An Energy-based Model for Feature Selection

H. Jair Escalante, HUGOJAIR@CCC.INAOEP.MX
Manuel Montes, MMONTESG@INAOEP.MX
and **Enrique Sucar** ESUCAR@INAOEP.MX
National Institute of Astrophysics, Optics and Electronics
Department of Computational Sciences
Tonantzintla, Puebla, 72840, México

Editor: Isabelle Guyon et al.

Abstract

In this paper we propose an energy-based model (*EBM*) for selecting subsets of features that are both causally and predictively relevant for classification tasks. The proposed method is tested in the causality challenge, a competition that promotes research on strengthen feature selection by taking into account causal information of features. Under the proposed approach, an energy value is assigned to every configuration of features and the problem is reduced to that of finding the configuration that minimizes an energy function. We propose an energy function that takes into account causal, predictive, and relevance/correlation information of features. Particularly, we introduce potentials that combine the rankings of individual feature selection methods, Markov blanket information and predictive performance estimations. The configuration with lower energy will be that offering the best tradeoff between these sources of information. Experimental results show that despite being simple, the *EBM* approach is able to select highly predictive features. In particular, the combined score of feature relevance and the predictive estimation resulted very useful.

Keywords: Feature selection; Causality Challenge; Energy-based modeling.

1. Introduction

Feature selection consists of choosing subsets of relevant features for building robust learning models. This task is very important because it can help to improve prediction performance of classifiers, to reduce the dimensionality of data (for both efficiency and visualization), and to provide a better understanding of the underlying process that generated the data (Guyon and Elisseeff, 2003). Commonly, the feature selection process is guided by the predictive effectiveness of features and by the correlation among features and between features with the target variable (Guyon and Elisseeff, 2003; Guyon et al., 2007). Despite features selected in this way can help to improve predictive performance of classifiers, it is not clear whether these features are indeed characteristics of the system or the result of experimental artifacts (Guyon et al., 2007). Even when predictive performance may be improved, experimental artifacts are useless for discovering interesting relations of features and for understanding the process that generated the data. For this reason, there is an increasing interest on strengthen feature selection methods by bringing into play causal information and causal discovery techniques (Guyon et al., 2007). Causality may help to improve the selection process by uncovering causal relationships between the variables involved, this could be particularly useful for understanding the underlying process that generated the data. Further, causal methods may provide robustness to violation of the identical and independent distributed (*iid*) assumption commonly made in machine learning. In this line is the causality challenge, a competition that encourages research on the use of causal information for enhancing the feature selection process (Guyon, 2008). The problem approached in this competition is that of selecting features that are both causally and predictively relevant by exploring a setting in which the *iid* assumption does not necessarily holds.

In this paper we report the results we obtained in the causality challenge. For this competition, we propose an energy-based model (*EBM*) that combines information of relevance, predictive effectiveness and causality of features. Under the proposed formulation, an energy value is assigned to every configuration of features and the problem is reduced to that of finding the configuration that minimizes an energy function. Since we are interested in features that are both causally and predictively relevant with respect to a target variable, we propose an energy function that takes into account causal, predictive, and relevance information of features. Relevance of features is measured by a score based on the output of several ranking-based feature selection methods; predictive information is quantified by the cross validation error of features for predicting the target variable; causal information is evaluated by considering the Markov blanket (*MB*) of the target variable. The configuration with lower energy will be that offering the best tradeoff between these sources of information. Experimental results show that despite being simple, the *EBM* approach is able to select highly predictive features. In particular, the combined score of feature relevance and the predictive estimation resulted very useful.

2. Energy-Based Model for Feature Selection

EBMs capture dependencies between variables by a associating a scalar energy to each configuration of them (LeCun et al., 2007). Inference in *EBMs* consists of finding the configuration of the variables that minimize the energy of the model. Problems that require of contextual information are specially well suited for *EBM*. For this reason, *EBMs* have been mainly applied to sequence prediction and other structured domains (LeCun et al., 2007; Bakir et al., 2007). Conditional random fields and several other well known statistical methods have in *EBM* a common theoretical framework. We considered *EBMs* for feature selection because they allow modeling dependencies between features and regarding diverse sources of information without a complicated design or implementation issues.

