한국정보전자통신기술학회논문지(jkiiect)'20-4, Vol.13 No.2
논문 20-13-02-95 http://dx.doi.org/10.17661/jkiiect.2020.13.2.95

Maximum Product Detection Algorithm for Group Testing Frameworks

Jin-Taek Seong*

Abstract In this paper, we consider a group testing (GT) framework which is to find a set of defective samples out of a large number of samples. To handle this framework, we propose a maximum product detection algorithm (MPDA) which is based on maximum a posteriori probability (MAP). The key idea of this algorithm exploits iterative detection to propagate belief to neighbor samples by exchanging marginal probabilities between samples and output results. The belief propagation algorithm as a conventional approach has been used to detect defective samples, but it has computational complexity to obtain the marginal probability in the output nodes which combine other marginal probabilities from the sample nodes. We show that the our proposed MPDA provides a benefit to reduce computational complexity up to 12% in runtime, while its performance is only slightly degraded compared to the belief propagation algorithm. And we verify the simulations to compare the difference of performance.

Key Words : Belief Propagation, Group Testing, Maximum a Posteriori Probability, Maximum Product

1. Introduction

Group Testing (GT) was introduced by Dorfman [1], and its application has been used in various fields for half a century [2]. GT began with a project to find all syphilis soldiers in the US Public Health service during World War II. At the time, syphilis inspection took blood samples of individual soldiers to check for syphilis infection. However, because the number of soldiers in the syphilis inspection was very large, the cost of the test was huge, and it took a lot of time to find a new test method [2]. Subsequently, it was first motivated by the development of the GT framework [1].

In the conventional GT, the syphilis inspection was performed using the following method. First, blood samples from several soldiers are mixed in one pool to see if they react to syphilis source of infection. And when the result is positive, it means that at least one soldier was infected with syphilis. On the other hands, in the case of a negative, it can be confirmed that all blood samples used for the syphilis infection were not infected with the syphilis. This is because most soldiers are not infected with syphilis, and only a handful of soldiers have syphilis. So the GT problem is mainly dealt with the following two directions. First, it is about how to pick samples to be included in a pool. The second is to use a detection algorithm to find a set of defective samples out of a large number of samples.

In this paper, the GT problem is clearly defined as follows. Let T be the number of

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the korean government (NRF-2017R1C1B5075823).
*Department of Convergence Software, Mokpo National University (jtseong@mokpo.ac.kr)
Received April 03, 2020 Revised April 05, 2020 Accepted April 14, 2020

tests required to find a set of defective sample when D samples of the N samples are defective. In this paper, we propose a maximum product detection algorithm (MPDA) using maximum a posteriori probability (MAP). And we see that the our proposed MPDA provides a benefit to reduce computational complexity, while its performance is only slightly reduced compared to a belief propagation algorithm which is a conventional approach. In addition, we show simulation results for both algorithms to compare how difference of performance between them.

The organization of this paper is as follows. In Section 2, we investigate the related works in detection algorithms of GT problems. And the GT problem is defined in detail in this paper in Section 3. In addition, the description of the detection algorithm proposed in this paper is provided in Section 4, as well as we show the simulation results and compare other theoretic results. Finally, in Section 5 we conclude that we have obtained meaning results and findings.

2. Related Work

A number of detection algorithms for GT problems have been proposed since the detection algorithm was first introduced by Dorfman. This section aims to review the some detection algorithms proposed in the GT problems.

The detection algorithm to be reviewed first is a binary splitting algorithm [2]. This algorithm is basically called the optimal adaptive algorithm in GT. The binary splitting algorithm is to find defect samples less than or equal to D in N samples, and is summarized as follows according to the size of N and D.

At the initial step, in case of $N \leq 2D-2$, to find D defective samples it is performed by individual testing. It means that individual testing is better than GT when there are many defective samples. Otherwise, set $L = N-D+1$ and $\alpha = \lfloor \log_2 L/D \rfloor$, respectively. At the second step, GT is performed by a set of samples with 2^α size for every testing. Here when the result of this GT is negative, all samples in this pool are determined as normal. An then, reset samples with $N = N-2^\alpha$ size and perform GT as same as the first step. On the other hands, using binary search, one defective sample and the other normal samples S are again set as follows, $N = N-1-S$ and $D = D-1$.

