

Correlated Binary Data for Machine Learning

Martí Llobet Turró

Master's student

Universitat Politècnica de Catalunya

Barcelona, Spain

marti.llobet@estudiantat.upc.edu

Margarita Cabrera-Bean

Dept. of Signal Theory and Communications

Universitat Politècnica de Catalunya

Barcelona, Spain

marga.cabrera@upc.edu

Abstract—Data sets containing instances that are assigned values by an ensemble of annotators of unknown accuracy are becoming increasingly common. Binary, potentially correlated data are frequent in a number of disciplines, and thus eligible to be exploited by ensemble meta-learners. A prior key step is testing the meta-learners with synthetic data sets featuring realistic correlation patterns, which is the main scope of this work. To achieve this goal, two challenges are faced: (i) finding out a new correlated pattern to model Bernoulli random variables, and (ii) obtaining a process to generate realistic synthetic data sets. A comparative analysis and performance results are provided for two methods of artificial data generation. The methods are also tested using two state-of-the-art binary ensemble meta-learners that consider inter-classifier dependencies.

Index Terms—Data sets, Unsupervised Ensemble Learning, Bernoulli correlated patterns

I. INTRODUCTION

Research into Unsupervised Ensemble Learning experienced a boost in recent years mainly due to the surge in crowdsourcing. Nonetheless, this Machine Learning discipline holds enormous potential in other fields, including Computational Biology, [1] [2] Computer Vision, [3] Natural Language Processing, [4] and 5G communications systems, [5] among others.

The topic of Unsupervised Ensemble Learning was already discussed by Dawid and Skene in as early as 1979. [6] The authors assumed conditional independence between classifiers, an assumption that has been maintained in similar, subsequent works. Such an assumption does not hold in many real-world scenarios, and it is thus crucial to continue the search for novel statistical methods considering this issue.

Extensive testing with synthetic data is a key step in the characterization of novel algorithms. However, artificial binary data sets where the conditional independence assumption is violated are scarce. Some previous works in the literature generate synthetic correlated binary data imposing several consistent relations on the joint annotators probability function as in [11], or by assuming a regression structure in the correlated binomial distribution parameters by using selected link functions as in [10].

Two recent works tackle the issue of inter-classifier dependencies, each developing an adapted meta-learner. In [7], the

EM algorithm is used to estimate the underlying probability distribution of the instances' ground truth labels; in [8], an algorithm identifying strong inter-classifier correlations is developed and ultimately used to adapt an already existing meta-learner. These two works were developed on the basis of two different approaches as to the modeling of inter-classifier dependencies.

We focus extensively on the approaches presented in [7] and [8]. Our work exploits the one in [7] to obtain an artificial data generation method with realistic correlation patterns. Based on the potentially existing relationships between the different annotators, an intuitive designation of their correlation structure is introduced. Plus, the overall degree of correlation is adjusted through an external input parameter. This provides greater adaptability with regard to possible testing environments for newly developed algorithms. A database containing pathway activation measurements is provided as a real example of highly correlated binary data to be emulated by our method.

The work is organized as follows: in Section II, the problem setup is laid out; in Section III, a novel method of artificial data generation is proposed and compared with an alternative approach; in Section IV, performance results for our artificial data generation method are presented; and in Section V, conclusions are drawn.

II. PROBLEM SETUP

A set of n instances have to be classified. Each instance x must be assigned a ground truth label $y \in \{0, 1\}$, an information that is not generally available. Each labeled instance may be modeled as a realization of the random variable (X, Y) . We have access to the instances' predicted labels provided by a set of m classifiers $\{f_j\}_{j=1}^m$. These results are stored in an n -by- m matrix with entries $z_{ij} = f_j(X_i)$. The classifiers' accuracies are fully characterized by the specificities η_j and sensitivities ψ_j , for $j = 1, \dots, m$, defined as

$$\begin{aligned} \eta_j &= Pr(f_j(X) = 0 \mid Y = 0) \\ \psi_j &= Pr(f_j(X) = 1 \mid Y = 1) \end{aligned} \quad (1)$$

Classifiers can be clustered in *correlation groups*, that is sets of classifiers showing a remarkable degree of correlation. Throughout this work, the number of classifiers belonging to the same correlation group will be written within braces. E.g., $\{5 \ 5 \ 5 \ 1\}$ denotes a 16-classifier ensemble, with 4 different

This work was supported by the project ROUTE56 (Agencia Estatal de Investigación, PID2019-104945GB-I00/AEI/10.13039/501100011033), and in part by the Grant 2017 SGR 578 (AGAUR, Generalitat de Catalunya).

correlation groups; the three first groups contain 5 classifiers each, whilst the fourth one contains a single, independent classifier. Correlation groups are indexed $k = 1, \dots, K$, so that K equals the number of correlation groups, obeying $K \leq m$.

