Ensemble Learning Based Sparse High-Order Boltzmann Machine for Unsupervised Feature Interaction Identification

Martin Renqiang Min, NEC Labs
Xia Ning, NEC Labs
Yanjun Qi, NEC Labs
Chao Cheng, Dartmouth College
Anthony Bonner, University of Toronto
Mark Gerstein, Yale University

***Abstract*— Identifying interpretable high-order feature interactions is important specially for biomedical applications. In an unsupervised setting, it is a challenging structure learning task. In this paper, we propose an Ensemble Learning based Sparse High-order Boltzmann Machine (ELSHBM) to identify interpretable high-order feature interactions. By converting the problem of data log-likelihood maximization into the one of data log-pseudo-likelihood maximization, we employ a novel ensemble learning based approach to explore the exponential search space of high-order feature interactions. We estimate the final structure of the high-order Boltzmann Machine using a sparse learning framework, and we use maximum likelihood estimation to learn the parameters given the estimated structure. We apply ELSHBM to a challenging bioinformatics problem of discovering complex Transcription Factor (TF) interactions from ChIP-Seq measurements in the ENCODE project [1]. We can successfully identify many more biologically meaningful interactions that are supported by literature and recent biological studies than by using conventional undirected Boltzmann Machines (BM) or directed Bayesian Networks (BN). More importantly, ELSHBM makes it possible for us to identify high-order interactions that better represent real biological complexes than traditional cliques with only pairwise interactions.**

I. Introduction

Identifying high-order feature interactions is important in machine learning, data mining, and data visualization. Complex feature interactions often convey essential information about the structures of the problem under consideration and reveal characteristic features of the datasets of interest.

Some machine learning and data mining methods implicitly encode complex feature interactions, and they have been applied to problems such as dimensionality reduction [10], [15] and kNN classification [14]. However, these methods are either based on deep neural networks, which are non-trivial to train, or undirected graphical models with hidden units, which are hard to interpret. Only very few effective methods have been developed to explicitly identify high-order feature interactions effectively.

In this paper, we extend the energy function of High-order Boltzmann Machines (HBM) as in [21] to have a combination of different orders of feature interactions up to an allowed maximum order. By converting the problem of data log-likelihood maximization into the one of data log-pseudo-likelihood maximization, we use Random Forests [2] to estimate the high-order interaction neighborhood of each feature variable. We identify the high-order feature interaction terms in the energy function of HBM based on ℓ_1-Regularized Logistic Regression (ℓ_1-LR), and we learn the parameters associated with different energy terms by maximizing the log-likelihood of observed data. We denote this new model for high-order feature interaction identification as Ensemble Learning based Sparse High-order Boltzmann Machine (ELSHBM).

We apply ELSHBM to discover complex Transcription Factor (TF) interactions and associations from ChIP-Seq measurements in the ENCODE project, which is an important problem in bioinformatics. A TF is a protein that controls the flow (or transcription) of genetic information from DNA to mRNA. TFs perform their main function (so-called "regulation") by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes. Thus, they are vital for many important cellular processes. TFs are multi-functional and typically regulate genes in a combinatorial fashion. These combinatorial interactions are critical to understanding TFs, as they provide a means by which the cell can integrate diverse signals, as well as increasing the sensitivity of transcriptional rates to TF concentration. Most relevant genome-wide TF studies focus on pairwise co-association analysis (independent analysis of pairs of TFs) which do not reveal higher-order dependencies, such as how the binding activity of one TF can affect the relationship between two other TFs [8].

II. Related Work

The literature on high-order interaction identification has been very limited, among which most work tackles the problem via structure learning with sparsity constraints. Dahinden *et al* [5] proposed a level-ℓ_1-regularized method to learn high-order interactions. They formulated the problem into a log-linear model with high-order interaction potentials included and group ℓ_1 regularizations. However, due to the iteractive nature of the bottom-up solution, this method is not scalable to large problems. Schmidt *et al* [18] addressed the high-order interaction problem using a hierarchical log-linear model, which has the constraint that if the parameters for a low-order interaction are all zero, then the parameters for a high-order interaction which contains the low-order interaction are also all zeros. Their model is over-parameterized. Schmidt *et al* [19] proposed a Conditional Random Fields (CRF) model to learn the interactions among data labels with block ℓ_1 regularization. Ding [6] proposed a method to learn the high-order interactions

[1] http://www.genome.gov/10005107

among data labels conditioned on features via group lasso with overlaps, that is, they did not learn interactions among features.

