "CORRAL", which were developed by (KJ97) to study feature selection; see Table II, which also specifies how we computed our accuracy values — based on 5-fold cross validation for small data, and holdout method for large data (Koh95). To deal with continuous variables, we implemented supervised entropy discretization (FI93). Table III (resp., Table IV) summarizes the results on complete (resp., incomplete) data.

Table II. Description of data sets used in the experiments (FGG97).

	Dataset	# Attributes	# Classes	# Instances	
				Train	Test
1	AUSTRALIAN	14	2	690	CV-5
2	BREAST	10	2	683	CV-5
3	CHESS	36	2	2130	1066
4	CLEVE	13	2	296	CV-5
5	CORRAL	6	2	128	CV-5
6	CRX	15	2	653	CV-5
7	DIABETES	8	2	768	CV-5
8	FLARE	10	2	1066	CV-5
9	GERMAN	20	2	1000	CV-5
10	GLASS	9	7	214	CV-5
11	GLASS2	9	2	163	CV-5
12	HEART	13	2	270	CV-5
13	HEPATITIS	19	2	80	CV-5
14	IRIS	4	3	150	CV-5
15	LETTER	16	26	15000	5000
16	LYMPHOGRAPHY	18	4	148	CV-5
17	MOFN-3-7-10	10	2	300	1024
18	PIMA	8	2	768	CV-5
19	SATIMAGE	36	6	4435	2000
20	SEGMENT	19	7	1540	770
21	SHUTTLE-SMALL	9	7	3866	1934
22	SOYBEAN-LARGE	35	19	562	CV-5
23	VEHICLE	18	4	846	CV-5
24	VOTE	16	2	435	CV-5
25	WAVEFORM-21	21	3	300	4700

Table III. Empirical accuracy of classifiers learned from *complete* data

	Data set	NB+OFE	NB+ELR	TAN+OFE	TAN+ELR	GBN+OFE	GBN+ELR
1	AUSTRALIAN	86.81±0.84	84.93±1.06	84.93±1.03	84.93±1.03	86.38±0.98	86.81±1.11
2	BREAST	97.21±0.75	96.32±0.66	96.32±0.81	96.32±0.70	96.03±0.50	95.74±0.43
3	CHESS	87.34±1.02	95.40±0.64	92.40±0.81	97.19±0.51	90.06±0.92	90.06±0.92
4	CLEVE	82.03±2.66	81.36±2.46	80.68±1.75	81.36±1.78	84.07±1.48	82.03±1.83
5	CORRAL	86.40±5.31	86.40±3.25	93.60±3.25	100.00±0.00	100.00±0.00	100.00±0.00
6	CRX	86.15±1.29	86.46±1.85	86.15±1.70	86.15±1.70	86.00±1.94	85.69±1.30
7	DIABETES	74.77±1.05	75.16±1.39	74.38±1.35	73.33±1.97	75.42±0.61	76.34±1.30
8	FLARE	80.47±1.03	82.82±1.35	83.00±1.06	83.10±1.29	82.63±1.28	82.63±1.28
9	GERMAN	74.70±0.80	74.60±0.58	73.50±0.84	73.50±0.84	73.70±0.68	73.70±0.68
10	GLASS	47.62±3.61	44.76±4.22	47.62±3.61	44.76±4.22	47.62±3.61	44.76±4.22
11	GLASS2	81.25±2.21	81.88±3.62	80.63±3.34	80.00±3.90	80.63±3.75	78.75±3.34
12	HEART	78.89±4.08	78.52±3.44	78.52±4.29	78.15±3.86	79.63±3.75	78.89±4.17
13	HEPATITIS	83.75±4.24	86.25±5.38	88.75±4.15	85.00±5.08	90.00±4.24	90.00±4.24
14	IRIS	92.67±2.45	94.00±2.87	92.67±2.45	92.00±3.09	92.00±3.09	92.00±3.09
15	LETTER	72.40±0.63	83.02±0.53	83.22±0.53	88.90±0.44	79.78±0.57	81.21±0.55
16	LYMPHOGRAPHY	82.76±1.89	86.21±2.67	86.90±3.34	84.83±5.18	79.31±2.18	78.62±2.29
17	MOFN-3-7-10	86.72±1.06	100.00±0.00	91.60±0.87	100.00±0.00	86.72±1.06	100.00±0.00
18	PIMA	75.03±2.45	75.16±2.48	74.38±2.81	74.38±2.58	75.03±2.25	74.25±2.53
19	SATIMAGE	81.55±0.87	85.40±0.79	88.30±0.72	88.30±0.72	79.25±0.91	79.25±0.91
20	SEGMENT	85.32±1.28	89.48±1.11	89.35±1.11	89.22±1.12	77.53±1.50	77.40±1.51
21	SHUTTLE-SMALL	98.24±0.30	99.12±0.21	99.12±0.21	99.22±0.20	97.31±0.37	97.88±0.33
22	SOYBEAN-LARGE	90.89±1.31	90.54±0.54	93.39±0.67	92.86±1.26	82.50±1.40	85.54±0.99
23	VEHICLE	55.98±0.93	64.14±1.28	65.21±1.32	66.39±1.22	48.52±2.13	51.95±1.32
24	VOTE	90.34±1.44	95.86±0.78	93.79±1.18	95.40±0.63	96.32±0.84	95.86±0.78
25	WAVEFORM-21	75.91±0.62	78.55±0.60	76.30±0.62	76.30±0.62	65.79±0.69	65.79±0.69