We consider the problem of feature selection in supervised learning, specifically for classification tasks. We are given a data set D with N samples, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, each $\mathbf{x}_i = x_1, \ldots x_d$ represents the observed values ($x_j \in \mathbb{R}$) of d random variables $X = \{X_1, \ldots, X_d\}$; X denotes the set of features related to the process of study. Each $y_i \in [-1, 1]$ represents the output of the underlying process associated with the observed values of the features $\mathbf{x}_i$; y_i is a realization of Y, the so called target variable. The problem we approach is that of selecting a subset of features $X' = \{X'_1, \ldots X'_m\}$, with $m \leq d$, such that the prediction performance of classifiers using X' for predicting Y is improved with respect of using X, the full set of features.

The user is asked to specify m, the expected size of the subset of relevant features. We define an *EBM* with m random variables $F = \{F_1, \ldots, F_m\}$, each F_i takes values from

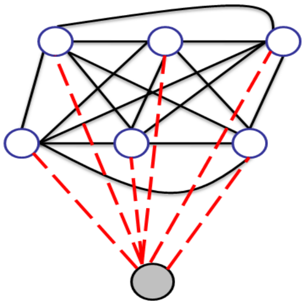

Figure 1: Graphical model of the proposed *EBM* for a value of $m = 6$. Unshaded nodes represent the set of variables of the *EBM* (i. e. $F_{1\ldots m}$) that take values from the set of available features X. The shaded node represents the target variable Y. The dashed lines represent the relationship of F_i with Y (measured by $\gamma(F_i, Y)$). While the solid lines represent the association of the F_i and F_j given Y (measured by $\psi(F_i, F_k, Y)$).

the set of features X, for clarity we abbreviate the assignment $F_i = X_j$ by F_i^j. Each F_i depends on the rest of variables in the model $(F_{j:j\neq i})$ and on the target variable (Y). F is a configuration of *m-unobserved* variables, which values must be inferred by the *EBM* using D, the training data; the target variable Y is *observed* in the training sample of data. In Figure 1 we show the graphical model of the proposed *EBM* for $m = 6$.

We define an energy function over the *EBM* as described in Equation (1).

$$U(F) = -\Big(\sum_{F_j^x \in F}^{m} \gamma(F_j^x, Y) + \sum_{F_j^x \in F}^{m} \sum_{F_k^z \in N_{F_j^x}} \psi(F_j^x, F_k^z, Y) \Big) \tag{1}$$

where $N_{F_i^x}$ is the set of neighboring sites of F_i^x, in our case $N_{F_i^x} = F^x_{j:j\neq i}$. γ and ψ are potentials that can be thought of as restrictions that will favor or punish certain configurations of F. Intuitively, $\gamma(F_i^x, Y)$ is a measure of how *good* is feature $F_i = X_x$ given target variable Y, $\gamma(F_i^x, Y)$ is called the observation potential. $\psi(F_i^x, F_j^z, Y)$, the so called association potential, measures how *good* is the combination of $F_i = X_x$ and $F_j = X_z$ given Y. In the context of the causality challenge, we define γ and ψ such that the minimization of the energy function results in maximizing the predictive power of F while giving preference to those features that are causally related to Y. The configuration F^* that minimizes Equation (1) is selected as the set of relevant features. For minimizing the energy function we considered a simple local-iterative procedure widely used for performing inference in both Markov and discriminative random fields: iterated conditioned modes, *ICM* (Winkler, 2006).

2.1 Observation potential

For measuring the *goodness* of an individual feature F_i^x given the target variable Y we consider its relevance, according several feature selection criteria, together with its predictive power. Both factors are weighted in turn by another one that determines the causal relevance of the feature. Diverse feature selection criteria are considered through the application of several feature selection techniques to the training data set D. For this purpose, we consider the feature selection methods available in the *CLOP*[1] machine learning toolbox

[1] http://clopinet.com/CLOP

ID	Method	Description
1	*s2n*	Signal-to-noise ratio for feature ranking
2	*gs*	Forward feature selection with Gram-Schmidt orth.
3	*relief*	Relief ranking criterion
4	*svcrfe*	Recursive feature elimination filter using svc
5	*auc*	Area under the curve criterion
6	*F-test*	F-test ranking criterion
7	*T-test*	T-test ranking criterion
8	*Pearson*	Pearson correlation coefficient criterion

Table 1: Feature selection methods from the CLOP toolbox that we considered in this work. We considered the default parameters of each method (Saffari and Guyon, 2006).