III. ARTIFICIAL DATA GENERATION METHODS

We present a new method to generate artificial correlated, binary data sets, denoted *Method 1*. This method is an upgraded version of the approach applied in [7], which has been entirely revisited to obtain easily tunable, realistic correlation patterns. The approach applied in [8], denoted *Method 2*, is also reviewed for comparative purposes.

A. Method 1

Method 1 estimates the probabilities of obtaining each of the possible combinations of classifier-estimated labels in each correlation group. To do so, the labels provided by the classifiers are modeled as a multivariate Bernoulli distribution, as presented in [9]. This approach requires computing the exponentially growing conditional probabilities of getting a certain label conditioned to previous labels in the same correlation group. In essence, correlation is introduced through hierarchical dependence within each correlation group, as shown in Fig. 1 (Left).

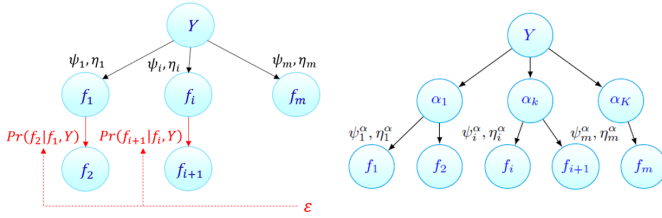


Fig. 1. Underlying inter-classifier dependence models of Method 1 (Left) and Method 2 (Right). Adapted from [8].

The labels provided by the m classifiers are modeled as a 1-by- m vector \mathbf{z} containing realizations of a Bernoulli random variable (r.v.). If there are K correlation groups containing m_k classifiers each s.t. $\sum_{k=1}^K m_k = m$, then $\mathbf{z} = [\mathbf{z}^1 \dots \mathbf{z}^K]$, where \mathbf{z}^k is a 1-by- m_k vector. Each \mathbf{z}^k can take $L_k = 2^{m_k}$ values, which are denoted by \mathbf{z}_l^k for $l = 0, \dots, L_k - 1$, so that \mathbf{z}_l^k may be thought of as the decimal to 1-by- m_k binary conversion of integer l . By extension, \mathbf{z} can take $L = 2^m$ values.

Each \mathbf{z}_l^k is assigned two of the $2L$ probabilities $p_{k,l}^0$ and $p_{k,l}^1$, defined as

$$\begin{aligned} p_{k,l}^0 &= Pr(\mathbf{z}^k = \mathbf{z}_l^k | Y = 0) \\ p_{k,l}^1 &= Pr(\mathbf{z}^k = \mathbf{z}_l^k | Y = 1) \end{aligned} \quad (2)$$

where $l = 0, \dots, L_k - 1$ and $k = 1, \dots, K$. Also, it must be ensured that $\{p_{k,l}^Y\}_{l=0}^{L_k-1}$ make up a probability simplex, i.e. $0 \leq p_{k,l}^Y \leq 1$ and $\sum_{l=0}^{L_k-1} p_{k,l}^Y = 1$.

The set of probabilities are generated in different steps. First, we generate the conditional probabilities associated to the occurrence that a certain instance

is assigned a label by a classifier in the correlation group k , conditioned to classifiers in the same correlation group. These conditional probabilities are denoted $v_{k,l,j}^Y := Pr(f_j^k = a_j | f_1^k = a_1, \dots, f_{j-1}^k = a_{j-1}, Y)$, for $j = 2, \dots, m_k$, where the combination of the classifiers' labels $\{a_1, \dots, a_{m_k}\}$ represents the decimal to binary conversion of the integer l . Secondly, considering the specificity and sensitivity of the first classifier in the correlation group k , η_1^k and ψ_1^k , the $2L_k$ probabilities $\{p_{k,l}^Y\}_{l=0}^{L_k-1}$ are computed as follows,

$$p_{k,l}^0 = \eta_1^k \prod_{j=2}^{m_k} v_{k,l,j}^0 \quad p_{k,l}^1 = \psi_1^k \prod_{j=2}^{m_k} v_{k,l,j}^1. \quad (3)$$

Designating which is the *first classifier* within a correlation group is completely arbitrary. Hierarchical inter-classifier dependence negatively impacts on correlation patterns, and it is thus crucial to randomize this designation, as explained later on in this section.