Table IV. Empirical accuracy of classifiers learned from *incomplete* data
(25 UCI benchmark datasets with "missing completely at random" at 0.25)

Data set	NB+ELR	NB+APN	NB+EM	TAN+ELR	TAN+APN	TAN+EM	GBN+ELR	GBN+APN	GBN+EM
AUSTRALIAN	$78.41_{\pm 1.01}$	$78.41_{\pm 0.96}$	$78.55_{\pm 1.01}$	$77.25_{\pm 0.59}$	$78.12_{\pm 0.74}$	$77.25_{\pm 0.59}$	$74.06_{\pm 1.06}$	$74.06_{\pm 1.06}$	$74.78_{\pm 0.74}$
BREAST	$95.59_{\pm 1.32}$	$96.03_{\pm 1.20}$	$96.03_{\pm 1.20}$	$96.03_{\pm 1.13}$	$95.88_{\pm 0.95}$	$96.18_{\pm 1.02}$	$94.12_{\pm 1.63}$	$94.85_{\pm 1.36}$	$94.85_{\pm 1.36}$
CHESS	$94.56_{\pm 0.69}$	$89.59_{\pm 0.94}$	$89.68_{\pm 0.93}$	$96.15_{\pm 0.59}$	$93.90_{\pm 0.73}$	$94.09_{\pm 0.72}$	$90.34_{\pm 0.90}$	$90.06_{\pm 0.92}$	$90.06_{\pm 0.92}$
CLEVE	$84.07_{\pm 1.90}$	$82.03_{\pm 2.05}$	$82.03_{\pm 2.05}$	$83.73_{\pm 1.57}$	$83.73_{\pm 1.57}$	$83.73_{\pm 1.57}$	$83.05_{\pm 1.93}$	$81.36_{\pm 2.34}$	$83.39_{\pm 1.89}$
CORRAL	$81.60_{\pm 3.25}$	$83.20_{\pm 3.67}$	$83.20_{\pm 3.67}$	$88.80_{\pm 3.67}$	$90.40_{\pm 1.60}$	$88.80_{\pm 2.65}$	$92.00_{\pm 1.79}$	$88.80_{\pm 2.65}$	$92.00_{\pm 1.79}$
CRX	$87.54_{\pm 1.43}$	$86.00_{\pm 1.67}$	$86.00_{\pm 1.67}$	$85.85_{\pm 1.43}$	$84.62_{\pm 1.29}$	$85.85_{\pm 1.43}$	$86.15_{\pm 1.67}$	$87.23_{\pm 1.10}$	$86.92_{\pm 0.97}$
DIABETES	$75.42_{\pm 1.84}$	$74.64_{\pm 1.83}$	$74.64_{\pm 1.83}$	$74.64_{\pm 2.06}$	$74.90_{\pm 2.19}$	$74.90_{\pm 2.19}$	$73.46_{\pm 1.99}$	$73.20_{\pm 1.99}$	$72.81_{\pm 1.79}$
FLARE	$83.00_{\pm 1.42}$	$82.35_{\pm 1.21}$	$82.44_{\pm 1.24}$	$82.54_{\pm 0.86}$	$82.35_{\pm 1.90}$	$82.54_{\pm 1.52}$	$82.63_{\pm 1.28}$	$82.63_{\pm 1.28}$	$82.63_{\pm 1.28}$
GERMAN	$74.50_{\pm 0.89}$	$74.10_{\pm 1.09}$	$74.00_{\pm 1.05}$	$72.70_{\pm 0.54}$	$74.00_{\pm 0.97}$	$72.90_{\pm 0.40}$	$73.70_{\pm 0.68}$	$73.40_{\pm 0.86}$	$73.70_{\pm 0.68}$
GLASS	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$	$35.71_{\pm 4.33}$
GLASS2	$79.38_{\pm 3.22}$	$77.50_{\pm 3.03}$	$77.50_{\pm 3.03}$	$76.25_{\pm 2.72}$	$76.25_{\pm 3.37}$	$76.25_{\pm 2.72}$	$78.13_{\pm 3.28}$	$77.50_{\pm 3.75}$	$78.13_{\pm 3.28}$
HEART	$75.19_{\pm 5.13}$	$74.81_{\pm 4.63}$	$74.81_{\pm 4.63}$	$72.22_{\pm 3.26}$	$73.33_{\pm 4.00}$	$73.33_{\pm 4.00}$	$73.70_{\pm 3.95}$	$73.33_{\pm 4.37}$	$73.33_{\pm 4.37}$