(Saffari and Guyon, 2006), these methods are described in Table 1. All of these methods rank features according different criteria and they return as relevant the $k-$features at the top of this rank. We consider ranking-based feature selection methods because they are very competitive in practice, they are faster, and simpler than other feature selection approaches (Guyon and Elisseeff, 2003). Commonly, when applied to a same data set D, these methods return different lists of ranked features. Our assumption in this work is that relevant features should appear at the top positions through the different lists, and, conversely, that irrelevant features will appear at the bottom positions. This factor of the observation potential attempts to exploit redundancy and diversity of features through the lists. The individual relevance score for each feature is given by Equation (2)

$$\gamma_1(F_i^x) = \frac{1}{\sum_{j=1}^{N_L} rs(F_i^x, L_j)} \tag{2}$$

where $rs(F_i^x, L_j)$ is the position, in ascending order of relevance, of feature F_i^x in ranked list L_j; N_L is the number of lists, in this work $N_L = 8$ (see Table 1).

We estimate the predictive power of feature F_i^x using the cross validation error of an arbitrary classifier C that uses F_i^x for predicting Y. A score is assigned to each feature according Equation (3)

$$\gamma_2(F_i^x) = \frac{1}{rs(F_i^x, L_{err})} \tag{3}$$

with rs as above and where L_{err} is the ranked list of features in ascending order of their balanced error rate (BER). The BER of classifier C using F_i^x for predicting Y is defined as $BER(C, F_i^x, Y) = \frac{E_+ + E_-}{2}$, where E_+ and E_- are the misclassifications rates for the positive and negative classes respectively.

In Equations (2) and (3) we have accounted for predictive and relevance/correlation information of individual features. We now introduce causal information into the selection process by weighting γ_1 and γ_2 by the individual causal factor described in Equation (4)

$$W_\gamma(F_i^x, Y) = 1 + \left(\mathbf{1}_{F_i^x \in LBM(Y,F)} + \mathbf{1}_{F_i^x \in GBM(Y,X)}\right) \tag{4}$$

where $\mathbf{1}_{F_i^x \in LBM(Y,F)}$ and $\mathbf{1}_{F_i^x \in GBM(Y,X)}$ are indicator functions that take the unit value when the feature F_i^x is in the local (LMB) or global (GMB) MB of Y. The MB of Y is defined as the set of variables that make Y independent of the rest of variables (Guyon et al., 2007). This set is composed of the causes of Y, its direct descendants and common causes of direct descendants of Y. We make a distinction between global and local MB because both sets differ in practice and both can provide information about Y. The GMB

of Y is just the MB obtained by considering the entire set of features X, while the LMB of Y is the MB obtained by considering only the features $F_1^{x_1} \dots F_m^{x_m}$. For computing both the LMB and the GMB we applied the *HITON-MB* algorithm from the Causal Explorer toolbox (Aliferis, 2005). *HITON-MB* is an efficient algorithm for finding the MB of target variables in large data sets with many of features (Aliferis et al., 2003).

The observation potential is a normalized combination of the scores described in Equations (2) and (3) weighted by the causal factor W_γ, as described in Equation (5).

$$\gamma(F_i^x, Y) = \log\big(\frac{\gamma_1(F_i^x, Y) + \gamma_2(F_i^x, Y)}{\sum_i^m \gamma_1(F_i^x, Y) + \gamma_2(F_i^x, Y)} \times W_\gamma(F_i^x, Y)\big) \qquad (5)$$

$\gamma(F_i^x, Y)$ weights the relevance of each individual feature F_i^x given Y by considering its predictive power, its relevance according different feature selection criteria and its causal factor. Configurations that maximize γ will be preferred in the inference process. We will see below, in Section 3, that this potential (in particular the non-causal part) resulted very useful for the selection of highly predictive features.