III. METHOD

A. Boltzmann Machine and High-order Boltzmann Machine

A fully-observable BM [1] is an undirected graphical model with symmetric weighted connections between "visible units" (features) $\mathbf{v} \in \{0,1\}^p$, where p is the number of visible units. The joint probability distribution of a configuration $\mathbf{v}$ is defined based on its energy function as follows,

$$-E(\mathbf{v}) = \sum_{ij} W_{ij} v_i v_j + \sum_i b_i v_i, p(\mathbf{v}) = \frac{1}{Z} exp(-E(\mathbf{v})), \quad (1)$$

where b_i's are biases, $Z = \sum_{\mathbf{u}} \exp(-E(\mathbf{u}))$ is the partition function, and W_{ij} is the connection weight between unit v_i and v_j. The BM presented above can only be used for modeling explicit pairwise interactions between input features. It was extended in [21] to have only third-order feature interactions in the energy function as follows,

$$-E(\mathbf{v}) = \sum_{ijl} W_{ijk} v_i v_j v_l, \quad (2)$$

and the resulting model was called HBM. However, due to the painfully slow Gibbs Sampling procedure calculating $O(p^3)$ feature interaction terms to get samples from the model distribution, the above dense third-order HBM has never been applied to any interesting practical problems.

B. Ensemble Learning based Sparse High-Order Boltzmann Machine

We further extend the energy function of HBM to have arbitrarily high-order feature interactions up to a maximum order k as follows,

$$-E(\mathbf{v}) = \sum_{i_1} W_{i_1} v_{i_1} + \cdots + \sum_{i_1 i_2 \cdots i_k} W_{i_1 i_2 \cdots i_k} v_{i_1} v_{i_2} \cdots v_{i_k}, \quad (3)$$

and the learning rule is,

$$\Delta W_{i_1 i_2 \cdots i_k} = \epsilon(\langle v_{i_1} v_{i_2} \cdots v_{i_k} \rangle_0 - \langle v_{i_1} v_{i_2} ... v_{i_k} \rangle_\infty), \quad (4)$$

where ϵ is a learning rate, $\langle \cdot \rangle_0$ denotes the expectation with respect to empirical data distribution and $\langle \cdot \rangle_\infty$ denotes the expectation with respect to the model distribution.

Due to the unfeasible computational complexity for learning using Equation 4, we perform structure learning. Following [13], the structure learning of the extended HBM defined in Equation 3 could be performed by solving the following ℓ_1-regularized optimization problem based on negative log-likelihood minimization,

$$\min_{\mathbf{W}} E(\mathbf{v}) + \log Z + \lambda ||\mathbf{W}||_1 \quad (5)$$

Since calculating the above negative log-likelihood and its gradient is intractable, we convert the problem of maximizing the log-likelihood of observed data into that of maximizing the log-pseudo-likelihood of the data as proposed in [11] for similar problems. Specifically, we optimize the following objective function,

$$\min_{\mathbf{W}} \sum_{n=1}^{N} \sum_{i=1}^{p} \log p(v_i^{(n)} | \mathbf{v}_{-i}^{(n)}, \mathbf{W}) + \lambda \cdot ||\mathbf{W}||_1, \quad (6)$$

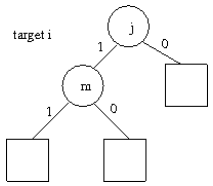

Fig. 1. An example of the decision trees.

where n indexes the data point, N is the size of input dataset, and $\mathbf{v}_{-i}$ is the set of visible units except v_i.

1) Generating High-Order Feature Interactions by Ensemble Learning: Due to the extremely large parameter space associated with high-order feature interactions, we approximate the log-pseudo-likelihood in Equation 6 further by utilizing a strategy proposed by Wainwright *et al* [22]. We first estimate the high-order feature interaction neighborhood of each visible unit. Then each sub-problem of Equation 6 is transformed into a high-order feature selection problem using each feature variable as a prediction target as follows,

$$\min_{\mathbf{W}} \sum_{n=1}^{N} \log p(v_i^{(n)} | \mathbf{v}_{-i}^{(n)}, \mathbf{W}) + \lambda \cdot ||\mathbf{W}||_1, \quad (7)$$

where $i = 1, \ldots, p$. The conditional distribution of v_i given other variables $\mathbf{v}_{-i} = \{v_1, v_2, \ldots, v_{(i-1)}, v_{(i+1)}, \ldots, v_p\}$ takes the following form,

$$p(v_i = 1 | \mathbf{v}_{-i}, \mathbf{W}) = \sigma(\sum_{s=2}^{k-1} \sum_{r_2, \cdots, r_s} W_{i, r_2 \cdots r_k} v_i v_{r_2} \cdots v_{r_s} - b_i), \quad (8)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$, k is the maximum allowed order of feature interactions, v_i can be viewed as the response variable y in a logistic regression where the interactions terms comprising other variables play the role of the Corette's $\mathbf{x}$.