HEPATITIS	$81.25_{\pm 7.65}$	$86.25_{\pm 5.00}$	$86.25_{\pm 5.00}$	$82.50_{\pm 5.00}$	$87.50_{\pm 3.95}$	$86.25_{\pm 5.00}$	$86.25_{\pm 3.64}$	$86.25_{\pm 3.64}$	$86.25_{\pm 3.64}$
IRIS	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$	$94.67_{\pm 0.82}$
LETTER	$75.28_{\pm 0.61}$	$67.24_{\pm 0.66}$	$67.14_{\pm 0.66}$	$81.86_{\pm 0.54}$	$85.25_{\pm 0.50}$	$84.07_{\pm 0.52}$	$72.80_{\pm 0.63}$	$69.81_{\pm 0.65}$	$68.60_{\pm 0.66}$
LYMPHOGRAPHY	$84.83_{\pm 2.80}$	$84.14_{\pm 1.38}$	$83.45_{\pm 1.29}$	$82.07_{\pm 3.84}$	$78.62_{\pm 2.01}$	$81.38_{\pm 3.87}$	$78.62_{\pm 2.29}$	$78.62_{\pm 2.29}$	$79.31_{\pm 2.18}$
MOFN-3-7-10	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$	$82.03_{\pm 1.20}$
PIMA	$74.90_{\pm 2.85}$	$74.90_{\pm 2.85}$	$74.90_{\pm 2.85}$	$74.25_{\pm 2.45}$	$73.99_{\pm 2.28}$	$73.99_{\pm 2.28}$	$73.99_{\pm 2.06}$	$74.64_{\pm 2.25}$	$74.77_{\pm 2.31}$
SATIMAGE	$84.90_{\pm 0.80}$	$81.85_{\pm 0.86}$	$81.90_{\pm 0.86}$	$87.70_{\pm 0.73}$	$87.80_{\pm 0.73}$	$87.70_{\pm 0.73}$	$73.95_{\pm 0.98}$	$76.35_{\pm 0.95}$	$76.30_{\pm 0.95}$
SEGMENT	$89.74_{\pm 1.09}$	$85.19_{\pm 1.28}$	$85.19_{\pm 1.28}$	$89.35_{\pm 1.11}$	$89.22_{\pm 1.12}$	$89.09_{\pm 1.12}$	$77.40_{\pm 1.51}$	$77.40_{\pm 1.51}$	$77.40_{\pm 1.51}$
SHUTTLE-SMALL	$99.17_{\pm 0.21}$	$99.07_{\pm 0.22}$	$99.07_{\pm 0.22}$	$99.28_{\pm 0.19}$	$99.17_{\pm 0.21}$	$99.17_{\pm 0.21}$	$99.22_{\pm 0.20}$	$98.04_{\pm 0.32}$	$98.04_{\pm 0.32}$
SOYBEAN-LARGE	$85.54_{\pm 1.79}$	$87.68_{\pm 1.77}$	$86.07_{\pm 2.37}$	$84.29_{\pm 1.25}$	$84.64_{\pm 1.34}$	$86.61_{\pm 0.80}$	$50.54_{\pm 1.61}$	$50.18_{\pm 1.75}$	$48.21_{\pm 2.43}$
VEHICLE	$62.72_{\pm 1.69}$	$57.28_{\pm 1.25}$	$57.51_{\pm 1.38}$	$64.85_{\pm 1.29}$	$62.49_{\pm 1.28}$	$62.60_{\pm 1.44}$	$49.94_{\pm 0.91}$	$44.73_{\pm 1.94}$	$44.73_{\pm 1.94}$
VOTE	$94.71_{\pm 0.86}$	$90.80_{\pm 1.54}$	$91.03_{\pm 1.52}$	$94.94_{\pm 0.86}$	$95.40_{\pm 0.51}$	$95.17_{\pm 0.67}$	$95.17_{\pm 0.76}$	$95.63_{\pm 0.92}$	$95.17_{\pm 0.76}$
WAVEFORM-21	$73.34_{\pm 0.64}$	$73.64_{\pm 0.64}$	$73.64_{\pm 0.64}$	$72.26_{\pm 0.65}$	$72.28_{\pm 0.65}$	$72.26_{\pm 0.65}$	$64.38_{\pm 0.70}$	$55.85_{\pm 0.72}$	$55.85_{\pm 0.72}$

