FEATURE SELECTION FOR COMPOSITE HYPOTHESIS TESTING WITH SMALL SAMPLES: FUNDAMENTAL LIMITS AND ALGORITHMS

Dayu Huang and Sean Meyn

CSL & ECE
University of Illinois at Urbana-Champaign
1308 West Main Street, Urbana, IL 61801, USA

ABSTRACT

This paper considers the problem of feature selection for composite hypothesis testing: The goal is to select, from m candidate features, r relevant ones for distinguishing the null hypothesis from the composite alternative hypothesis; the training data are given as L sequences of observations, of which each is an n-sample sequence coming from one distribution in the alternative hypothesis.

What is the fundamental limit for successful feature selection? Are there any algorithms that achieve this limit? We investigate this problem in a small-sample high-dimensional setting, with $n = o(m)$, and obtain a tight pair of achievability and converse results:

(i) There exists a function $f(L, n, r, m)$ such that if $f(L, n, r, m) \downarrow 0$, then no asymptotically consistent feature selection algorithm exists;

(ii) We propose a feature selection algorithm that is asymptotically consistent whenever $f(L, n, r, m) \uparrow \infty$.

Index Terms— Feature selection, high-dimensional model, composite hypothesis testing, small sample, supervised learning

1. INTRODUCTION

Composite hypothesis testing problems arise from many anomaly detection applications such as surveillance and network security, medical and public heath (See the survey paper [1] for many applications). The statistics of normal behavior are given and the goal is to infer whether a sequence of test data is from the normal behavior (null hypothesis) or from one of many possible abnormal behaviors (alternative hypothesis).

In many applications, a large number of features are measured, while in fact only a small subset of features are relevant for inferring which hypothesis is true. By identifying the set of relevant features and only using them in the test, it is possible to improve the test performance of the test (see Lemma 1). The technique of finding these relevant features is known as *feature selection*.

In this paper, we study the problem of feature selection for composite hypothesis testing. Our goal is evaluate the performance of different feature selection algorithms, and our criterion is sample complexity. We first derive a necessary condition, also called a hardness result or lower-bound, on the number of samples required for successfully identifying the relevant features. We then design an algorithm whose sample complexity almost achieves this lower-bound.

To be more specific, in the composite hypothesis testing problem considered in this paper, the null distribution π of the normal behavior is known, or can be estimated precisely from records of the normal behavior. The set of alternative distributions Π_m is unknown, and the amount of data recording abnormal behaviors is limited.

The number of recorded abnormal behaviors L, the number of data samples n for each anomaly, the number of candidate features m, and the number of relevant features r are quantities of interest when characterizing the performance of feature selection algorithms. We consider a small-sample high-dimension model in which $L, n, m, r \to \infty$ and $n = o(m)$, and ask how (L, n) should depend on (m, r) in order for the algorithm to be asymptotically consistent, i.e., the probability of correctly learning the relevant features converges to one.

The main contribution of this paper is identifying the fundamental limit, described by a function $f(L, n, r, m)$ and proposing a near optimal algorithm:

(i) Hardness result: If $f(L, n, r, m) \downarrow 0$, then no asymptotically consistent feature selection algorithm exists;

(ii) Achievability result: The proposed feature selection algorithm is asymptotically consistent whenever $f(L, n, r, m) \uparrow \infty$.

Precise statements are given in Theorem 1.

Financial support from the National Science Foundation (NSF CCF 07-29031 and CCF 08-30776), ITMANET DARPA RK 2006-07284 and AFOSR grant FA9550-09-1-0190 is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, DARPA or AFOSR.

1.1. Related work

There are many approaches to feature selection (See [2]). The statistic used in the proposed feature selection algorithm bears some similarity to the chi-square statistic commonly used in feature selection techniques (e.g. [3]). It is not clear whether the feature selection algorithm based on the chi-square statistic also achieves the fundamental limit given in Theorem 1.

