A Novel Approach to Improve Quality Control by Comparing the Tagged Sequences of Product Traceability

Jing-Doo Wang^{1,*}

¹ Dept. of Computer Science and Information Engineering, Asia University, Taiwan.

Abstract. Quality control is an essential issue for manufacture, especially when the manufacture is towards intelligent manufacturing that is associated with "Internet of thing"(IOT) and "Artificial Intelligence"(AI) to speed up the rate of product line automatically nowadays. To monitor product quality automatically, it is necessary to collect and monitor the data generated by sensors, or to record parameters by machine operators, or to save the types (brands) of materials used when producing products. In this study, it is assumed that the sequences of the traceability of unqualified products are different from that of qualified ones, and these different values (or points) within the sequences result in these products qualified or unqualified. This approach extracts maximal repeats from the tagged sequences of product traceability, and meanwhile computes the class frequency distribution of these repeats, where the classes, e.g. "qualified" or "unqualified", are derived from the tags. Instead of inspecting all of the sequences of product traceability aimlessly, quality control engineers can filter out those maximal repeats whose frequency distributions are unique to specific classes and then just check the corresponding processes of these repeats. However, from the practical point of view, it should be estimated as a big-data problem to extract these maximal repeats and meanwhile compute their corresponding class frequency distribution from a huge amount of tagged sequential data. To have this work practical, this study uses one previous work that is based on Hadoop MapReduce programming model. and has been applied for an U.S.A patent (US Patent App. 15/208,994). Therefore, it is expected to be able to handle a huge amount of sequences of product traceability. With this approach that can narrow down the range for identifying false points (processes) within product line, it is expected to improve quality control by comparing tagged sequences of product traceability in the future.

1 Introduction

To speed up the processes of producing products and to reduce the cost of human labors, nowadays manufactories are toward to industry 4.0 [1] that adopt robots and combine with internet of things (IOT) [2]. It is predictable that the increasing rate of producing products in industry 4.0 will overwhelm the ability of quality control engineers to handle the problem in time when defective products occur and they need to identify what factors result in that products. To promote the ability of quality control [3] to meet the requirement of industry 4.0, it is necessary and essential to set up sensors

* e-mail: jdwang@asia.edu.tw

Figure 1. The traceability of one product attached with a tag.

to collect automatically the parameters from both environment and machines. That is, quality control (assurance) engineers can use these data to trace back the producing processes of those defective products and to identify false steps within the product line if abnormal products are detected or reported. Due to the increasing rate of producing products using robots, it is beyond the ability of quality control engineers to identify the false processes that result in these abnormal (false) products efficiently without extra analysis tools. It is desired to have novel approach to overcome above big data computational problem [4] to promote the power of quality control of industry in the future.

To improve quality control for reducing the response time of false points identification within the product line, in this study, the sequences of the traceability of products are collected first of all. The traceability of one product, as shown in Fig.1, is a ordered list (sequence) of elements or records that consist of data generated by sensors, or parameters of machines, or the types of materials collected sequentially when that product is produced. In this study it is assumed that the tag (class) of one product, given by quality control, is certainly concerned with its traceability. In other words, all of the sequences of traceability of those false or abnormal products with the same tag are supposed to contain some specific subsequences, called "markers", that result in of those products. Intuitively, the sequence of product traceability of one product determines the outcome of one product; the makers, some subsequences hidden within its traceability, are the keys to decide its tag that is the same as the quality control labels with. One class marker, therefore, is expected to be highly correlated with its own class (tag) such that it can provide clues to quality control engineers to inspect and identify false points from the subsequence representing that markers, instead of scanning the whole sequence of its traceability.

The goal of this paper is to extract class markers by comparing the class frequency distribution of maximal repeats extracted from the tagged sequences of product traceability via the scalable approach in the previous work[5]. The approach proposed in this study is to reduce the range of searching factors that result in unqualified products by comparing the tagged sequences of product traceability, where the traceability of one product is the sequential data generated while that product is produced, e.g. values collected by sensors, machine parameters, materials or components used or the identifier of operators; the tags are labels given by quality control staff. It is believed that these class markers, if they can be extracted efficiently, are valuable to quality control engineers because they can focus on inspecting those steps that are corresponding to these markers within the sequences of product traceability to speed up the processes of locating the false points within product line, instead of surveying the whole steps aimlessly.

The remainder of this paper is organized in the following. Section 2 describes how to generate the tagged sequences for experiments and the method to extract class markers. Section 3 shows experimental results. Section 4 has discussions and Section 5 gives conclusions and future works.