Table V. UCIrvine Datasets with Missing Information

dataset information *	# instances	# attributes	# classes		missing ratio	missing total/attris
AGARICUS-LEPIOTA	8124	22	2	CV5	1.39%	2480/1
ALLBP	2800/972	29	3	train/test	5.54%	4556+1508
ALLHYPER	2800/972	29	5	train/test	5.54%	4556+1508
ALLREP	2800/972	29	4	train/test	5.54%	4556+1508
ANNEAL	798	38	6	CV5	64.94%	19692/28
BANDS	540	29	2	CV5	1.93%	302
BREAST-CANCER	699	10	2	CV5	0.23%	16
CLEVE	303	13	2	CV5	0.18%	7
CRX	690	15	2	CV5	0.65%	67/7
DERMATOLOGY	366	34	6	CV5	0.06%	8/1
DIS	2800	29	2	CV5	5.61%	4556
HORSE-COLIC	368	22	2	CV5	23.80%	1927
HYPOTHYROID	3163	25	2	CV5	6.74%	5329
IMPORTS-85	205	25	7	CV5	1.15%	59/7
MONK1-CORRUPT	288/144	6	2	train/test	30.17%	521+261
PRIMARY-TUMOR	339	17	22	CV5	3.90%	225/5
SICK	2800	29	2	CV5	5.61%	4556
SICK-EUTHYROID	3163	25	2	CV5	6.74%	5329
SOYBEAN-LARGE	307/376	25	2	train/test	4.32%	705/33
WATER-TREATMENT	523	38	13	CV5	2.97%	591/31

D. Dealing with Missing Data

We ran a body of experiments over the 20 UCIrvine datasets shown in Table V, and found that, when dealing with NB, ELR was significantly better than either APN or EM: NB+ELR $\Leftarrow_{(p<0.00559)}$ NB+EM and NB+ELR $\Leftarrow_{(p<0.026125)}$ NB+APN. However, there was no statistical significance when considering TAN: TAN+ELR $\leftarrow_{(p<0.083164)}$ TAN+EM and TAN+ELR $\leftarrow_{(p<0.077631)}$ TAN+APN. All of the data appears in Table VI.

Table VI. Results on UCI Datasets with missing information

errors	NB+EM	NB+APN	NB+ELR	TAN+EM	TAN+APN	TAN+ELR
AGARICUS-LEPIOTA	4.41±0.3	4.35±0.32	0±0	0.01±0.01	0±0	0±0
ALLBP	4.22±0	4.22±0	3.09±0	4.12±0	4.12±0	3.5±0
ALLHYPER	2.78±0	2.78±0	1.85±0	2.37±0	1.85±0	1.75±0
ALLREP	3.5±0	3.6±0	3.29±0	2.47±0	2.67±0	2.78±0
ANNEAL	5.79±1.66	4.65±1.84	1.76±0.67	6.54±1.64	5.16±1.86	1.89±0.4
BANDS	30±1.96	29.81±1.79	25.56±1.39	25.37±2.06	24.63±2.24	26.48±2.24
BREAST-CANCER	2.59±0.84	2.59±0.84	3.74±1.14	5.18±0.89	5.76±1.27	5.04±0.85
CLEVE	15.67±3.23	15.67±3.23	16±2.72	18±1.62	17.33±1.63	18±1.62
CRX	14.06±1.11	14.06±1.04	13.33±0.93	15.22±0.51	15.07±0.77	15.22±0.51
DERMATOLOGY	2.19±0.82	2.19±1.11	1.92±0.7	4.66±1.11	3.29±1.11	3.29±1.27
DIS	2.11±0.56	2.11±0.6	1.39±0.26	1.71±0.27	1.57±0.3	1.43±0.22
HORSE-COLIC	19.73±1.66	19.73±1.66	17.81±1.15	18.08±1.01	18.36±0.82	19.73±0.93
HYPOTHYROID	2.25±0.6	1.99±0.63	1.96±0.54	2.31±0.6	2.24±0.6	2.15±0.55
IMPORTS-85	37.56±4.27	37.56±3.33	40±2.51	34.63±1.79	34.15±3.45	33.17±3.05
MONK1-CORRUPT	36.11±0	36.11±0	34.72±0	22.92±0	22.22±0	16.67±0
PRIMARY-TUMOR	51.64±2.69	50.15±3.01	50.45±2.89	51.94±4.51	54.93±3.38	51.94±4.51
SICK	4.71±1.21	4.89±1.36	4.11±0.77	4.46±0.85	4.46±0.88	4.18±0.71
SICK-EUTHYROID	7.03±0.93	6.96±0.89	6.36±0.99	7.25±0.89	7.15±0.91	6.46±1.1
SOYBEAN-LARGE	11.97±0	7.71±0	8.51±0	8.78±0	10.11±0	10.37±0
WATER-TREATMENT	47.31±1.91	47.31±1.91	47.31±1.91	47.31±1.91	47.31±1.91	47.31±1.91
average	**15.2815**	**14.922**	**14.158**	**14.1665**	**14.119**	**13.568**