Equation (3) is equivalent to multiplying one node in each layer of a m_k -layer tree diagram of conditional probabilities. See a graphic example in Fig. 2 with $m_k = 3$ and $Y = 0$. The j -th layer, for $j = 1, \dots, m_k$, is denoted by a 1-by- 2^{j-1} vector $\mathbf{v}_{k,j}^Y$, with entries

$$\mathbf{v}_{k,j}^Y [i] = \begin{cases} Y \psi_1^k + (1 - Y) \eta_1^k & \text{if } j = 1 \\ v_{k,i,j}^Y & \text{if } j > 1 \end{cases}, \quad (4)$$

where $i = 1, \dots, 2^{j-1}$ denotes the binary to decimal conversion of the set of classification results $\{a_1, \dots, a_j\}$. Consequently, $p_{k,l}^Y$ are the joint probabilities that the annotators' labels $\{f_1, \dots, f_{m_k}\}$ are equal to the entries of \mathbf{z}_l^k , for a given k and conditioned to Y . In Fig. 2, this designation for $p_{k,l}^0$ is used. Note that two tree diagrams must be generated eventually, one for each possible ground truth label $Y \in \{0, 1\}$.

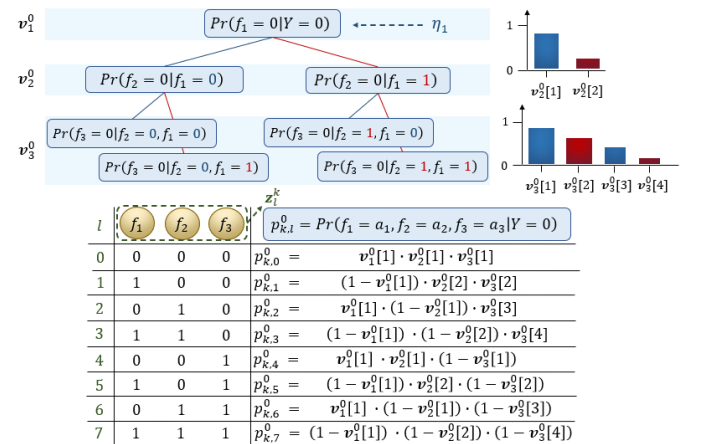


Fig. 2. Tree diagram of conditional probabilities of a three-classifier correlation group, conditioned to $Y = 0$. For the sake of simplicity, the classifier index k is not always included in this figure.

The generation of a synthetic data set following the procedure presented in this work is carried out in two stages.

Firstly, all the probabilities provided in (2) are generated. Secondly, the binary labels are assigned to each instance by the m classifiers.

The first stage begins with the generation of the entire set of vectors $\mathbf{v}_{k,j}^0, \mathbf{v}_{k,j}^1$, for $j = 1, \dots, m_k$, and $k = 1, \dots, K$. The entries of the vectors $\{\mathbf{v}_{k,j}^Y; j = 2, \dots, m_k\}$ are assigned 2^{j-1} realizations of a uniformly distributed r.v. in $[0,1]$. For the case $Y = 0$, the realizations are sorted from highest to lowest in each vector. Contrarily, for $Y = 1$, they are sorted in the opposite order. This logic ensures that the first and last vector entries, $\mathbf{v}_{k,j}^Y[1]$ and $\mathbf{v}_{k,j}^Y[2^{j-1}]$, for $j > 1$, are assigned the conditional probabilities $Pr(f_j^k = 0(1) | f_1^k = 0(1), \dots, f_{j-1}^k = 0(1), Y = 0(1))$ and $Pr(f_j^k = 1(0) | f_1^k = 1(0), \dots, f_{j-1}^k = 1(0), Y = 0(1))$. Consequently, these probabilities always correspond to the highest (lowest) and lowest (highest) values, respectively. See a graphic depiction in Fig. 2 for the case $Y = 0$.

