F-011

特徴選択のための一貫性指標について

On consistency measures for feature selection

中永 健太郎† 兵頭 俊紀† 森川 優† 簗瀬 大悟† 申 吉浩†
Kentaro Nakanaga Toshiki Hyodo Yu Morikawa Daigo Yanase Kilho Shin

1. まえがき

特徴選択は,機械学習を行う前に特徴の数を減らし,少ない説明変数で現象を説明する良いモデルを作ることを目的とする.

特徴選択では,特徴選択指標による評価値に基づいてデータセットの中から特徴集合を抽出する.典型的な特徴選択指標とその適用例を述べる.相互情報量は確率変数間の相関を表す指標で,個別の特徴がどの程度クラスを決定するかを測る物差しとして利用できる.データセット中の全ての特徴に対して,クラスとの相互情報量を計算し,その上位を選択することにより,効率的に特徴選択が行えるように思える.しかしながら,実際には,特徴の間に相互作用がある時には,この手法ではうまくいかない.

特徴間の相互作用について説明する.Table1 は,4 つの特徴 F1,F2,G1,G2 の確率分布を定める.更に,クラスは F1 と F2 の値の排他的論理和により定められるとする.まず,特徴 F1 と F2 は互いに独立であり,C がクラスを表すとして,F1 と C,F2 と C も互いに独立である.すなわち,F1 と F2 は単独では C と相関を持たない.一方,C は F1 と F2 の排他的論理和として決定されるので,特徴集合{F1, F2}は,C の値を一意に決定している.この時,「F1 と F2 は相互に作用して C を決定する」という.ついで,特徴 G1 と G2 を考える.Table 1 から分かるように,G1 および G2 は単独で C に対して相関をもつが,特徴集合{G1,G2}としての相関は 1/2＋ε程度,すなわち,{G1,G2}を選択して分類アルゴリズムを適用すると,確率 1/2 に対して ε 程度しか優位性をもたない.つまり,{F1, F2}と{G1,G2}のどちらかを選ばなければならないのならば,{F1, F2}を選ぶべきなのであるが,F1 や F2 は単独ではクラスに相関がないので,特徴単独の相関によって順位付けを行うと,F1 および F2 は最下位となってしまう.即ち,特徴単独の相関のみを見て特徴を選択する場合は,{F1, F2}を選ばず,{G1, G2}を選ぶ結果となる.

このように,特徴間の相互作用を考慮した特徴選択を行うためには,特徴集合のクラスに対する相関を見る必要がある.その手段として一貫性指標がある.

一貫性指標とは,特徴集合が完全にクラスを決定している時,かつその時に限り,0 を返し（決定性）,評価値が小さいほど,特徴集合とクラスとの相関が高いという性質を持つ指標である.言い換えると,一貫性指標が返した値が小さいほど,特徴集合がクラスラベルを一意に決定できる状態に近い.よって,一貫性指標を評価に用いる特徴選択では,可能な限り評価値の小さい特徴集合を抽出する.

より正確には,一貫性指標は,決定性と単調性の二つの性質を有する指標として定義される(Shin et al, 2012).

(Molina et al, 2002)では,文献中で報告されている特徴選択指標がリストアップされており,(Shin et al, 2012)ではそれらの指標を含めた 19 個の指標を調査し,そのうち,17 個は一貫性指標の定義を満たすことを示した.つまり,文献中で使用されている指標の大多数は一貫性指標であることが分かった.

しかしながら,今まで,これらの一貫性指標の間を精度の観点から比較する研究は知られていない.本論文では初めての試みとして,これらの一貫性指標の精度の観点からの比較を行う.具体的には,各一貫性指標を用いて,複数のデータセット上で特徴選択を行い,選択された特徴により分類を行ったときどれだけの精度を出せるかを,交差検定により評価する.