E. `ELR` vs other Learning Algorithms

Table VII summarizes the experimental results (on complete data) obtained from the following papers.

GSSZ04	Greiner et al. (2004)
GZ02	Greiner and Zhou (2002)
GD04	Domingos et al. (2004)
FGG97	Friedman et al. (1997)

In short, we found that x+`ELR` performed comparably to C4.5 and SNB.

We next compared `ELR` to SVM-light (Joa02). Here, tried a number of parameter setting before settling on the values "`c 0.05 poly 2 (t=1 d=2)`", which we found had the best average performance, over all of the datasets. (Note we just considered the datasets with binary classes.) When using this single-best setting, we found `ELR` was best, for any of the structures:

`NB+ELR` $\Leftarrow_{(p<0.023)}$ SVM-light (best_ave)
`TAN+ELR` $\Leftarrow_{(p<0.036)}$ SVM-light (best_ave)
`GBN+ELR` $\leftarrow_{(p<0.0078)}$ SVM-light (best_ave)

Table VIII presents these results. (Table IX provides timing information.)

Table VII. ELR vs Other Learning Algorithms (from other papers)

	GSSZ04			GZ02		GD04			FGG97	
Data set	GBN+ELR	NB+ELR	TAN+ELR	NB+ELR	TAN+ELR	NB+ELR	TAN+ELR	C4.5	C4.5	SNB
AUSTRALIAN	86.81	84.93	84.93	84.93	85.07	85.12	82.77	84.90	85.65	86.67
BREAST	95.74	96.32	96.32	95.54	96.12	96.61	96.49	93.90	94.73	96.19
CHESS	90.06	95.40	97.19	95.40	97.09	94.00	96.25	99.50	99.53	94.28
CLEVE	82.03	81.36	81.36	82.33	81.33	83.40	78.36	79.40	73.31	78.06
CORRAL	100.00	86.40	100.00	90.40	100.00	87.27	92.29	98.50	97.69	83.57
CRX	85.69	86.46	86.15	84.64	85.07	84.95	83.97	86.10	86.22	85.92
DIABETES	76.34	75.16	73.33	75.69	75.95	75.81	76.16	74.10	76.04	76.04
FLARE	82.63	82.82	83.10	82.72	82.35	81.87	82.20	82.70	82.55	83.40
GERMAN	73.70	74.60	73.50	74.00	73.60	75.44	73.91	72.90	72.20	73.70
GLASS	44.76	44.76	44.76	41.90	41.90	57.80	49.82	59.30	69.62	71.98
GLASS2	78.75	81.88	80.00	77.50	76.25	80.62	77.51	76.10	76.67	79.17
HEART	78.89	78.89	78.15	79.26	80.00	84.50	81.53	78.20	81.11	81.85
HEPATITIS	90.00	86.25	85.00	85.16	85.16	87.06	86.98	82.50	86.25	90.00
IRIS	92.00	94.00	92.00	95.33	95.33	95.15	92.37	96.00	94.00	94.00
LETTER	81.21	83.02	88.90	83.54	88.90	69.32	82.48	87.80	77.70	75.36
LYMPHO-GRAPHY	78.62	86.21	84.83	83.45	79.31	85.30	82.16	78.40	77.03	77.72
MOFN-3-7-10	100.00	100.00	100.00	100.00	100.00	86.33	100.00	84.00	85.55	87.50
PIMA	74.25	75.16	74.38	75.42	75.69	74.95	76.16	74.10	75.13	74.86
SATIMAGE	79.25	85.40	88.30	85.50	88.60	82.70	85.80	82.30	83.15	82.05
SEGMENT	77.40	89.48	89.22	89.74	89.74	92.99	94.29	91.80	93.64	93.25
SHUTTLE-SMALL	97.88	99.12	99.22	99.28	99.38	99.17	99.48	99.40	99.17	99.28
SOYBEAN-LARGE	85.54	90.54	92.86	92.65	92.65	90.80	93.37	91.10	92.00	92.89
VEHICLE	51.95	64.14	66.39	62.72	64.97	65.47	72.73	68.30	69.74	61.36
VOTE	95.86	95.86	95.40	96.09	95.40	96.30	95.13	94.70	95.63	94.71
WAVEFORM-21	65.79	78.55	76.30	78.45	76.74	82.28	74.66	65.10	74.70	76.53