To control the correlation degree, the parameter $\epsilon \in [0,1]$ is introduced. For $Y = 0$, the first and last vector entries, $\mathbf{v}_{k,j}^Y[1]$ and $\mathbf{v}_{k,j}^Y[2^{j-1}]$, for $j > 1$, are recomputed as follows,

$$\begin{aligned} \mathbf{v}_{k,j}^0[1] &= \min\{\epsilon u, \mathbf{v}_{k,j}^0[1]\} \\ \mathbf{v}_{k,j}^0[2^{j-1}] &= \max\{1 - \epsilon u, \mathbf{v}_{k,j}^0[2^{j-1}]\}, \end{aligned} \quad (5)$$

where u is a uniformly distributed r.v. in $[0,1]$. For $Y = 1$, the values assigned to the vector entries are interchanged. Following this logic, the lower the value of ϵ , the higher the degree of correlation.

Once the vectors $\mathbf{v}_{k,j}^Y$ have been obtained, the probabilities $p_{k,l}^0, p_{k,l}^1; l = 1, \dots, 2^{m_k}$ are computed as in (3). Next, the discrete Cumulative Distribution Function (CDF) $F_{k,l}^Y := Pr\{l_k \leq l | Y\}$ for the set of decimal values $l_k = 0, \dots, 2^{m_k} - 1$ in each correlation group k is computed as follows,

$$F_{k,l}^Y = \sum_{l_k=0}^l p_{k,l_k}^Y. \quad (6)$$

In the second stage, given a set of n instances and their corresponding ground truth values $\{y_i\}_{i=1}^n$, each instance x_i is assigned labels by the m classifiers. In each correlation group k , we proceed by generating an n -by-1 vector \mathbf{u} , with entries corresponding to n runs of a uniformly distributed r.v. in $[0,1]$. The entries of the vector \mathbf{u} are assigned the decimal values $l = 0, \dots, L_k - 1$, following the criteria provided by

$$F_{k,l-1}^Y \leq u_i \leq F_{k,l}^Y. \quad (7)$$

Note that (7) will change for each instance depending on its ground truth label y_i .

Upon allocation of the decimal values l to the n instances, performing a decimal to binary transformation produces the corresponding sets of classification results, represented by the n -by- m_k matrix $[\mathbf{z}_{l,1}^k; \dots; \mathbf{z}_{l,n}^k]$. Repeating for each correlation group k , an n -by- m matrix Z is eventually obtained.

One last step is the random permutation of the labels assigned to each instance n by the m_k classifiers in each

correlation group k . This tackles the arbitrary designation of the critical role of *first classifier* for each correlation group.

The software implementation of Method 1 is summarized in Algorithm 1.

Algorithm 1 Artificial Data Generation: Method 1

Input: $\{y_i\}_{i=1}^n, \{\eta_1^k, \psi_1^k\}_{k=1}^K, \epsilon$

Output: Z

```

1: for  $k = 1, \dots, K$  do
2:   Compute  $\mathbf{v}_{k,1}^Y$  as in (4)
3:   for  $j = 2, \dots, m_k$  do
4:     Generate entries  $\mathbf{v}_{k,j}^Y[i] \sim [0, 1]$ 
5:     Depending on  $y_i$ , sort all vector entries, and re-
compute  $\mathbf{v}_{k,j}^Y[1], \mathbf{v}_{k,j}^Y[2^{j-1}]$  as in (5)
6:   end for
7:   Compute  $p_{k,l}^0, p_{k,l}^1$  as in (3)
8:   Compute the CDF as in (6)
9:   Generate a  $n$ -by-1 vector with entries  $u_i \sim [0, 1]$ 
10:  for  $l = 0, \dots, L_k - 1$  do
11:    Replace entries  $u_i$  according to (7)
12:  end for
13:  for  $i = 1, \dots, n$  do
14:    Obtain  $\mathbf{z}_{l,i}^k$  by performing a decimal-to-binary con-
version on  $u_i$ 
15:    Randomly permute the entries of  $\mathbf{z}_{l,i}^k$ 
16:  end for
17: end for

```

B. Method 2

Denoted *Method 2*, we review the procedure to artificially generate binary, correlated data presented in [8] to model inter-classifier dependencies.

In essence, this approach is a relaxation of the inter-classifier conditional independence assumption in the original model developed in [6]. Method 2 introduces inter-classifier dependencies through a series of unobserved binary latent variables, one for each correlation group. Classifiers do not directly depend on the ground truth, but through the latent variables $\{\alpha_k\}_{k=1}^K$. The set of unobserved binary latent variables are conditionally independent when conditioned to Y . Classifiers depending on different latent variables α_k are conditionally independent, whereas classifiers depending on the same α_k show correlated results. Classifiers depend on Y *only* through their corresponding α_k .