The number of tests T for the generalized binary splitting algorithm with respect to $p > 0$, N and D, is required as $T = (\alpha+2)D+p-1$, where in case high N/D, T converges to $D\log_2(N/D)$ [2].

Next, let us investigate a COMP (Combinatorial Orthogonal Matching Pursuit) algorithm [3]. This algorithm is a class of the nonadaptive GT algorithms. The COMP algorithm works as follows. First, each entry of the group matrix is assumed with the i.i.d.(independent and identically distributed) Bernoulli probability distribution with the probability $1/D$ for 1, and $1-1/D$ for 0. The core idea of the detection algorithm is to combine the columns of the group matrix corresponding to the samples participated in a pool. As the conventional GT problems, the results are determined to be positive or negative depending on the existence of defective samples. The number of test T for the COMP algorithm with any constant $\epsilon > 0$

and that the average error probability is less then or equal $N^{-\epsilon}$, is as follows, $T \geq eD(1+\epsilon)\ln(N)$ [3].

Another algorithm that is a extended version of the COMP algorithm is as called DD (Definite Defectives) to vanish false positive error [4]. The performance of the DD algorithm improves that of the COMP algorithm. The detection method of the DD algorithm exploits useful attributes of the COMP algorithm. Note that the normal samples obtained from the COMP algorithm are surely detected without false negative error. Therefore the DD algorithm only generates false negatives compared to the COMP algorithm.

The SCOMP (Sequential COMP) algorithm is an algorithm that takes advantage of the fact that the DD algorithm does not cause errors until the last step [6]. All remaining samples are assumed to be normal. Let K be the set of samples detected to be defective. If the test contains at least one defective sample from the set K, a positive result is obtained. Note that it cannot be said that the set of defective samples detected by the DD algorithm occurs all positive results. This means that test results that cannot be clearly identified have to contain one hidden defect sample. Simulation results using the SCOMP algorithm showed results close to the optimal ones [4]. The other results of adaptive and noisy GT problems are presented in [5]-[7].

3. Group Testing Framework

In this section, we define the GT framework in detail. Let $X=(x_1,x_2,\cdots,x_N)^T$ be the binary input vector with size N where $X\in\{0,1\}^N$. If the jth sample of X is defective, then we present as $x_j=1$, otherwise $x_j=0$. So all the entries of the binary input vector X are represented as 0 or 1. We assume in this paper that each entry of the vector X has the following i.i.d. Bernoulli probability distribution,

$$\Pr\{x_j=\theta\}=\begin{cases}1-\epsilon & \text{if } \theta=0\\ \epsilon & \text{if } \theta=1\end{cases} \quad (1)$$

where $\epsilon := D/N$ denotes the defective rate, and θ is a dummy variable.

The group matrix $A \in \{0,1\}^{T\times N}$ has T rows and N columns. For $i\in\{1,2,3,...,T\}$ and $i\in\{1,2,3,...,T\}$, if the ith group includes jth sample, the corresponding entry A_{ij} of the group matrix A is represented as $A_{ij}=1$, otherwise, we express as $A_{ij}=0$. In other words, when the entry of the group matrix is 1, GT is performed including the jth sample indicating the corresponding column [2].

The following describes the mathematical expression of GT in more detail. The binary input vector X and the group matrix A defined above are defined by the following GT:

$$Y=A\oplus X \quad (2)$$

where Y is the result vector and if the result of the ith group is positive, then we say as $y_i=1$, otherwise it is 0. And the symbol $\oplus$ denotes logical operation with AND and OR. The following equation (3) shows a simple example of the mathematical expression of GT, $Y=A\oplus X$,

$$\begin{bmatrix}1\\1\\0\end{bmatrix}=\begin{bmatrix}1&1&1\\0&1&1\\1&0&0\end{bmatrix}\oplus\begin{bmatrix}0\\1\\0\end{bmatrix} \quad (3)$$

From (3), the first entry of the result $[1\ 1\ 0]^T$ is obtained as (1 1 1)$\oplus$(0 1 0) = (1 AND 0) OR (1 AND 1) OR (1 AND 0) = 1. It is positive. In the same manner, we calculate all the entries of the output vector as a result of

the GT. As shown in the example, if at least one defective samples of the vector included in each group exist, the result is positive. This mathematical expression of GT takes advantage into easy handling states of X, A, and Y in our proposed detection algorithm.