The results in [4] and [5] both theoretically characterize sample complexity of feature selection algorithms for the binary classification problem, in which there is one distribution in each class. Their results do not apply to the composite hypothesis testing problem studied in this paper, and the proof techniques are also different.

The high-dimensional model considered in this paper is similar to those investigated in [6, 7] and the converse result in this paper is based on a similar proof technique.

2. MODEL AND DEFINITIONS

2.1. Composite hypothesis testing

An i.i.d. sequence $\boldsymbol{Z}_1^t = \{Z_1, \dots, Z_t\}$ is observed, where $Z_i \in [m] := \{1, 2, \dots, m\}$. Each element in $[m]$ is a feature. Let $\mathcal{P}(\mathsf{Z})$ denote the set of distributions over $[m]$. Under normal behavior, Z_i has a known distribution $\pi \in \mathcal{P}(\mathsf{Z})$. Under abnormal behaviors, Z_i is distributed according to some $\mu \in \Pi_m \subset \mathcal{P}(\mathsf{Z})$, where Π_m is set of possible abnormal behaviors. A test is a sequence of binary-valued functions $\{\phi_t\}$ with $\phi_t : [m]^t \to \{0, 1\}$.

A fixed set $\mathcal{S} \subseteq [m]$ represents the set of "relevant features". Relevancy is made precise as follows: The set of alternative distributions Π_m is a subset of

$$\{\mu \in \mathcal{P}(\mathsf{Z}) : d_{\mathcal{S}}(\mu, \pi) \geq \epsilon, d_{[m]\setminus\mathcal{S}}(\mu, \pi) = 0\}, \qquad (1)$$

where the pseudo-metric $d_{\mathcal{S}}$ is defined as

$$d_{\mathcal{S}}(\mu, \pi) = \sum_{j \in \mathcal{S}} |\mu(j) - \pi(j)|.$$

Two assumption are made when Π_m is assumed to take the form in (1): 1) The anomaly behaviors are sufficiently different from normal behavior, modeled by $d_{\mathcal{S}}(\mu, \pi) \geq \epsilon$; 2) Only features in $\mathcal{S}$ are relevant, modeled by $d_{[m]\setminus\mathcal{S}}(\mu, \pi) = 0$. The second assumption can be relaxed, as discussed in Section 6.

How does the cardinality of $\mathcal{S}$, which we denote by r, affect the performance of tests? This is studied in [6] where $\mathcal{S} = [m]$ and π is the uniform distribution. The results in that paper can be extended to the current case where $\mathcal{S} \subseteq [m]$:

Lemma 1. *Suppose $n = o(m)$, π is uniform, and Π_m is given by* (1). *Then*

(i) *There exists an asymptotically consistent test if*

$$\lim_{m\to\infty} \frac{t^2 \varepsilon^4}{m(r/m)^4} = \infty.$$

(ii) *No asymptotically consistent test exists if*

$$\lim_{m\to\infty} \frac{t^2 \varepsilon^4}{m(r/m)^4} = 0.$$

2.2. Feature selection

Suppose the normal behavior π is known exactly, a reasonable assumption when the amount of data for normal behavior is large. Suppose L independent anomaly data sequences $\{\boldsymbol{Y}_1^{n,(l)}, 1 \leq l \leq L\}$ are given, each representing a different abnormal behavior: Each sequence $\boldsymbol{Y}_1^{n,(l)} = \{Y_i^{(l)}, \dots, Y_n^{(l)}\}$ is i.i.d with $\mu^{(l)} \in \Pi_m$. The assumptions on $\mu^{(l)}$ are given in Section 3. We note that the distribution of the test sequence might differ from the distribution of any training sequence $\{\boldsymbol{Y}_1^{n,(l)}, 1 \leq l \leq L\}$.