Table1　確率分布の例

F_1	G_1 ／ G_2 ＼ F_2	0 0	0 1	1 0	1 1	Sum
0	0	$\frac{1}{16}+\frac{\varepsilon}{12}$	$\frac{1}{16}+\frac{\varepsilon}{12}$	$\frac{1}{16}+\frac{\varepsilon}{12}$	$\frac{1}{16}-\frac{\varepsilon}{4}$	$\frac{1}{4}$
0	1	$\frac{1}{16}-\frac{\varepsilon}{12}$	$\frac{1}{16}-\frac{\varepsilon}{12}$	$\frac{1}{16}-\frac{\varepsilon}{12}$	$\frac{1}{16}+\frac{\varepsilon}{4}$	$\frac{1}{4}$
1	0	$\frac{1}{16}-\frac{\varepsilon}{12}$	$\frac{1}{16}-\frac{\varepsilon}{12}$	$\frac{1}{16}-\frac{\varepsilon}{12}$	$\frac{1}{16}+\frac{\varepsilon}{4}$	$\frac{1}{4}$
1	1	$\frac{1}{16}+\frac{\varepsilon}{12}$	$\frac{1}{16}+\frac{\varepsilon}{12}$	$\frac{1}{16}+\frac{\varepsilon}{12}$	$\frac{1}{16}-\frac{\varepsilon}{4}$	$\frac{1}{4}$
C=0		$\frac{1}{8}+\frac{\varepsilon}{6}$	$\frac{1}{8}+\frac{\varepsilon}{6}$	$\frac{1}{8}+\frac{\varepsilon}{6}$	$\frac{1}{8}-\frac{\varepsilon}{2}$	$\frac{1}{2}$
C=1		$\frac{1}{8}-\frac{\varepsilon}{6}$	$\frac{1}{8}-\frac{\varepsilon}{6}$	$\frac{1}{8}-\frac{\varepsilon}{6}$	$\frac{1}{8}+\frac{\varepsilon}{2}$	$\frac{1}{2}$
Sum		$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1

記号の説明

本論文で使う記号を下表に示す.

記号	定義
X	特徴ベクトル
C	クラスを表す確率変数
ε	有限データセット,ε が与えられた時は P は ε から誘導される確率分布を表す
a[X]	a∈ε の X に対応する値ベクトル
a[C]	a∈ε のクラス
$\varepsilon_{X=x}$	$=\{a \in \varepsilon \mid a[X]=x\}$
$\varepsilon_{X=x,C=\xi}$	$=\{a \mid a[X]=x, a[C]=\xi\}$

2. 一貫性指標

まず,一貫的特徴集合を定義する.

データセットの特徴集合が一貫的であるとは,特徴集合の値がクラスを一意に決定すること,即ち,$|\varepsilon_{X=x,C=\xi}|$の値は 0

†兵庫県立大学　応用情報科学研究科

か$|\varepsilon_{X=x}|$のいずれかであることと定義する（Shin et al, 2012）.一貫性指標は,与えられた特徴集合が,一貫性のある状態から,どれだけ乖離しているか,その「距離」を与える指標と考えることができる.特徴集合に含まれる各特徴がクラスに無相関だとしても,集合的に強い相関をもつ場合には,一貫性指標の値は正しく集合としての相関を反映する.一貫性指標は,決定性と単調性を満足する指標として定義される.

（決定性）F が一貫的であることと,μ(F)=0 が成り立つことは同値である.

（単調性）F⊇G ならば,μ(F)≦μ(G)が成り立つ.

直感的に述べれば,この二つの性質は,一貫性指標が小さいほど,一貫的特徴集合に近いということを意味している.したがって,特徴選択を行う時には,閾値をパラメータとして与え,その閾値を超えないという制約のもとで,要素数が最も少ない特徴集合を選ぶことができる.

(Shin et al, 2012)の論文では 17 個の一貫性指標を特定したが,本論文では特定された指標を更に詳細に調査し,重複を取り除くことで,以下の 8 つの一貫性指標を得た.この論文ではこれら 8 つの一貫性指標を比較する.