The above optimization problem is generally unfeasible for more than a few variables. Following the approximation approach in [7], we first perform a high-order feature interaction exploration by growing an ensemble of decision trees predicting each feature variable on multiple random subsamples of η percent of the original input dataset. Each path of each resulting decision tree produces a collection of high-order feature interactions that we will consider as candidate feature interaction terms. We can take advantage of existing fast algorithms for producing decision tree ensembles. Thereby, we have a Random Forest predicting each feature variable i based on all the other feature variables. Figure 1 shows an example of a decision tree generated by the feature interaction exploration predicting target feature variable i, in which squares are leaf nodes. In this tree, we have four paths corresponding to two unique feature interactions v_j and $v_j v_m$ for predicting v_i.

The high-order feature interactions corresponding to any path in the tree are given by the product of all the variables in that path. After we generate all the interaction terms from the forest of decision trees for predicting target variable v_i, we use ℓ_1-regularized logistic regression to filter these interactions to identify the final discriminative interactions predictive of the target variable.

2) Filtering Feature Interactions by ℓ_1-regularized Logistic Regression: We filter feature interactions using ℓ_1-regularized logistic regression as in Equation 9 by considering each feature as a data point and the interactions across features as indicated by non-zero and sparse regression relations (i.e., $\mathbf{w}$) among the data points.

$$\min_{\mathbf{w},b} f(\mathbf{w},b) = L(\mathbf{w},b) + \lambda||\mathbf{w}||_1, \tag{9}$$

We use the Projected Scaled Sub-Gradient (PSSG) method [20] to solve the ℓ_1-regularized logistic regression problem, where an L-BFGS update is performed on the non-zero values in $\mathbf{w}$ and a diagonally-scaled pseudo-gradient update is performed on the zero values in $\mathbf{w}$. In the end, orthant projects are applied on the weights so as to introduce sparsity into $\mathbf{w}$.

3) Learning Parameters based on Maximum Likelihood Estimation: After the high-order interaction neighborhood of each feature variable is finalized, we add corresponding high-order feature interaction terms into the energy function of the extended HBM in Equation 3. For example, if $v_j v_m$ and v_j are discriminative interaction terms of v_i, we accordingly add $v_i v_j v_m$ and $v_i v_j$ into $E(\mathbf{v})$ as in Fig. 1. Then we use Maximum-Likelihood Estimation update as in Equation 4 to learn the weights associated with the identified high-order feature interaction terms. We call the resulting HBM Ensemble Learning based Sparse High-order Boltzmann Machine (ELSHBM). The final weight updates for the weights associated with the identified high-order feature interaction terms require drawing samples from the model distribution of ELSHBM. We use damped mean-field updates to get the samples from the model distribution as follows,

$$r^{(t)}(v_i) = \lambda r^{(t-1)}(v_i) + (1-\lambda) p(v_i|\mathbf{v}_{-i}, \mathbf{W}) \tag{10}$$

where $r^{(t)}(v_i)$ is the mean-field approximation to the sampled feature value v_i in iteration t, $t = 1, \cdots, T$, $p(v_i|\mathbf{v}_{-i}, \mathbf{W})$ is the conditional probability given the neighborhood interactions, and $r^{(0)}(v_i)$ is initialized as a data vector. We use $r^{(T)}$ as the final sample from the model distribution.

IV. Experimental Results

A. Datasets

We use a TF-gene regulatory interaction dataset to test the performance of our method. The dataset is downloaded from Gerstein *et al* [9], which defines TF-gene regulatory interactions based on ChIP-seq experiments. In the dataset, the binding scores of genes with TFs were calculated based on the ChIP-seq signals in their promoter regions [4] and based on them the most confident target genes were determined for 116 human TFs. The dataset is represented as a binary matrix with 9,322 rows (genes) and 116 columns (TFs). The (i,j)-th entry of the matrix indicates whether gene i is regulated by TF j.