Latent variables are fully characterized by the $2K$ probabilities

$$Pr(\alpha_k = 0 | Y = 0) \quad Pr(\alpha_k = 1 | Y = 1), \quad (8)$$

for $k = 1, \dots, K$. Similarly, classifiers are characterized by the $2m$ probabilities

$$\begin{aligned} \eta_j^\alpha &= Pr(f_j(X) = 0 | \alpha = 0) \\ \psi_j^\alpha &= Pr(f_j(X) = 1 | \alpha = 1). \end{aligned} \quad (9)$$

Note that the total number of probabilities to be computed is $2K + 2m$, a number that is remarkably lower than in Method 1

for large correlation groups. In Method 1, $2L$ probabilities are to be computed, where $L = 2^m$, with the number of classifiers in a correlation group m_k at the exponent.

The software implementation of Method 2 is an immediate application of the tree diagram in Fig. 1 (Right)—two layers of binary labels to be generated, one consisting of k elements, and the other of m elements, obeying $K \leq m$. The elements in the k -label layer are generated according to the probabilities $Pr(\alpha_k = 0 | Y = 0)$ or $Pr(\alpha_k = 1 | Y = 1)$, for $k = 1, \dots, K$, depending on the ground truth label Y . On the other hand, the m -label layer is generated according to the probabilities $\eta_j^{\alpha_k}$ or $\psi_j^{\alpha_k}$, for $j = 1, \dots, m$, depending on the previously generated k labels. Repeating this process over n instances, using probabilities that generate strongly enough dependencies between layers of labels, an n -by- m data set is eventually obtained.

IV. RESULTS

The performance of synthetic data sets generated using Method 1 is evaluated with a series of correlation structures. The results for Method 2 are included as well, for comparative purposes.

The evaluated correlation structures are listed in Table I. Denoted as Cases 1–12, a total of 12 different structures are tested. Cases 1–10 feature one single correlation group, with a total of $m = 20$ classifiers. The number of classifiers in the correlation group m_1 increases with the case number—Case 1 features $m_1 = 1$, whereas Case 10 has $m_1 = 10$. Note that Case 1 corresponds to the all-independent structure, featuring $K = 20$ correlation groups.

TABLE I
LIST OF CASE NUMBERS AND CORRESPONDING CORRELATION STRUCTURE

Case No.	1	2	3	4	5	6	7	8	9	10
m_1	1	2	3	4	5	6	7	8	9	10
K	20	19	18	17	16	15	14	13	12	11
Case No.	11					12				
Corr. Str.	{5 5 5 1}					{15 10 1 ... (×25) ... 1}				
m	16					50				
K	4					27				

We measure the averaged degree of correlation featured by synthetic data sets generated using Methods 1 and 2 for Cases 1–12. The covariance matrix characterizing a database containing binarized pathway activation measurements is included to assess the methods' ability to resemble an actual, real-world case. Also, we present testing results using two state-of-the-art binary ensemble meta-learners—the Correlated Expectation-Maximization (CEM) and the Latent Spectral Meta-Learner (LSM), presented in [7] and [8], respectively. We include the relative errors featured by both meta-learners using both methods for Cases 1–12.

A. Degree of correlation

Synthetic correlated data sets can be characterized by the degree of correlation and overall performance of their m -by-

m inter-classifier covariance matrices R , with entries $r_{ij} = \mathbb{E}[(f_i - \mathbb{E}[f_i])(f_j - \mathbb{E}[f_j])]$.

We use the correlation coefficient ρ , defined as

$$\rho := \frac{\sum_{j=1}^m \sum_{j'=1}^{j-1} |\hat{R}(j, j')|}{\sum_{j=1}^m |\hat{R}(j, j')|}, \quad (10)$$

to provide an objective measurement on the overall degree of correlation in a data set. Fig. 3 plots averaged measurements of ρ for Methods 1 and 2, for Cases 1–12 of Table I and $n = 1,000$ instances. Measurements for Method 1 include results for several values of the parameter $\epsilon \in [0, 1]$. The simulation parameters are randomly generated from uniformly distributed r.v. in the ranges $\eta_1^k, \psi_1^k \in [0.5, 0.8]$, $Pr(\alpha_k | Y) \in [0.5, 0.8]$, and $\eta_j^\alpha, \psi_j^\alpha \in [0.7, 0.9]$. We also assume null class imbalance, that is $Pr(Y = 0) = Pr(Y = 1)$.