非一貫性比率（ベイズリスク）(Zhou et al, 2006)

$$\mu_{icr}(X)=\frac{\sum_x(|\varepsilon_{X=x}|-max\,\xi|\varepsilon_{X=x,C=\xi}|)}{|\varepsilon|}$$

ラフセット一貫性指標(Pawlak, 1991)

$$\underline{X}F=\{\varepsilon_{X=x}|\varepsilon_{X=x}\subseteq F\}$$

$$\mu_{rs}(X)=1-\sum_{\xi}\frac{|\underline{X}\varepsilon_{C=\xi}|}{\varepsilon}$$

非一貫的例ペア指標(Arauzo-Azofra et al,2006)

$$\mu_{ie}(X)=\frac{\sum_x\sum_{\xi\neq\eta}|\varepsilon_{X=x,C=\xi}|\cdot|\varepsilon_{X=x,C=\eta}|}{|\varepsilon|^2}$$

条件付きエントロピー(Shin & Xu, 2009)

$$\mu_{ce}(X)=H(C|X)$$

$$=\sum_{x,\xi}-P(X=x,C=\xi)log\frac{P(X=x,C=\xi)}{P(X=x)}$$

Kolmogorov（Molina et al, 2002）

正規化の場合：$\mu_{km}=1-\frac{\sum_x|p(x)-q(x)|}{2}$

非正規化の場合：$\mu_{pkm}=1-\sum_x|p(x)-q(x)|$

Bhattacharyya（Molina et al, 2002）

正規化の場合：$\mu_{bh}=\sum_x\sqrt{p(x)q(x)}$

非正規化の場合:$\mu_{pbh}=\sum_x\sqrt{p(x)q(x)}$

Bhattacharyya と Kolmogorov と表記された指標は,二つの分布 p と q の距離を与える Bhattacharyya ダイバージェンスと Kolmogorov ダイバージェンスから誘導される.p と q の選び方には,正規化（周辺化）と非正規化の二つがあるので,Bhattacharyya と Kolmogorov の両方で合計 4 種類の指標を定める.

正規化の場合の確率は,

$$p(x)=P(X=x|C=0),q(x)=P(X=x|C=1)$$

非正規化の場合の確率は,

$$p(x)=P(X=x,C=0),q(x)=P(X=x,C=1)$$

と与えられる.

一貫性指標の比較としては,(Shin et al. 2012)において,「篩としての目の細かさ」という観点から,理論的な比較がなされている.すなわち,閾値が与えられたとき,一貫性指標は与えられた閾値より小さな評価値を持つ特徴集合を選ぶ「篩」と考えることが出来る.この観点から,下の結果が得られている.

$$\mu_{ie}\ >\ \mu_{icr}\sim\mu_{pkm}\ >\ \mu_{pbh},\mu_{bh},\mu_{kol},\mu_{rs},\mu_{ce}$$

即ち,μ_{ie} が最も目が粗く,μ_{icr} と μ_{pkm} とは同等であり,μ_{pbh}, μ_{bh}, μ_{kol}, μ_{rs}, μ_{ce} は最も目が細かい.μ_{pbh}, μ_{bh}, μ_{kol}, μ_{rs}, μ_{ce} 間の比較は未解決である.

3. 特徴選択アルゴリズム

まえがきで述べたように,特徴間の相互作用を考慮することは重要であり,一貫性指標は,特徴集合に対してクラスへの相関を与える指標となる.

しかし,データセットに含まれる特徴の数を n 個とすると,特徴集合の総数は 2^n 個となり,これらの全てに対して指標値を計算することは計算量的に不可能となる.そこで,適切な探索アルゴリズムを用いて,指標により評価する特徴集合の範囲を狭めることが重要になる.

INTERACT (Zhao et al, 2007)は,一貫性指標に基づく特徴選択アルゴリズムであり,元の特徴集合から特徴の数を減らしていく.閾値を設定し,最後尾の特徴の有無だけが異なる二つの特徴セットの一貫性指標の差が,この閾値を越えるか否かで最後尾の特徴を除去するか否かを判定する.このアルゴリズムの計算量は特徴数に対して線形で,非常に効率的であるが,良好な精度を示す点で重要である.INTERACT の計算量は $O(n)$となる.