2.2 Association potential

The association potential ψ is defined similarly as γ, with the difference that now the factors that compose the potential should take as input pairs of features. We consider only the predictive power of pairs of features and a causal factor defined over pairs of features. The predictive power of a pair of features (F_i^x, F_j^z) given Y is determined by

$$BER(C, \{(F_i^x, F_j^z)\}, Y) = \frac{E_+ + E_-}{2} \qquad (6)$$

we take the value instead of the position in a raked list because for some data sets it may be infeasible listing the full set of pairs of features. The causal factor for the association potential is described in Equation (7)

$$W_\psi(F_i^x, F_j^z, Y) = 1 + |\{F_i^x, F_j^z\} \in LBM(Y, F)| + |\{F_i^x, F_j^z\} \in GBM(Y, X)| \qquad (7)$$

The causal factor is now governed by the number of features that are in the LMB and GMB of Y, the maximum value of W_ψ is 5. The association potential is then defined as follows

$$\psi(F_i^x, F_j^z, Y) = \log\big(\frac{1}{BER(C, \{(F_i^x, F_j^z)\}, Y)} \times W_\psi\big) \qquad (8)$$

The association potential is summed over the set of neighboring variables. As with γ, good configurations of features are expected to obtain high values of ψ. We can clearly see that the energy function described in Equation (1) will assign low energy values to configurations of features that are predictively powerful, causally important (according the MB information) and relevant (according diverse feature selection methods). The configuration of features that minimizes Equation (1) is selected as the relevant subset of features of size m.

3. Experimental Results

The data sets considered in the causality challenge are described in Table 2. Four data sets in three different versions are provided to participants, in version '0' the training and testing

Data set	Features	Training	Testing
REGED	999	500	20000
SIDO	4932	12678	10000
CINA	132	16033	10000
MARTI	1024	500	20000

Table 2: Data sets used in the causality challenge. There are three versions of each data set ('0','1','2'). In version '0' training and testing data come from the natural distribution; in versions '1' and '2' the data has been manipulated (Guyon, 2008)

data come from the so called natural distribution (i. e. the *iid* assumption holds), while in versions '2' and '3' training and testing data come from different distributions[2]. For each subset the participant is asked to provide a ranked list of features and nested subsets of features sorted by relevance, together with predictions for the target variable in each nested subset, for details about the challenge we refer the reader to (Guyon, 2008).

We predefined the following subset sizes $M = \{2, 4, 8, 16, 32, 64, 256, ...\}$ for creating the nested subsets of features. The *EBM* is used for selecting features up to size 64, for subsets of higher size we used the rank from Equation (5). For each subset size $m_i \in \{2, \dots, 64\}$ we ran *ICM* for 100 iterations for minimizing the energy function described in Equation (1). The configuration of features that minimizes the energy $F^*_{m_i} = X^*_{mi}$ is selected as the subset of relevant features of size m_i. Then, for each subset of features $X^*_{m_i}$ we build a classifier using $X^*_{m_i}$ in the training set for predicting Y. The trained model is then used on the testing set using the subset of selected features $X^*_{m_i}$. This process is repeated for each subset size, the nested predictions are submitted to the challenge, as well as a ranked list of features *slist*. The *slist* is obtained by merging the nested subsets of features in a similar way as it is done in Equation (2).

A naive Bayes classifier is used for computing the cross validation error in Equations (3) and (6). For predictions in the testing subsets we used a kernel ridge regression classifier in most of our runs. Only for two runs we performed model selection for each subset size. In these runs we applied *PSMS*, i. e. particle swarm model selection (Escalante et al., 2008), a population-based search technique for the selection of classifier, preprocessing method, feature selection and hyperparameter optimization; this method was run for 25 iterations. In Table 3 we show the results of the runs submitted to the challenge, we show the leading measure for ranking participants *Tscore*, which is the area under the *ROC* curve (Guyon, 2008). We show results for the REGED and MARTI data sets, because only in this data sets we could run most of the experiments. For SIDO and CINA we had some difficulties for computing the *MB*.