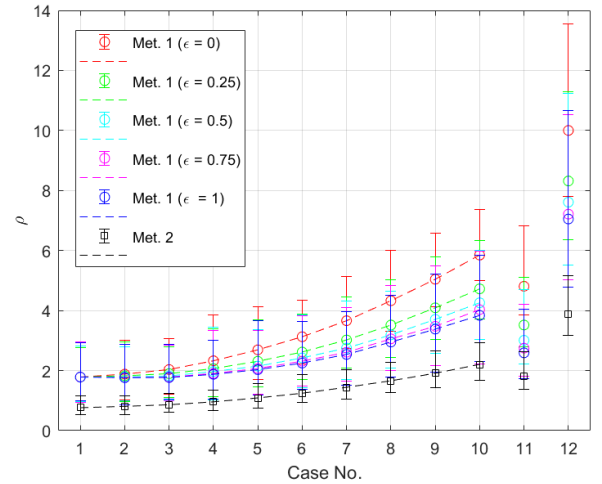


Fig. 3. Averaged correlation coefficient ρ for Methods 1 and 2, and Cases 1–12 of Table I.

Note that Method 1 always features a higher averaged measurement of ρ than Method 2. This occurs regardless of the value of ϵ , which regulates the degree of correlation (see Section II). This behavior is a consequence of the higher degrees of correlation intrinsic to Method 1.

B. Comparing with a real case

We compare the behavior and appearance of the covariance matrix R of a real database containing pathway activation measurements, as presented in [12]. This type of genomic data is easily binarized, and serves as a good example of a real data set. The covariance matrices of the synthetic data sets generated using Methods 1 and 2 are included for comparative purposes. Specifically, we generate the correlation structure $\{3 \ 5 \ 2 \ 3 \ 5\}$, with $\eta_1^k, \psi_1^k = 0.5$, $Pr(\alpha_k | Y) = 0.7$, and $\eta_j^\alpha, \psi_j^\alpha = 0.8$. Note that the degree of correlation within each correlation group is more heterogeneous for Method 1, a more realistic characterization compared to Method 2.

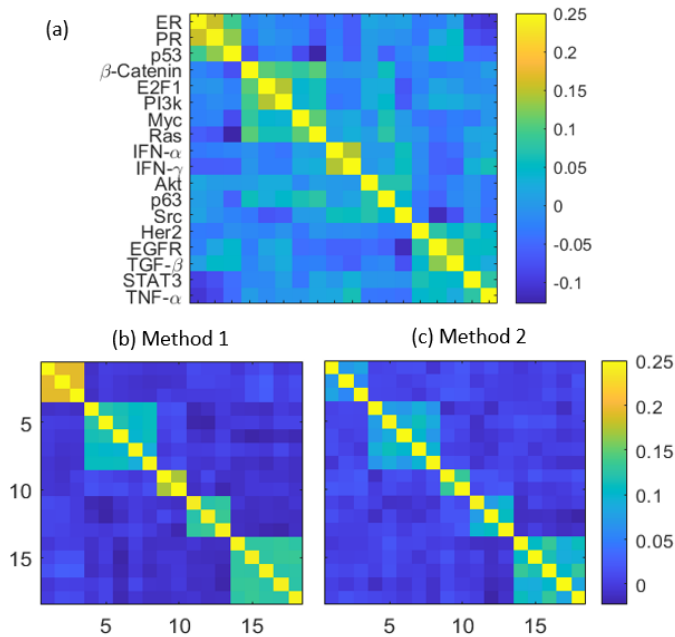


Fig. 4. (a) Covariance matrix obtained from binarized pathway activation measurements in [12], and similar artificial correlation structures obtained using (b) Method 1 and (c) Method 2.

C. Testing the meta-learners

The performances of the two state-of-the-art meta-learners, the CEM and LSM, are evaluated using synthetic data sets generated with Methods 1 and 2. We test all the correlation structures of Table I, assuming that this is an *a priori* knowledge. Performance is quantified using the relative error $\epsilon_r = \frac{FP+FN}{n}$, where *FP* and *FN* stand for the number of *false positives* and *false negatives*, respectively, and *n* is the number of instances in the data set. Simulation parameters are the same as Fig. 3, with $\epsilon = 0.5$.