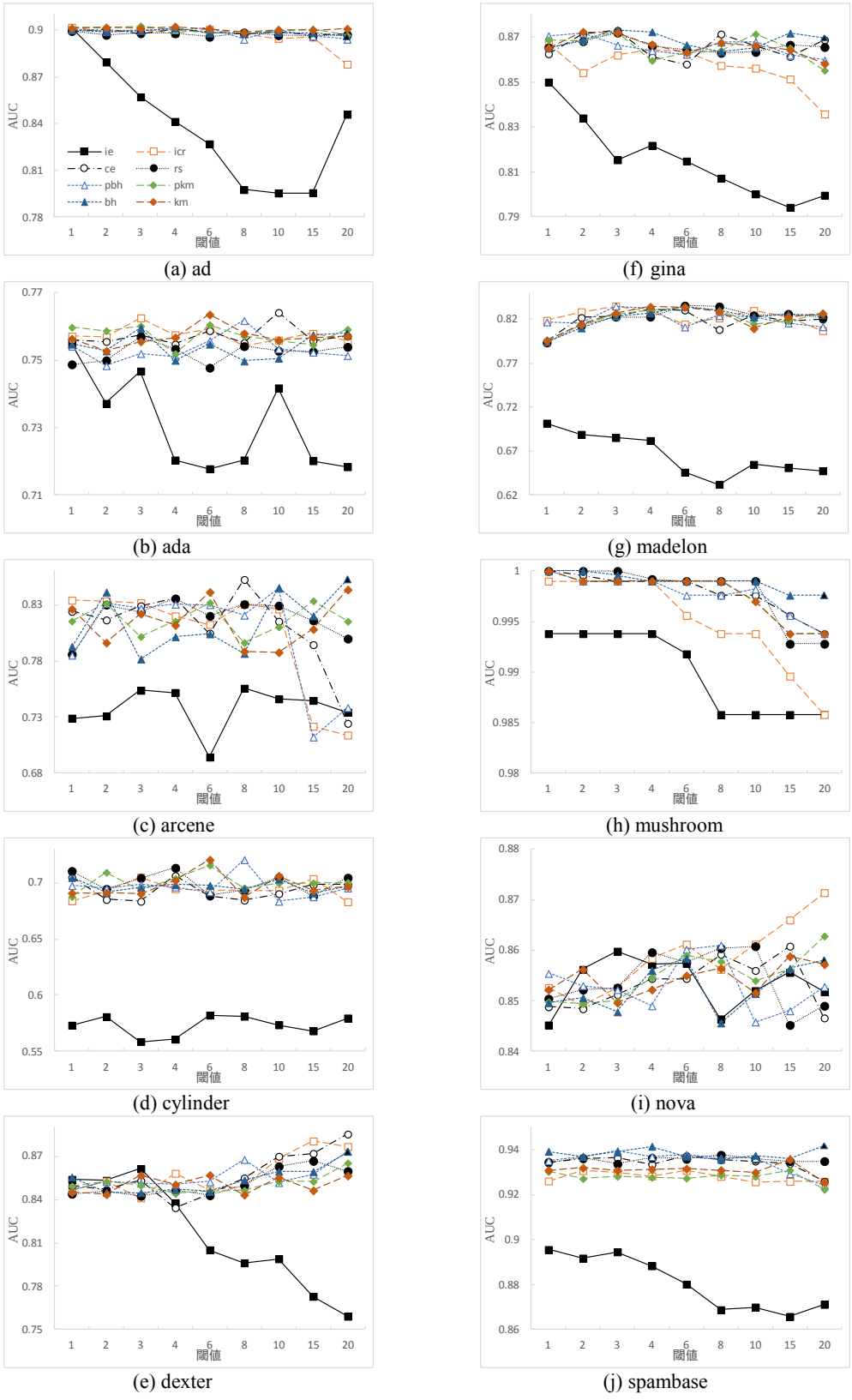
(Shin & Xu, 2009)は INTERACT の理論的な欠陥を指摘し,ある種のデータセットを入れると答えを返せないことを示すと共に,この欠陥を改善したアルゴリズム LCC を提案している.LCC では,より簡潔に,一貫性指標が閾値を下回っていればその特徴集合には含まれていない一つ後ろの特徴を取り除き,上回っていればその特徴を残す.この手順を,特徴集合の削減が閾値で止まるまで繰り返す.最後に残った極小の特徴集合は閾値より小さい一貫性指標を持つ.特徴を簡単に捨てないのが LCC の長所であり,INTERACT より良い精度が得られることが報告されている.