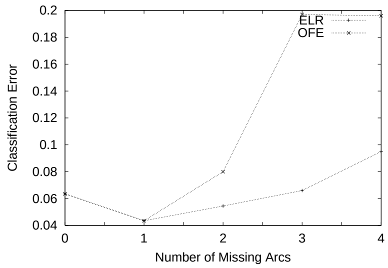
Table VIII. ELR vs SVM

Data set	NB+ELR	TAN+ELR	GBN+ELR	svm-light*
AUSTRALIAN	$84.93_{\pm 1.06}$	$84.93_{\pm 1.03}$	$86.81_{\pm 1.11}$	$70.29_{\pm 9.11}$
BREAST	$96.32_{\pm 0.66}$	$96.32_{\pm 0.70}$	$95.74_{\pm 0.43}$	$93.97_{\pm 1.21}$
CHESS	$95.40_{\pm 0.64}$	$97.19_{\pm 0.51}$	$90.06_{\pm 0.92}$	$97.65_{\pm 0.00}$
CLEVE	$81.36_{\pm 2.46}$	$81.36_{\pm 1.78}$	$82.03_{\pm 1.83}$	$72.54_{\pm 4.39}$
CORRAL	$86.40_{\pm 3.25}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$96.80_{\pm 5.22}$
CRX	$86.46_{\pm 1.85}$	$86.15_{\pm 1.70}$	$85.69_{\pm 1.30}$	$70.15_{\pm 8.34}$
DIABETES	$75.16_{\pm 1.39}$	$73.33_{\pm 1.97}$	$76.34_{\pm 1.30}$	$69.28_{\pm 5.77}$
FLARE	$82.82_{\pm 1.35}$	$83.10_{\pm 1.29}$	$82.63_{\pm 1.28}$	$82.06_{\pm 3.81}$
GERMAN	$74.60_{\pm 0.58}$	$73.50_{\pm 0.84}$	$73.70_{\pm 0.68}$	$66.20_{\pm 1.75}$
GLASS2	$81.88_{\pm 3.62}$	$80.00_{\pm 3.90}$	$78.75_{\pm 3.34}$	$79.37_{\pm 8.45}$
HEART	$78.89_{\pm 4.08}$	$78.15_{\pm 3.86}$	$78.89_{\pm 4.17}$	$76.67_{\pm 2.81}$
HEPATITIS	$86.25_{\pm 5.38}$	$85.00_{\pm 5.08}$	$90.00_{\pm 4.24}$	$86.25_{\pm 5.23}$
MOFN-3-7-10	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
PIMA	$75.16_{\pm 2.48}$	$74.38_{\pm 2.58}$	$74.25_{\pm 2.53}$	$70.59_{\pm 4.03}$
VOTE	$95.86_{\pm 0.78}$	$95.40_{\pm 0.63}$	$95.86_{\pm 0.78}$	$93.10_{\pm 1.15}$
average	85.43	85.92	86.05	81.66

* We tried many settings, and found the setting *[c=0.05, poly 2 (t=1, d=2)* produced the best average for SVM. (As this is based on ALL data, it does give svm-light a slight advantage.)

Table IX. Training Time in seconds, on AMD/MP2600-2048

	ELR without cross-tuning		ELR with 5-fold cross-tuning			
dataset	NB+ELR	TAN+ELR	NB+ELR	TAN+ELR	SVM-light c 0.05 poly 2	
AUSTRALIAN	593.23	1204.16	1345.39	2965.11	2460	CV5
BREAST	478.16	850.11	1067.95	1997.25	60	CV5
CLEVE	214.85	500.27	473.33	1028.93	60	CV5
CORRAL	25.33	118.97	110.67	300.29	60	CV5
CRX	701.17	1296.07	1591.67	3648.29	1020	CV5
DIABETES	518.7	666.61	1107.73	2617.4	60	CV5
FLARE	919.35	1057.65	2090.83	5080.25	180	CV5
GERMAN	951.43	2151.93	3755.08	10346.66	60	CV5
GLASS2	109.62	49.18	229.08	435.69	60	CV5
HEART	245.18	170.88	530.64	1055.6	60	CV5
HEPATITIS	138.79	95.4	308.16	587.67	60	CV5
PIMA	705.64	330.59	1361.73	3345.17	120	CV5
VOTE	776.67	468.1	1878.65	1976.25	60	CV5
CHESS	1932.27	2180.17	4784.38	5574.15	60	Train/Test
MOFN-3-7-10	70.78	85.86	140.74	232.32	60	Train/Test
Average	969.61	5861.2	2968.08	28460.81	296	