By inspection of Fig. 5, it may be observed that the CEM outperforms the LSM when using artificial data generated with Method 1. The opposite occurs for Method 2, with the LSM outperforming the CEM. However, both wins occur by a remarkably small margin. Averaged error measurements show lower values in the case of Method 1 compared to Method 2. Also, greater variability is observed in the first case.

Artificially generated data sets using Method 1 show larger values of the correlation coefficient, positively impacting on the meta-learners' performances. The higher variability is also a consequence of the intrinsic higher variability of Method 1, and this is reflected onto the meta-learners' performance results.

V. CONCLUSIONS

An adaptable, easily tunable method for generating binary, correlated synthetic data sets is presented in this work.

The underlying flexibility behind the method, offers better characterization of developing binary meta-learners, providing more varied testing scenarios. For instance, the introduced

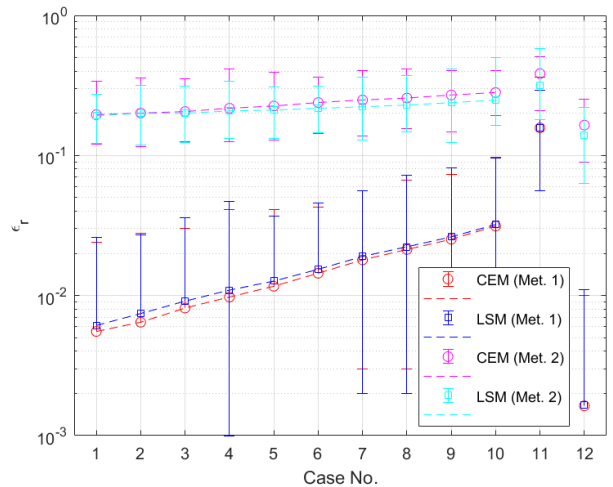


Fig. 5. Averaged relative error ϵ_r for Methods 1 and 2, and Cases 1–12 of Table I.

tuning parameter allows to change the overall degree of correlation featured by inter-dependent classifiers.

The well performance of the method has been assessed by direct comparison with a real, Genomics-related database, and through extensive testing using ensemble meta-learners.

REFERENCES

- [1] M. Micsinai, F. Parisi, F. Strino, P. Asp, B. Dynlacht and Y. Kluger, "Picking ChIP-seq peak detectors for analyzing chromatin modification experiments," *Nucleic Acid Research*, vol. 40, no. 9, p. e70, 2012.
- [2] V. Raykar et al., "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. Apr, p. 1297–1322, 2010.
- [3] J. W. T.-f. Whitehill, J. Bergsma, J. Movellan and P. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Advances in neural information processing systems*, p. 2035–2043, 2009.
- [4] R. Snow, B. O'Connor, D. Jurafsky and A. Ng, "Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 254–263, 2008.
- [5] M. Usman and K. Insoo, "Sensor network-based spectrum sensing for cognitive radio network," *Proc. of the International Conference on Intelligent Systems Engineering*, p. 19–25, 2016.
- [6] A. P. Dawid and A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," *Journal of The Royal Statistical Society Series C-applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.
- [7] M. Cabrera-Bean, A. Pagès-Zamora, and C. Díaz-Vilor, "Unsupervised Ensemble Classification With Correlated Decision Agents," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1085–1089, 2019.
- [8] A. Jaffe, E. Fetaya, Ethan, B. Nadler, T. Jiang, and Y. Kluger, "Unsupervised Ensemble Learning with Dependent Classifiers," *Proc. of Artificial Intelligence and Statistics Conference*, pp. 351–360, 2016.
- [9] B. Dai and S. Ding, and G. Wahba, "Multivariate Bernoulli distribution," *Bernoulli*, vol. 19, no. 4, pp. 1465–1483, 2013.
- [10] R. M. Pires, C.A.R. Diniz, "Correlated Binomial Regression Models," *Computational Statistics and Data Analysis*, vol. 56, no. 8, pp. 2513–2525, 2012.
- [11] M. Hisakado, K. Kitsukawa, and S. Mori, "Correlated binomial models and correlation structures," *Journal of Physics A: Mathematical and General*, vol.39, no. 50, 2006.
- [12] M. Gatza et al., "A pathway-based classification of human breast cancer," *Proc. of the National Academy of Sciences of the USA*, vol. 107, no. 15, p. 6994–6999, 2010.