4. 本論文で示すこと

本論文は一貫性指標と呼ばれる一群の特徴選択指標について,その精度に関する比較を実験で行う.(Shin et al, 2012)では,「篩としての目の細かさ」という観点から指標の比較を行っているが,この比較が特徴選択指標の精度としての比較と一致するかについては明らかでない.本論文では両者の比較が一致するか否かという観点も踏まえる.

Table2　**8 つの指標間の順位**

	ie	icr	ce	rs	pbh	pkm	bh	km
平均	0.79 (7.7)	0.847 (4.4)	0.849 (4.5)	0.85 (3.4)	0.848 (5.1)	0.849 (3.75)	0.85 (3.3)	0.849 (3.85)



(a) ad　(f) gina

(b) ada　(g) madelon

(c) arcene　(h) mushroom

(d) cylinder　(i) nova

(e) dexter　(j) spambase

Figure1　8 つの指標の各データセットにおける AUC

5. 実験

5. 1 実験方法

10 個のデータセットに対して 5 分割交差検定を行う.交差検定の学習フェーズで特徴選択と SVM の学習を行い,テストデータの予測から AUC の平均値を計算し比較する.

Table 3　使用したデータセット一覧

データセット	特徴数	サンプル数	出典
ad	1,558	3,279	(Blake and Merz, 1998)
ada	48	4,147	(WCCI,2006)
arcene	10,000	100	(NIPS,2003)
cylinder	40	512	(Blake and Merz, 1998)
dexter	20,000	300	(NIPS,2003)
gina	970	3,153	(WCCI,2006)
madelon	500	2,000	(NIPS,2003)
mushroom	22	8,124	(Blake and Merz, 1998)
nova	19,969	1,754	(WCCI,2006)
spambase	58	4,601	(Blake and Merz, 1998)

1. 任意の指標と閾値の組み合せに対し,LCC を用いて特徴を選択する.
2. 選択される特徴集合は,指標の評価値が閾値を超えない範囲で極小の特徴集合となる.

5. 2 実験結果

Figure1 では,閾値（横軸）を 0.001 から 0.020 まで動かした時の AUC の値を示す.

- データセット ada,cylinder-bands,dexter,madelon-c,spambase では,ie がどの閾値に対しても AUC が悪く,他の指標の間では大きな差はない.
- データセット ad,gina では ie がすべての閾値において AUC が悪く,icr が閾値 20 辺りにおいて他の指標より値が悪い.
- データセット arcene では ie がすべての閾値において AUC が悪く,他の指標では icr,ce,pbh が閾値 15 から 20 辺りで AUC が悪くなっている.
- データセット mushroom では ie と icr が閾値 8 から 20 の間で AUC が悪くなっている.
- データセット nova では icr が閾値 15 から 20 辺りで少し AUC が良くなっている.

5. 3 結果の検証と考察

このように,ie と他の指標との差は顕著であったが,残りの指標間で差があるかどうかははっきり分からなかった.そのため,検定により指標間の差の有意性を統計的に調べることとした.検定に当たっては,(Demsar et al, 2006)に指摘されているように,複数のデータセットを用いた実験値では,ばらつきに正規分布を仮定できないため,ノンパラメトリック検定手法を用いる.更に,3 個以上の指標を比較するため,多重検定を用いる.具体的には,(Demsar et al, 2006)の推奨に従って,Friedman 検定と Hommel 検定を用いることとした.