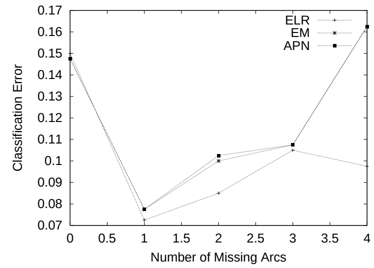


Figure 3. "Correctness of Structure": Comparing ELR to OFE, on increasingly incorrect structures for (a) Complete Data; (b) Incomplete Data

F. "Correctness of Structure" Study

The NaïveBayes-assumption, that the attributes are independent given the classification variable, is typically incorrect. This is known to handicap the NaïveBayes classifier in the standard OFE situation; see the paper and (DP96).

The paper demonstrated above that ELR is more robust than OFE, in that it is not as handicapped by an incorrect structure. We designed the following simple experiment to empirically investigate this claim.

We used synthesized data, to allow us to vary the "incorrectness" of the structure. Here, we consider an underlying distribution P_0 over the $k+1$ binary variables $\{C, E_1, E_2, \ldots, E_k\}$ where (initially) we made NaïveBayes-assumptions and set[1]

$$P(+c) = 0.9 \qquad P(+e_i \,|\, +c) = 0.2 \qquad P(+e_i \,|\, -c) = 0.8 \tag{2}$$

and our queries were all complete; *i.e.*, each instance of the form $\mathbf{E} = \langle \pm e_1, \pm e_2, \ldots, \pm e_k \rangle$.

We then used OFE (resp., ELR) to learn the parameters for the NaïveBayes structure from a data sample, then used the resulting BN to classify additional data. As the structure was correct for this P_0 distribution, both OFE and ELR did quite well, efficiently converging to the optimal classification error.

We then tried to learn the CPtables for this NaïveBayes structure, but for distributions that were *not* consistent with this structure. In particular, we formed the m-th distribution P_m by asserting that $E_1 \equiv E_2 \equiv \ldots \equiv E_m$ (*i.e.*, $P(+e_i \,|\, +e_1) = 1.0$, $P(+e_i \,|\, -e_1) = 0.0$ for each $i = 2..m$) in addition to Equation 2. Hence, P_0 corresponds to the $m = 0$ case. For $m > 0$, however, the m-th distribution cannot be modeled as a NaïveBayes structure, but could be modeled using that structure augmented with $m-1$ links, connecting E_{i-1} to E_i for each $i = 2..m$.

Figure 3(a) shows the results, for $k = 5$, based on 400 instances. As predicted, ELR can produce reasonably accurate CPtables here, even for increasingly wrong structures. However, OFE does progressively worse.

[1] For binary variables, we let "$+c$" represent $c =$ True, and "$-c$" represent $c =$ False.

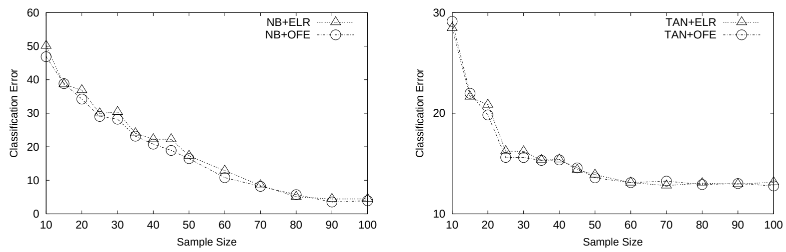


Figure 4. $G > T$ situations, complete data. (a) Model is NB; Truth is $C \equiv E_1$; (b) Model is TAN; Truth is NaïveBayes. (Each point is averaged over 10 runs)

"Correctness of Structure", Incomplete Data: We next degraded this training data by randomly removing the value of each attribute, within each instance, with probability 0.5. Figure 3(b) compares ELR with the standard systems APN and EM; again we see that ELR is more accurate, in each case.