In the feature interaction exploration step based on ensemble learning, we generate 200 decision trees by randomly sampling 80% of the training data, and set the maximum allowed order of feature interactions to 6. Given the TF-gene interaction matrix, genes are considered data points, and TFs are considered as features. Each data point is represented as a binary interaction profile of the corresponding gene with all TFs, with 1 indicating interactions and 0 indicating non-interactions. The dataset contains 27,901 TF-gene interactions corresponding to a density of 2.58%.

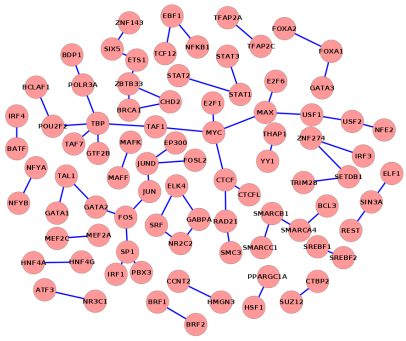

Fig. 2. TF interactions identified from ELSHBM

B. Interaction Identification

We first applied ELSHBM and BN to the TF-gene regulatory dataset for the identification of high order interactions among human TFs. Figure 2 presents the top 65 pairwise interactions identified by ELSHBM.

As shown, ELSHBM identifies many interactions that have been experimentally identified in previous studies. For example, ELSHBM identifies the MYC-MAX interaction as the second most significant. Indeed, MYC and MAX are known to be members of the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors. They form a MAX/MYC heterodimer, in which both of the two subunits can bind DNA and act as transcriptional activators [3]. However, BN is not able to identify the interaction between them. Another example is the interactions of STAT1 with STAT2 and STAT3, which is identified by ELSHBM but not by BN. STAT1, STAT2 and STAT3 are members of the signal transducers and activators of transcription family of transcription factors. In response to IFN-α or IFN-β stimulation, STAT1 and STAT2 form an heterodimer that can bind the ISRE (Interferon Stimulated Response Element) promoter element [12]. The interaction between STAT1 and STAT3 has also been described in previous studies [16]. Some other higher-order interactions identified by ELSHBM which have literature support include the interactions between ESR1, FOXA1 and GATA3; GTF2F1, USF1 and USF2; and E2F1, FO3 and SP1.

We evaluated the identified pairwise interactions by comparing them to an experimental dataset containing 5238 TF-TF physical interactions between 1400 human TF's from Ravsi *et al* [17] and human protein reference database [2]. We use enrichment as the metric for comparison, which is defined as follows,

$$enrichment = \frac{\text{\#identified interactions}}{\text{\#expected interactions}}, \tag{11}$$

where #identified interactions is the number of correctly identified physical interactions, and #expected interactions is the number of expected physical interactions, which is calculated as follows,

$$\text{\#expected interactions} = N \times \frac{\#PPI}{\binom{\#TF}{2}}, \tag{12}$$

where N is the number of TF-TF interactions under consideration, $\#TF$ is the total number of human TFs, and $\#PPI$ is

[2]http://www.hprd.org/

the total number of physical interactions among human TFs. Enrichment is used to test how good the identification is. The higher the enrichment value is, the better the set of identified interactions is.

TABLE I. COMPARISON OF ENRICHMENT

method	50	60	70	80	90	100	110	119
ELSHBM	37.39	37.39	34.72	30.38	27.01	24.30	22.10	25.14
BM	33.65	28.04	26.71	25.71	22.85	24.30	25.49	25.14
BN	3.74	3.12	5.34	9.35	8.31	7.48	6.80	6.28
GTN	22.44	18.70	16.03	16.36	14.54	14.96	15.30	15.58

Each column corresponding to 50, 60, 70, etc, is the enrichment for top 50, top 60, top 70, etc, interactions.