多重検定を用いるに当たっては,比較対象の数が多すぎると検出力が下がるため,比較対象を限定する.Table 2 に示されるように,rs,pkm,bh,km の 4 指標は AUC の平均も順位の平均もほぼ同じであるので,検定力を維持する必要性を考慮し,この中から代表を一つ選ぶ.ダイバージェンスから誘導される指標としては pbh を比較対象にふくめるため,rs を代表とすることとした.Friedman 検定の p 値は,1.603E-4 であり,危険率 5%で帰無仮説を棄却することができ,5 つの指標間に差があると結論できる.

Table 4　5 つの指標間の順位

	ie	icr	rs	pbh	ce
平均	0.79	0.847	0.85	0.848	0.849
	(4.8)	(2.6)	(2.0)	(3.0)	(2.6)

続いて,順位平均の最もよい rs をコントロールとして Hommel 検定を行った.

Table 5　Hommel Post-Hoc Test (P-Values)

ie	icr	rs	pbh	ce
0.00	0.396	-	0.315	0.396

危険率を 5%と置くと,ie と rs の比較において帰無仮説を棄却できる.すなわち rs は ie に対して有意に優れている.

rs と残りの指標においての比較は,p 値が危険率より大きいので,帰無仮説を棄却することはできないが,p 値は大きくないので,更にデータセットの数を増やして検定すれば差を検出できる可能性がある.

さらに詳しく見ると,icr と ce との比較における信頼度は 40%程度あるのに対し,pbh との比較における信頼度は,30%強である.このことは,より多いデータセットで実験を行えば,icr,ce のグループ と rs,pbh のグループの間に,有意な差が検出できる可能性を示しており,(Shin et al. 2012)で述べられた結果をサポートする可能性がある.(Shin et al. 2012)では rs のグループに属する指標間の順序を定めていないが,pbh との比較の信頼度が比較的小さいことから rs と pbh の間に有意な差がある可能性もある.

6. 今後の展望

Hommel 検定において,rs と ie 以外の残りの指標間の帰無仮説は棄却できなかったが,p 値は大きくないので,データセットの個数を増やして更に調べる必要がある.また,SVM 以外に,J48,NaiveBayes といった分類器を用いて精度を計測する実験も行う予定である.

参考文献

[1] Pawlak, Z., "Rough Sets, Theoretical aspects of reasoning about data." Proc. Kluwer Academic Publishers (1991)
[2] Molina, L., L. Belanche, and A. Nebot, "Feature selection algorithm: A survey and experimental evaluation.", Proc. IEEE International Conference on Data Mining, pp.306-313 (2002).
[3] Arauzo-Azofra.,A, J.M. Benitez, and J. L Castro, "Lerning boolean concepts in the presence of many irrelevant features." Proc. Journal of Intelligent Information Systems 30:3,pp.273-292 (2006)
[4] Demasar, J., "Statistical comparisons of classifiers over multiple data sets." Proc. Journal of Machine Learning Research 7, pp.1-30 (2006).
[5] Zhao, Z., and Liu.H"Searching for Interacting features." Proc. International Joint Conference on Artificial Intelligence, pp.1145-1161 (2007).
[6] Shin, K., D. Fernandes, and S. Miyazaki, "Consistency Measures for Feature Selection: A Formal Definition, Relative Sensitivity Comparison and a Fast Algorithm" Proc. International Joint Conference on Artificial Intelligence, pp.1491-1497 (2012).
[7] Zhou, M.Q., and Y.You (2008) "Hierarchical Bayes Small Area Estimation for the Canadian Community Health Survey"
[8]K.Shin and X.M. Xu. Consistency-based feature selection. In 13th International Conference on Knowledge-Based and Intelligent Information & Engineering system,(2009).
[9] BLAKE, C. S., and C. J. MERZ. UCI repository of machine learning databases. Technical report, University of California, Irvine. (1998)
[10] NIPS. Neural Information Processing Systems Conference 2003: Feature selection challenge. (2003) http://www.nipsfsc.ecs.soton.ac.uk/.
[11] WCCI. IEEE World Congress on Computational Intelligence 2006: Performance prediction challenge. (2006) http://www.modelselect.inf.ethz.ch/.