4. Detection of Defective Samples

4.1 Proposed Detection Algorithm

In this section, we proposes a Maximum Product detection algorithm (MPDA) for the GT problems. This MPDA is based on using MAP. Note that the GT problem of finding the optimal MAP solution is NP-hard. Although this argument is difficult to find the optimal solution in fact, many researchers have tried to find suboptimal approaches close to the optimal one. Among them, the performance of the belief propagation algorithm introduced by Mackey in [8] showed results close to the Shannon bounds in channel coding theory.

To treat our proposed MPDA, we assume the following: Each sample x_j, $j=1,2,\cdots,N$, has a priori probability of defective and normal state given by (1), under the system assumption that samples, group matrix, and result vector are independent with each other, the GT problem is to find the MAP combination $\hat{X}$ of samples given the observed output Y of result vectors. This is formulated as

$$\begin{aligned}\hat{X} &= \arg\max \Pr\{X|Y\} \\ &= \arg\max \prod_{j=1}^{N} \Pr\{x_j|Y\}\end{aligned} \tag{4}$$

where the second equality comes from the independent assumption of priori samples.

Using Bayes' rule, the conditional probability $\Pr(x_j|Y)$ in (4) can be rewritten by where the

$$\begin{aligned}\Pr\{x_j|Y\} &= \sum_{X\setminus\{x_j\}} \Pr\{X,Y\} \\ &= \sum_{X\setminus\{x_j\}} \Pr\{Y|X\}\Pr\{X\} \\ &= \sum_{X\setminus\{x_j\}} \prod_{i=1}^{T} \Pr\{y_i|X\} \prod_{j=1}^{N} \Pr\{x_j\}\end{aligned} \tag{5}$$

symbol $\setminus$ refers to exclusion from a set, and independent assumptions lead to equality in (5). The aim of the proposed algorithm is to find the maximized marginal probability for each sample in (5).

Next, we describe the key idea of the MPDA proposed in this paper. Before explaining our algorithm, the graphical representation of for one example of GT in (3) is introduced as shown in Figure 1. There are 3 samples, x_1,x_2,x_3 and 3 outputs of GT, y_1,y_2,y_3. Since the first row of A in (3) is (1 1 1), there exists 3 edges between 3 samples, x_1,x_2,x_3, and y_1. In the same way, other edges between samples and outputs can be draw as shown in Figure 1.

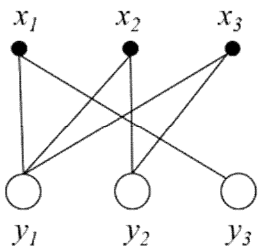

Fig. 1. Graphical representation with 3 samples and 3 output results for the group matrix A in the given example of (3).

Let $\ell(i)=\{x_j : A_{ij}=1\}$ be the set of samples participating in the ith group, and $\mathfrak{I}(j)=\{y_i : A_{ij}=1\}$ be the set of groups participating in the jth sample. We also use $\ell(i)\setminus\{j\}$ to denote the set $\ell(i)$ excluded the j th sample, and $\mathfrak{I}(j)\setminus\{i\}$ to denote the set $\mathfrak{I}(j)$ excluded the ith group. The MPDA proposed in this paper is mainly described as a process in

which two probabilities exchange information in each iteration. Note that we aim to find the maximum conditional probability for each sample as in the last line of (5). In other words, the two conditional probabilities, $\Pr\{x_j|Y\}$ and $\Pr\{y_i|X\}$, are exchanged with each other to maximize the posteriori probability. Let ξ_{ji} be the upward message from the sample x_j to the output y_i, and δ_{ij} be the downward message from the output y_i and the sample x_j.

Now the MPDA updates the messages ξ_{ji} and δ_{ij} associated with each edge between a sample x_j and an output y_i. There are 3 steps to estimate each input sample $\hat{x_j}$: initialization, update the messages ξ_{ji} and δ_{ij}, tentative decoding as to check if the constraint condition $Y=A\oplus\hat{X}$ where $\hat{X}$ denotes the estimated state of the unknown samples. In the initialization step, we define the probability distribution of X in (1), generate the group matrix A with random design, i.e., low-density parity check codes in [7], and obtain the result output vector Y from the given A and X. We aim to find the original input vector X by using known A and Y. In addition, the initial upward message ξ_{ji} can be obtained from the priori probability $\Pr(x_j=\theta)$ assuming that the downward messages for 0 and 1 are equally distributed.