For the first run *LM* we used the ranking score of Equation (5) with $W_\gamma = 1$ for creating the nested subsets of features. That is, we only considered the merged relevance rank and the predictive power estimation. We can clearly appreciate that *LM* is a strong baseline, since there is not a significant difference with entries that considered further information, not even when we performed model selection. *MB* is a run in which we combine the *LM* ranked list of features with the list of features in the *GMB* of Y, this run is not as good as *LM*. *EBM* is the application of the *EBM* as described in Section 2, we can see that there is not an improvement with respect to *LM*, actually, accuracy decreases for some data sets

[2] In this work we make no distinction in the different versions of the subsets and we treat them equally.

Desc.	Run ID	R0	R1	R2	R-m	M0	M1	M2	M-m
LM	Naive LM all	**0.9997**	**0.9512**	0.8239	**0.9249**	0.9673	0.7867	0.7764	0.8435
MB	MB-LM	0.9995	0.921	0.7596	0.8934	0.9673	0.8154	0.7176	0.8334
EBM	DfRF-MB-LM-10-10	0.9996	0.9389	**0.8355**	0.9246	0.9673	0.8029	0.7607	0.8436
EBM (LM)	DRF-LM-64-kridge-only-LM	0.9995	0.9416	0.8242	0.9217	0.9673	0.8013	0.7513	0.84
EBM - (no GMB)	DRF-LM-MB-not-local	0.9996	0.9389	**0.8355**	0.9246	0.9673	0.7867	0.7777	0.8439
EBM + PSMS	DRF-LM-MB-PSMS Final Run 1	**0.9997**	0.9447	0.7512	0.8985	**0.9675**	0.8052	**0.7858**	0.8528
EBM (LM) + PSMS	DRF-LM-PSMS Final Run 2	0.9996	0.9448	0.7512	0.8985	0.9673	**0.8636**	0.7764	**0.8691**

Table 3: Results of the entries we submitted to the causality challenge for the Reged (**R**) and Marti (**M**) data sets. The first column indicates the settings of our method, columns 3-10 show the *Tscore* (AUC).

by applying this method. The *EBM* with other settings do not resulted in a significant improvement over *LM*.

For the causality challenge, only the last complete run is ranked in the final list. As a result our entry *DRF-LM-PSMS Final Run 2* was the one evaluated. For this run[3] we applied *PSMS* at the end of the feature selection process according our *EBM* do not considering any causal factor (i. e. $W_\gamma = W_\psi = 1$). In Table 4 we show detailed results of this run. Results of this entry are mixed, however we can see that for the data sets from versions '0' the performance of our method is close to that of the top ranked entry. For those data sets in which the *iid* assumption is violated (i. e. versions '2' and '3') performance is not so close. Though for some data sets like *REGED2* and *MARTI2* our method obtains good results. Regarding *FNUM*, the best subset size for making predictions, we can see that the *EBM* worked well for the REGED data set, where small subsets of features performed better. However, for the SIDO, CINA and MARTI the best results are obtained by using a large number of features. A notable exception is MARTI2 for which only 8 features where enough for obtaining performance so close to that obtained by the top ranked entry. We should emphasize that for SIDO and CINA we had some difficulties with the *HITON-MB* algorithm and therefore the *EBM* method could not be applied successfully for these data sets. Results shown in Tables 3 and 4 show that despite being simple, the *EBM* could be useful for selection of predictively relevant features. The causal information we consider resulted useless, and the score obtained by merging the outputs of several ranking-based feature selection methods and the predictive performance estimation resulted very useful.

4. Conclusions

We can conclude two interesting facts from the results obtained in the causality challenge. First, the combination of the output of several feature selection methods with predictive performance estimations resulted very useful for ranking features and making good predictions. Information fusion has proved to be very effective in a number of fields, most notably in machine learning (ensembles, boosting, bagging) and information retrieval (multi-modal retrieval of video and images). Experimental results give evidence that the fusion of feature selection methods has practical advantages. This motivates further research for studying the best way of combining the outputs and for theoretically justifying this idea. Second, the flexibility and generality of the *EBM* framework can be very useful for the problem of feature selection. However, it is not trivial to provide effective ways for measuring the causal

[3] We used this settings for REGED and MARTI, for SIDO and CINA we submitted the results of the *LM* run, see Table 3.