Our task is to design a feature selection algorithm $\psi = \{\psi_1, \dots, \psi_n\}$, given by a sequence of set-valued functions: $\psi : [m]^{n\times L} \to 2^{[m]}$. The feature selection algorithm is asymptotically consistent, if for any collection of $\{\mu^{(l)}\}$,

$$\lim_{m\to\infty} \mathsf{P}\{\psi_n(\{\boldsymbol{Y}_1^{n,(l)}, 1 \leq l \leq L\}) = \mathcal{S}\} = 1,$$

3. MAIN RESULTS

We begin with assumptions on the training data used for feature selection: First, the number of samples per anomaly and the number of relevant features are both small:

Assumption 1. $n = o(m)$, $\limsup_{m\to\infty} r/m < 1$.

Second, no feature plays a dominant role:

Assumption 2. *There exists $c_2, c_1 > 0$ such that $c_2/m \leq \pi(j) \leq c_1/m, \mu^{(l)}(j) \leq c_1/m$.*

Third, the information on the relevant features given by the training data increases with L:

Assumption 3. $\eta := \min_{j\in\mathcal{S}} \frac{1}{L} \sum_{l=1}^{L} |\mu^{(l)}(j) - \pi(j)| \asymp \varepsilon/r$.

Notation $h \asymp g$ means $0 < \liminf h/g \leq \limsup h/g < \infty$. Note that we do not exclude the case where $\mu^{(l)}(j) = \pi(j)$ for some l and $j \in \mathcal{S}$, i.e., some abnormal behaviors are different from the normal behavior only on a subset of the relevant features.

Our main result answers the following question: How fast should L and n grow in order for a feature selection algorithm to be asymptotically consistent?

Theorem 1. *Suppose Assumption 1-3 hold.*

(i) *There exists a feature selection algorithm that is asymptotically consistent, if*

$$\lim_{m\to\infty} \frac{n^2 \varepsilon^4 L}{(r/m)^4 m^2 \log m} = \infty. \qquad (2)$$

(ii) *No feature selection algorithm is asymptotically consistent, if*

$$\lim_{m\to\infty} \frac{n^2 \varepsilon^4 L}{(r/m)^4 m^2 \log m} = 0. \qquad (3)$$

We have the following remarks:

(i) Consider the case $n = m/\mathrm{polylog(m)}$, and $\varepsilon \asymp r/m$. Theorem 1 implies the number of observed abnormal behaviors L should increase with m as $L = \mathrm{polylog(m)}$, i.e., the relevant features can be correctly identified even when the number of irrelevant features is much larger than the number of observed anomalies.

(ii) Consider the cases $\varepsilon = \alpha r/m$ where $\alpha = o(1)$: The difference between the normal and abnormal behaviors becomes less concentrated as α decreases. Theorem 1 implies n^2L needs to increase proportionally to $1/\alpha$, indicating that less concentration makes it harder to learn the features.

(iii) Note that n, L enters $f(L, n, r, m)$ as n^2L. Suppose we have a fixed budget of total number of samples nL. This result implies that in the small sample case $n = o(m)$, it is advantages to spend more samples on each abnormal behavior instead of observing more abnormal behaviors.

4. PROOF OF THE HARDNESS RESULT

We focus on the case where π is the uniform distribution; extensions to the non-uniform case are straightforward. We first construct a hypothesis testing problem that is no harder than the feature selection problem, i.e., an asymptotically consistent feature selection algorithm gives an asymptotically consistent test. We then show that no asymptotically consistent test exists for the testing problem when (3) holds.