Next we consider a downward message δ_{ij} from an output y_i to a sample x_j. In the conventional belief propagation algorithm [7], this message δ_{ij} is obtained as follows

$$\delta_{ij}(\theta)=\sum_{\{X:\, y_i=A_i\oplus X\}}\ \prod_{j'\in T(i)\setminus\{j\}}\xi_{j'i}(\theta) \qquad (6)$$

In this step of our proposed MPDA, we reduce the complexity of computation for the message δ_{ij} by replacing summation to maximum operation. This is why taking maximum operation instead of summation allows us to easily obtain one single value without consideration of constraint condition such as $y_i=A_i\oplus X$ at the output node. This variant has reduction of computational complexity, which results in lower performance than the conventional belief propagation algorithm. Despite this weakness, when using the algorithm in practice, a reduction in computational complexity is more significant.

Table 1. Maximum Product Detection Algorithm.

Algorithm 1: Maximum Product Detection Algorithm
Input: Priori probability $\Pr(X)$ in (1) Group matrix A Result vector Y
Output: Estimated $\hat{X}$
Initialization: Set the probability distribution: $\Pr(X)$ Put the initial message: $\xi_{ji}(\theta)\leftarrow\Pr(x_j=\theta)$
while $Y=A\oplus\hat{X}$ or Maximum iterations **do** 1) Update the message δ_{ij}: $\delta_{ij}(\theta)\leftarrow\max\left(\prod_{j'\in\ell(i)\setminus\{j\}}\xi_{j'i}(\theta)\right)$ 2) Update the message ξ_{ji}: $\xi_{ji}(\theta)\leftarrow\lambda\Pr(x_j=\theta)\prod_{i'\in\Im(j)\setminus\{i\}}\delta_{i'j}(\theta)$ 3) Tentative decoding: $\Pr(\hat{x}_j=\theta)\leftarrow\Pr(x_j=\theta)\prod_{j\in\ell(i)}\delta_{ij}(\theta)$
return Estimated $\hat{X}$

To handle the downward message δ_{ij}, we use the following equation instead of (6)

$$\delta_{ij}(\theta)=\max\left\{\prod_{j'\in T(i)\setminus\{j\}}\xi_{j'i}(\theta)\right\} \qquad (7)$$

where constraint condition is satisfied as $y_i=A_i\oplus X$.

And each upward message ξ_{ji} can be written by

$$\xi_{ji}(\theta)=\lambda\Pr(x_j=\theta)\prod_{i'\in\Im(j)\setminus\{i\}}\delta_{i'j}(\theta) \qquad (8)$$

where the variable λ uses for normalization of the total probability. Let $\Pr(\hat{x}_j=\theta):=\Pr(x_j|Y)$

be the posteriori probability for a sample x_j. We finally determine a maximum probability for $\theta=0$ or 1 as defined in (5),

$$\Pr(\hat{x_j}=\theta)=\Pr(x_j=\theta)\prod_{j\in\ell(i)}\delta_{ij}(\theta) \quad (9)$$

Using (7) and (8), the proposed MPDA iteratively updates the messages ξ_{ji} and δ_{ij}, and check if the constraint condition is satisfied as $Y=A\oplus\hat{X}$. That is, during each iteration, the MPDA stops if the condition is met. Otherwise, the MPDA continues to the maximum iterations as set in advance. Table 1 shows a version of pseudo code for our proposed MPDA.

4.2 Simulation Results

In this section, we evaluate the performance of the MPDA for the GT frameworks. To this end, we set the simulation environment as follows: the defective samples are generated from the probability distribution (1) with $\epsilon=0.02$, and the group matrix comes from the low-density parity-check [8] with 5 constant weights. As shown in Table 1, we evaluate the MPDA with the number of iterations, i.e., 20 and 50. In this paper, the length of X is 500, $N=500$, and the simulation is performed. We evaluate the probability of failure for the detection performance of the GT frameworks.