Data set	Fnum	Fscore	Tscore (Ts)	Top Ts	Max Ts	Rank
REGED0	32/999**	0.8778	$0.9996^{+}_{-}0.0010$	1.0000	1.0000	
REGED1	128/999 **	0.7996	$0.9448^{+}_{-}0.0039$	0.9980	0.9980	6
REGED2	64/999 **	0.7638	$0.7512^{+}_{-}0.0060$	0.8600	0.9534	
SIDO0	4096/4932 **	0.8442	$0.9355^{+}_{-}0.0077$	0.9443	0.9467	
SIDO1	4932/4932 **	0.4675	$0.6913^{+}_{-}0.0134$	0.7532	0.78930	8
SIDO2	4932/4932 **	0.4675	$0.6157^{+}_{-}0.0128$	0.6684	0.7674	
CINA0	132/132 **	0.9550	$0.9670^{+}_{-}0.0035$	0.9788	0.9788	
CINA1	132/132 **	0.4982	$0.7873^{+}_{-}0.0049$	0.8977	0.8977	8
CINA2	128/132 **	0.4982	$0.5481^{+}_{-}0.0044$	0.8157	0.8910	
MARTI0	1024/1024 **	0.5446	$0.9673^{+}_{-}0.0036$	0.9996	0.9996	
MARTI1	512/1024 **	0.4711	$0.8636^{+}_{-}0.0054$	0.9470	0.9542	5
MARTI2	8/1024 **	0.7055	$0.7764^{+}_{-}0.0061$	0.7975	0.8273	

Table 4: Final entry submitted to the challenge: *DRF-LM-PSMS Final Run 2*. *Fnum* is the best number of features used to make predictions with *slist*. *Fscore* is an indicator of causal discovery performance, *Tscore* is the area under the *ROC* curve, Top *Ts* is the top valued entry among those ranked, Max *Ts* is the maximum Tscore that can be achieved using ground truth data, finally rank is the position in the raked list of entries.

and predictive effectiveness of configurations of features. This is because while the observation potential resulted very useful, the pairwise potentials we defined were not helpful at all. We should investigate better ways for defining the association potential of the *EBM*, we could, for example, try to learn the energy function from the data, instead of using an ad-hoc defined energy function. Further, we can use different optimization algorithms for making inference in the *EBM.*

Acknowledgments

We would like to thank the organizers of the causality challenge for the support provided. The first author received support from the UNIPEN foundation and from the National Science Foundation under grant ECCS-0725746.

References

C.F. Aliferis. The causal explorer. $http://discover1.mc.vanderbilt.edu/discover/public/causal_explorer/$, 2005.

C.F. Aliferis, I. Tsamardinos, and A. Statnikov. Hiton: A novel markov blanket algorithm for optimal variable selection. In *Proceedings of the FLAIRS conference*, 2003.

G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskas, and S.V.N. Vishwanathan, editors. *Predicting Structured Data.* MIT Press, 2007.

H. J. Escalante, M. Montes, and L. E. Sucar. Particle swarm full model selection. *J. Mach. Learn. Res.*, page accepted, 2008.

I. Guyon. The causality challenge. http://www.causality.inf.ethz.ch/, 2008.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

I. Guyon, C.Aliferis, and A. Elisseeff. Causal feature selection. *Tech. Report Clopinet*, 2007.

Y. LeCun, S. Chopra, R. Hadsell, M. A. Ranzato, and F. J. Huang. *Predicting Structured Data*, chapter Energy-Based Models, pages 191–246. Advances in Neural Information Processing Systems. MIT Press, 2007.

A. Saffari and I. Guyon. Quickstart guide for clop. Technical report, Graz University of Technology and Clopinet, May 2006. http://www.ymer.org/research/files/clop/QuickStartV1.0.pdf.

G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods.* Number 27 in Applications of Mathematics. Springer, 2nd. edition, 2006.