We first construct the distributions used to define the two hypotheses in the hypothesis testing problem. Assume $(m-r+2)/4$ is an integer; extensions to other cases are straightforward. Let $T = m - r + 2$. For each $y \in [T/2], z \in \{-1, 1\}$, define the distribution $\mu_{y,z}$ as follows:

$$\mu_{y,z}(j) = \begin{cases} \frac{1}{m} - \eta & j \text{ is odd}, m \geq j > T \\ \frac{1}{m} + \eta & j \text{ is even}, m \geq j > T \\ \frac{1}{m} - \eta z & j = 2y-1 \\ \frac{1}{m} + \eta z & j = 2y \\ \frac{1}{m} & \text{otherwise.} \end{cases}$$

Note that the distribution $\mu_{y,z}$ is in the set Π_m with $\mathcal{S} = \{2y, 2y-1\} \cup \{T+1, \ldots, m\}$. Let $\mathbf{z} \in \{-1, 1\}^L$, and denote $\mathbf{j} = \{\mathbf{j}_1, \ldots, \mathbf{j}_L\}$. Define the following distribution on $[m]^L$:

$$\mu_{y,\mathbf{z}}(\mathbf{j}) = \prod_{l=1}^{L} \mu_{y,\mathbf{z}_l}(\mathbf{j}_l).$$

Define two sets of distributions $\mathcal{A}$ and $\mathcal{B}$ as

$$\mathcal{A} = \{\mu_{y,\mathbf{z}} : 1 \leq y \leq \frac{T}{4}, \mathbf{z} \in \{-1,1\}^L\},$$

$$\mathcal{B} = \{\mu_{y,\mathbf{z}} : \frac{T}{4} + 1 \leq y \leq \frac{T}{2}, \mathbf{z} \in \{-1,1\}^L\}.$$

Consider the following binary composite hypothesis testing problem: L sequences of observations of length n $\{\boldsymbol{Y}_1^{n,(l)}, 1 \leq l \leq L\}$ are given, where $Y_i^{(l)}$ is i.i.d. with marginal $\mu^{(l)}$. Our task is to decide between the following two hypothesis:

$$\overline{H0} : \mu \in \mathcal{A}, \quad \overline{H1} : \mu \in \mathcal{B}. \tag{4}$$

Roughly speaking, both hypotheses agree that $\{T+1, \ldots, m\}$ is a subset of relevant features. The rest two features are in the first half and second half of $[T/2]$ for $\overline{H0}$ and $\overline{H1}$, respectively.

Suppose ψ is an asymptotically consistent feature selection algorithm. Then the test that decides in favor of $\overline{H1}$ if $\big(\psi_n(\{\boldsymbol{Y}_1^{n,(l)}\}) \cap [T/2] = \emptyset$ is asymptotically consistent.

We now show that there is no asymptotically consistent test for the problem (4). Our main tool is the following results in [8]. Let $\mathsf{conv}(\cdot)$ denote the convex hull of a set. The set $\mathcal{A}_n$ is the set of nth-order product of distributions in $\mathcal{A}$, i.e., $\mathcal{A}_n = \{\mu^n : \mu \in \mathcal{A}\}$.

Lemma 2. *If there are a sequence of distributions* $\{\nu^n\} \in \mathsf{conv}(\mathcal{A}_\mathsf{n})$, $\{\bar{\nu}^n\} \in \mathsf{conv}(\mathcal{B}_\mathsf{n})$, *such that*

$$\lim_{n \to \infty} \|\nu^n - \bar{\nu}^n\|_1 < 2,$$

then no asymptotically consistent test exists for the binary hypothesis testing problem (4).

The key is then to construct the two distributions ν^n and $\bar{\nu^n}$. We use the "mixture measure" technique. Let $\mu_{y,\mathbf{z}}^n$ denote the nth-order product of $\mu_{y,\mathbf{z}}$, i.e. the distribution of a length-n sequence generated i.i.d. with $\mu_{y,\mathbf{z}}$. We construct two mixing distributions, one from $\mathsf{conv}(\mathcal{A}_\mathsf{n})$ and the other from $\mathsf{conv}(\mathcal{B}_\mathsf{n})$:

$$\nu^n = \frac{4}{T}\sum_{y=1}^{T/4} \frac{1}{2^L} \sum_{\mathbf{z}} \mu_{y,\mathbf{z}}^n, \quad \bar{\nu}^n = \frac{4}{T}\sum_{y=T/4+1}^{T/2} \frac{1}{2^L} \sum_{\mathbf{z}} \mu_{y,\mathbf{z}}^n.$$

We can show that

Lemma 3.

$$\|\nu^n - \bar{\nu}^n\|_1 \leq 8\exp\{4n^2m^2(\varepsilon/r)^4L - \log(m-r)\}.$$

Therefore, the condition in Lemma 2 is satisfied when (3) holds. We conclude that no asymptotic consistent feature selection algorithm exists.