Figure 2 shows the simulation results when there are 10 defective samples ($D=10$) out of 500 samples ($N=500$). As shown in Figure 2, the probability of failure is obtained according to the number of tests T in the GT frameworks. We see that the more the number of tests T, the lower the probability of failure, but, on the other hands, the probability of failure increases. In addition, in case of $N=500$ and $D=10$, the lower bound on the number of tests based on information-theoretic approach in [9] is 70. It is shown that the gap of the performance between our proposed MPDA and the belief propagation algorithm is very small. And Table 2 shows the runtime for both algorithms to compare the computational complexity when the number of maximum iterations is 20 and 50, respectively. The MPDA provides reduction of runtime up to 12% compared to the belief propagation algorithm.

Table 2. Runtime (sec) of the belief propagation algorithm and our proposed MPDA.

	No. Iteration	Belief Prop. Algorithm [8]	MPDA
$N=500$ $D=10$	20	21.3	18.9
	50	53.8	48.1
$N=1000$ $D=100$	20	48.5	43.4
	50	123.3	108.5

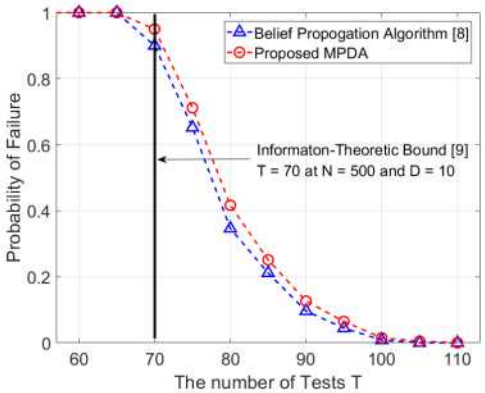

Fig. 2. Simulation results of the probability of failure with $N=500$ and $D=10$ compared to other results in [8] for the belief propagation algorithm and the bound [9] for the information-theoretic result.

5. Conclusion

In this paper, we consider the GT framework and also propose the MPDA for finding defective samples out of a set of large samples. The proposed detection algorithm is based on

maximum a posteriori probability, and is performed so that the posteriori probability of the output signal is maximized by exchanging marginal probabilities between samples and output results. As a result, we showed that there is slightly small gap between the conventional belief propagation algorithm and the our proposed MPDA. However, it provides us to reduce the computational complexity up to 12% in the aspect of runtime of both algorithms.

REFERENCES

[1] R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436-440, Dec. 1943.

[2] D.-Z. Du and F.K. Hwang, Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing, World Scientific, 2006.

[3] C.L. Chan, P.H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: near-optimal bounds with efficient algorithms," *49th Annual Allerton Conference on Communication, Control, and Computing,* pp. 1832-1839. Sep. 2011.

[4] M. Aldridge, L. Baldassini, and O. Johnson, "Group Testing Algorithms: Bounds and Simulations," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3671-3687, Jun. 2014.

[5] L. Baldassini, O. Johnson, M. Aldridge, "The capacity of adaptive group testing," *IEEE International Symposium on Information Theory*, pp. 2676-2680, Oct. 2013.

[6] G.K. Atia and V. Saligrama, "Boolean Compressed Sensing and Noisy Group Testing," *IEEE Transactions on Information Theory,* vol. 58, no. 3, pp. 1880-1901. Mar. 2012.

[7] J. Scarlett, "Noisy Adaptive Group Testing: Bounds and Algorithms," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3646-3661. Jun. 2019.

[8] M.C. Davey and D. Mackey, "Low-density parity-check codes over GF(q)," *IEEE Communications Letters*, vol. 2, no. 6, pp. 165-167, Jun. 1998.

[9] J.-T. Seong, "A New Upper Bound for Finding Defective Samples in Group Testing," *IEICE Transactions on Information and Systems*, advance publication, Feb. 2020.

Author Biography

Jin-Taek Seong **[Member]**

- Aug. 2014: Dep. Information & Communication Eng., GIST (Ph.D.)
- Mar. 2008 ~ Dec. 2010: LG Elec., Junior Researcher
- Sep. 2014 ~ Sep. 2016: DGMIF, Researcher
- Sep. 2016 ~ Mar. 2017: DAPA, Program Manager
- Mar. 2018 ~ Current: Dep. Convergence Software, Mokpo National University, Assistant Professor

〈Research Interests〉 Information Theory, Machine Learning, Communication Theory