G. Model is More Complex than Truth ($G > T$)

Section 5.1 focused on the common situation where G (the BN-structure being instantiated) is presumedly *simpler* than the "truth" — *e.g.*, we used naïve-bayes when there probably were dependencies between the attributes. This section considers the opposite situation, where we allow the model "more degrees of freedom" than the truth. As this is atypical, we could only consider artificial data.

In our first experiment, we attempt to learn the parameters for a naïve-bayes model, when the truth is $C \equiv E_1$ — *i.e.*, the other attributes $E_2, \ldots, E_k$ are each irrelevant. We focus on $k = 6$ and $k = 7$ attributes, where all variables are binary. When the data is complete, we used first OFE and then ELR to instantiate the parameters of a given NaïveBayes model. Figure 4(a) shows the learning curve as we increase the sample size, over 10 different runs. (Each run used its own training sample.) We see that NB+OFE is consistently slightly better than NB+ELR: averaged over all of the runs, this is significant at $p < 0.002$.

We also weakened the $C \equiv E_1$ condition, to simply require that C be highly correlated with E_1. Using the same set-up show above, when the correlation is 0.96, we found NB+OFE $\Leftarrow_{(p<0.001)}$ NB+ELR. When the correlation is 0.80, the dominance is even more: NB+OFE $\Leftarrow_{(p<0.0001)}$ NB+ELR.

The second experiment "reverses" the situations shown in Appendix F above. Here, the truth corresponds to a naïve-bayes structure (with no dependencies between the evidence E_i variables, conditioned on the class variable), but we attempt to find the parameters for a "P_m-based structure" — *i.e.*, a TAN

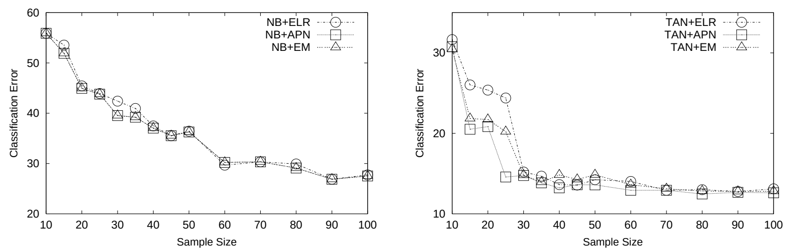


Figure 5. $G > T$ situations, *incomplete* data. (a) Model is NB; Truth is $C \equiv E_1$; (b) Model is TAN; Truth is NaïveBayes. (Each point is averaged over 10 runs.)

structure that links $E_1 \equiv E_2 \equiv \ldots \equiv E_m$. These results appear in Figure 4(b), again this is averaged over 10 runs. (This difference is not significant.)

We next considered the same two situations, but in the *incomplete* data case. In particular, here we blocked a value of any entry with probability 0.2.

The results, shown in Figure 5, show that the generative measures (NB+APN and NB+EM) dominated the discriminative NB+ELR: NB+APN $\Leftarrow_{(p<0.02)}$ NB+ELR and NB+EM $\Leftarrow_{(p<0.015)}$ NB+ELR. (Moreover, NB+EM $\Leftarrow_{(p<0.025)}$ NB+APN.) The generative approach is also superior in the other sitation (Figure 5(b)): TAN+APN $\Leftarrow_{(p<0.025)}$ TAN+ELR, and TAN+EM $\Leftarrow_{(p<0.05)}$ TAN+ELR.

In a nutshell, we observed that discriminative ELR learning typically did worse than the generative learners in this "model is more complex than truth" situation, when dealing with either complete or incomplete data.

Note, of course, that we had to produce a carefully constructed experiment to illustrate this point. As this "$G > T$" situation is very uncommon, we continue to advocate using ELR in general.

References

C. Blake and C. J. Merz. UCI repository of machine learning databases. Technical report, Dept. Info. & Comp. Sci., Univ. Calif. at Irvine, 2000. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, 1990.

A. Darwiche. A differential approach to inference in Bayesian networks. In *UAI'00*, 2000.

P. Domingo and M. Pazzani. Beyond independence: conditions for the optimality of the simple Bayesian classier. In *Proc. 13th International Conference on Machine Learning*, 1996.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning Journal*, 29:131–163, 1997.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteeth International Joint Conference on Artificial Intelligence*, pages 1022–1027, San Francisco, CA, 1993. Morgan Kaufmann.

Russell Greiner, Adam Grove, and Dale Schuurmans. Learning Bayesian nets that perform well. In *Uncertainty in Artificial Intelligence*, 1997.

Russell Greiner, Xiaoyuan Su, Bin Shen, and Wei Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *Machine Learning*, 2005.

Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer, 2002.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 1997.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143, San Francisco, CA, 1995. Morgan Kaufmann.