5. A FEATURE SELECTION ALGORITHM

We propose the following feature selection algorithm and show that it is asymptotically consistent when (2) holds. Denote the empirical distribution for the sequence $\boldsymbol{Y}_1^{n,(l)}$ by $\Gamma^{(l)}(j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{Y}_i^{(l)} = j\}$. Denote

$$g_j = \tfrac{1}{L}\textstyle\sum_{l=1}^{L} g_j^{(l)},$$

where

$$g_j^{(l)} = \begin{cases} \frac{1}{2}n^2\pi(j)^2, & n\Gamma^{(l)}(j) = 0 \\ -n\pi(j), & n\Gamma^{(l)}(j) = 1 \\ 1, & n\Gamma^{(l)}(j) \geq 2. \end{cases}$$

The proposed feature selection algorithm is

$$\psi = \{j \in [m] : g_j \geq \tau\}. \tag{5}$$

To bound the probability of error, we begin with the expectation and variance of g_j when $j \in \mathcal{S}$ and $j \notin \mathcal{S}$ are given in Lemma 4.

Lemma 4. *If $j \notin \mathcal{S}$, then*

$$\mathsf{E}[g_j] = O(n^3/m^3).$$

If $j \in \mathcal{S}$, then

$$\mathsf{E}[g_j] = O(n^3/m^3) + \tfrac{1}{2}\frac{1}{L}\sum_{l=1}^{L}[(\mu^{(l)}(j) - \pi(j))^2].$$

Regardless of whether $j \in \mathcal{S}$, we have

$$\mathsf{Var}\,(g_j^{(l)}) = O(n^2/m^2).$$

Since $\sum_{l=1}^{L}(\mu^{(l)}(j) - \pi(j))^2/L \geq \eta^2$, we choose $\tau = \eta^2/4$. Applying Chebyshev's inequality leads to an achievability result that is not as good as (2): Instead of $m^2 \log m$, we would have m^3 in the denominator of (2). To improve it to $m^2 \log m$, we need to apply the Chernoff bound. Denote the log-moment-generating function

$$\Lambda_{j,(l)}(\theta) = \log(\mathsf{E}[\exp(\theta g_j^{(l)})].$$

Lemma 5. *The following holds for fixed θ,*

$$\sum_{l=1}^{L}\Lambda_{j,(l)}(\theta) = \theta\mathsf{E}[g_j] + \tfrac{1}{2}\theta^2 O((\frac{n}{m})^2). \tag{6}$$

Proof for Lemma 5. The following well-known expansion of the log-moment-generating function can be obtained via the mean value theorem:

$$\Lambda_{j,(l)}(\theta) = \theta\mathsf{E}[g_j^{(l)}] + \tfrac{1}{2}\mathsf{Var}_{\check{\nu}_j^{(l)}}[g_j^{(l)}]$$

where the *twisted distribution* $\check{\nu}_j^{(l)}$ is defined as follows: For random variable h,

$$\mathsf{E}_{\check{\nu}_j^{(l)}}[h] = \frac{\mathsf{E}_{\mu^{(l)}}[e^{\bar{\theta}g_j^{(l)}}h]}{\mathsf{E}_{\mu^{(l)}}[e^{\bar{\theta}g_j^{(l)}}]},$$

where $\bar{\theta}$ satisfies $|\bar{\theta}| \leq |\theta|$. Our choice of $g_j^{(l)}$ satisfies $g_j^{(l)} = O(1)$. Therefore, for $\theta = O(1)$, we obtain

$$\mathsf{Var}_{\check{\nu}_j^{(l)}}[g_j^{(l)}] = O(\mathsf{Var}\,[g_j^{(l)}]).$$