We examined the TF-TF interactions identified by ELSHBM, BN and BM, respectively. Since BN only identifies 119 interactions while ELSHBM and BM can identify more, we only consider the enrichment from the top 119 interactions from each method. Out of the top 119 pairwise interactions identified by ELSHBM, 15 are true physical interactions, which is enriched by 25.13 folds compared to the expected number of interactions in the dataset. In contrast, there are only 4 true physical interactions out of the 119 TF-TF interactions identified by BN, which is enriched by only 6.28 folds compared to the expected number of interactions. There are also 15 true physical interactions out of the top 119 interactions identified by BM. Table I presents the enrichment of method ELSHBM, BM, BN and GTN for top 50 up to top 119 interactions, compared with true physical interactions as ground truth. GTN is the method that is developed by Gerstein *et al* in their Nature paper [9] with average interaction size 3.[3] In GTN, the interaction identification problem is formulated as a supervised classification problem using Rulefit3[4], where the negative instances are generated by randomly permuting the TF-gene interactions for each TF (column-wise shuffling in our setting), and the positive data are the real TF-gene interaction data. Gerstein *et al* suggested an average interaction size 6 in their Nature paper for identifying higher-order interactions involving 5 or more TFs, but these interactions are hard to evaluate for comparisons and none of the above 3-order interactions identified by ELSHBM with literature support was found in the ones output by GTN. With average interaction size 6, GTN identifies 48 pairwise interactions in total with an enrichment of 7.79, which is much lower than the enrichment from ELSHBM. Clearly, ELSHBM has higher enrichment for top interactions than BM and BN, and consistently outperforms the GTN method by a large margin. The model in [18] is the state of the art on small datasets but not scalable to our TF dataset, on which it was run for more than three days without any output.

V. CONCLUSION

In this paper, we present an ensemble learning based sparse undirected graphical model called ELSHBM for unsupervised high-order feature interaction identification. We apply it to discover TF interactions and associations from experimental ChIP-Seq measurements. Experimental results demonstrate that our method successfully identifies much more biologically meaningful high-order TF interactions than conventional undirected BM and the popular directed BN.

[3]The method GTN works best using this parameter setting for identifying pairwise interactions based on our tuning.

[4]http://statweb.stanford.edu/ jhf/R_RuleFit.html

REFERENCES

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Alberto Cascón and Mercedes Robledo. Max and myc: a heritable breakup. *Cancer research*, 72(13):3119–3124, 2012.

[4] Chao Cheng, Renqiang Min, and Mark Gerstein. A probabilistic method for identifying transcription factor target genes from chip-seq binding profiles. *Bioinformatics*, 2011.

[5] C. Dahinden, G. Parmigiani, M.C. Emerick, and P. Bühlmann. Penalized likelihood for sparse contingency tables with an application to full-length cdna libraries. *BMC bioinformatics*, 8(1):476, 2007.

[6] Shilin Ding, Grace Wahba, and Xiaojin (Jerry) Zhu. Learning higher-order graph structure with features by structure penalty. In *NIPS*, pages 253–261, 2011.

[7] J.H. Friedman and B.E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

[8] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science's STKE*, 303(5659):799, 2004.

[9] Mark B. Gerstein and *et al*. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, September 2012.

[10] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[11] H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906, 2009.

[12] Michael G Katze, Yupeng He, and Michael Gale. Viruses and interferon: a fight for supremacy. *Nature Reviews Immunology*, 2(9):675–687, 2002.

[13] S.I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of markov networks using l1 regularization. In *In NIPS*. Citeseer, 2006.

[14] Martin Renqiang Min, David A Stanley, Zineng Yuan, Anthony Bonner, and Zhaolei Zhang. A deep non-linear feature mapping for large-margin knn classification. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 357–366. IEEE, 2009.

[15] Martin Renqiang Min, Laurens van der Maaten, Zineng Yuan, Anthony J. Bonner, and Zhaolei Zhang. Deep supervised t-distributed embedding. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 791–798, 2010.

[16] Sara Pensa, Gabriella Regis, Daniela Boselli, Francesco Novelli, and Valeria Poli. Stat1 and stat3 in tumorigenesis. *JAK-STAT Pathway in Disease*, page 100, 2009.

[17] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

[18] M. Schmidt and K. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[19] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. *CVPR. IEEE Computer Society*, 2008.

[20] Mark Schmidt. *Graphical model structure learning with l1-regularization*. PhD thesis, UNIVERSITY OF BRITISH COLUMBIA, 2010.

[21] Terrence J Sejnowski, Paul K Kienker, and Geoffrey E Hinton. Learning symmetry groups with hidden units: Beyond the perceptron. *Physica D: Nonlinear Phenomena*, 22(1):260–270, 1986.

[22] M.J. Wainwright, P. Ravikumar, and J.D. Lafferty. High-dimensional graphical model selection using l1-regularized logistic regression. *Advances in neural information processing systems*, 19:1465, 2007.