The conclusion follows from the independence of the sequence $\{g_j^{(l)}, 1 \leq l \leq L\}$. □

Proof for the achievability result in Theorem 1. Applying the Chernoff bound, we obtain the following upper-bound:

$$\mathsf{P}\{g_j \geq \tau | j \notin \mathcal{S}\} \leq \inf_{\theta}\exp\{-\theta L\tau + \sum_{l=1}^{L}\Lambda_{j,(l)}(\theta)\}. \tag{7}$$

Substituting (6) into (7), we obtain for any $\theta = O(1)$, there exist $\kappa > 0$ such that for large enough n,

$$\mathsf{P}\{g_j \geq \tau | j \notin \mathcal{S}\} \leq \exp\{-\theta L\tau + \tfrac{1}{2}\theta^2\kappa L(\frac{n}{m})^2\}.$$

Take $\theta = m^2\tau/(n^2\kappa)$, and note that $\tau = \eta^2/4$, we obtain

$$\mathsf{P}\{g_j \geq \tau | j \notin \mathcal{S}\} \leq \exp\{-n^2m^2\eta^4 L/(32\kappa)\}. \tag{8}$$

Similarly, we can bound

$$\mathsf{P}\{g_j \leq \tau | j \in \mathcal{S}\} \leq \exp\{-n^2m^2\eta^4 L/(32\kappa)\}. \tag{9}$$

Apply the union bound with (8) and (9), we obtain

$$\mathsf{P}\{\psi \neq \mathcal{S}\} \leq \exp\{\log(m) - n^2m^2\eta^4 L/(32\kappa)\}.$$

If the assumption $\lim_{m\to\infty}\frac{n^2\varepsilon^4 L}{(r/m)^4 m^2 \log m} = \infty$ holds, then the exponent on the right-hand side decreases to $-\infty$, thus the feature selection function ψ is asymptotically consistent. □

6. CONCLUSIONS AND DISCUSSIONS

We have shown there is a fundamental limit on the sample complexity of the feature selection algorithm for composite hypothesis testing and propose an algorithm that nearly achieves this limit. Possible extensions of results include

The assumption that $d_{[m]\setminus\mathcal{S}}(\mu, \pi) = 0$ can be relaxed. For example, one can assume that $|\mu(j) - \pi(j)| \leq \bar{\varepsilon}$ for $j \notin \mathcal{S}$. The same feature selection algorithm will work with a different choice of τ (See Equation (5)).

The feature selection algorithm is proposed for the case n/m is small. To extend the algorithm to the case where n/m is moderate, we could assign different coefficients for $n\Gamma^{(l)}(j) = k$ for $k \geq 2$, rather than aggregating $n\Gamma^{(l)}(j) \geq 2$ together. Those coefficients can be determined using a finer asymptotic expansion of $\mathsf{E}[g_j]$.

7. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, pp. 15 – 1 – 15 – 58, Jul. 2009.

[2] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *7th IEEE International Conference on Tools with Artificial Intelligence*. IEEE, 1995, pp. 388 – 391.

[4] A. Y. Ng, "On feature selection: Learning with exponentially many irrelevant features as training examples," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 404 – 412.

[5] V. Y. F. Tan, M. Johnson, and A. S. Willsky, "Necessary and sufficient conditions for high-dimensional salient feature subset recovery," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, Jun. 2010, pp. 1388 – 1392.

[6] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct. 2008.

[7] B. G. Kelly, T. Tularak, A. B. Wagner, and P. Viswanath, "Universal hypothesis testing in the learning-limited regime," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, Jun. 2010, pp. 1478 – 1482.

[8] L. M. L. Cam, *Asymptotic methods in statistical decision theory*. Springer, 1986.