2 Materials and Methods

This study assumes that each of products is labeled with a tag (ClassID), e.g. one of these five classes as "normal", "abnormal I", "abnormal II", "abnormal III" or "abnormal IV" by quality control (assurance) engineers, and there is a sequence of traceability generated and collected when that product is produced. Most important of all, for simplicity, one "abnormal" tagged sequence of product traceability is expected to contain one specific subsequence, called "marker" in this study, that result in the same class labeling as quality control engineers do. That is, whenever one product is generated and its traceability sequence contain one of these markers, then its class labeling is determined according to which one of these markers. In this following, section 2.1 shows how to generate tagged sequences for experiments and section 2.2 describes the approach to extract maximal repeat briefly.

2.1 Tagged Sequences of Product traceability

To have the sequences of product traceability as experimental resource for demonstration in this study, there are 50 sequences of product traceability generated randomly and then each of these sequences is inserted with one class marker manually. Table 1 shows, for example, 50 sequences ($N=50$) of product traceability containing 20 steps ($k=20$) in which each of steps can be one of five symbols as "a", "b", "c", "d" and "e". For example, the value of the 2nd step in the sequence of traceability for first product "P1" is "S2,b". Each of these sequences is associated with one "ClassID" of "C1", "C2", "C3" and "C4", and is hidden with one marker, as shown in Table 2. For example, the "C3" contains 10 products and all of these products are hidden with the marker "S2,b S3,a S4,a". Note that these markers are supposed to be unknown and hidden within the sequences of product traceability in advance and the lengths of markers are not fixed but variable.

2.2 Extracting Class Markers from Tagged Sequences

A maximal repeat[8] is estimated to be a marker of one class if that repeat is with highly biased class frequency distribution, e.g. appearing in only that class and most of the sequences belonging to that class contain that marker. In other words, a marker of one specific class is a representative pattern for that class. In this study, the sequences of product traceability for experiments are generated randomly, and the markers are inserted into these sequences manually according to their corresponding tags individually.

To extract these distinctive markers hidden in classes that are derived from the tags, this study adopts a scalable approach developed in the previous work[5] that based on Hadoop MapReduce programming model [6, 7]. That approach has been proved to be capable to handle a huge amount of maximal repeat extraction from sequential data[9, 10] and is applying for a patent [11]. To have discussion systematical, there are definitions given in the section 2.2.1 and the approach of maximal repeat extraction is described briefly in the section 2.2.2.

2.2.1 Definition: Class Marker

In this study, products with their sequences of traceability are divided into k classes C_i , $1 \leq i \leq k$, where the types of classes are derived from the tags according to what kind of frequency distribution of subsequences is expected to be biased among these classes. Let the length of the sequence of one product traceability is " k ". Let $CF(MR_j)$ be the number of classes that contain the maximal repeat MR_j . The number of product traceability in class C_i is defined as n_{C_i} and the total number of product traceability is $N = \sum_{i=1}^{i=k} n_{C_i}$. Let $df(MR_j, C_i)$ be the number of product traceability of class C_i that

	ClassID	ProductID	Product Traceability																			
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	C3	P1	S1,e	S2,b	S3,a	S4,a	S5,b	S6,b	S7,d	S8,d	S9,d	S10,a	S11,c	S12,b	S13,a	S14,d	S15,e	S16,a	S17,d	S18,e	S19,c	S20,a
2	C1	P2	S1,a	S2,e	S3,d	S4,a	S5,b	S6,d	S7,e	S8,c	S9,e	S10,b	S11,c	S12,c	S13,c	S14,c	S15,a	S16,c	S17,a	S18,b	S19,c	S20,b
3	C4	P3	S1,a	S2,c	S3,b	S4,a	S5,a	S6,b	S7,e	S8,c	S9,d	S10,d	S11,b	S12,d	S13,b	S14,c	S15,e	S16,c	S17,a	S18,e	S19,d	S20,c
4	C2	P4	S1,d	S2,d	S3,e	S4,a	S5,b	S6,a	S7,a	S8,d	S9,d	S10,b	S11,c	S12,c	S13,d	S14,e	S15,b	S16,b	S17,a	S18,e	S19,e	S20,c
5	C1	P5	S1,a	S2,e	S3,d	S4,a	S5,c	S6,d	S7,c	S8,e	S9,d	S10,d	S11,d	S12,c	S13,c	S14,d	S15,e	S16,b	S17,a	S18,b	S19,b	S20,c
6	C2	P6	S1,c	S2,d	S3,c	S4,c	S5,c	S6,e	S7,e	S8,a	S9,c	S10,d	S11,e	S12,d	S13,e	S14,e	S15,b	S16,b	S17,a	S18,b	S19,d	S20,a
7	C2	P7	S1,a	S2,d	S3,b	S4,e	S5,c	S6,a	S7,a	S8,d	S9,d	S10,a	S11,a	S12,d	S13,d	S14,e	S15,b	S16,b	S17,a	S18,a	S19,d	S20,c
8	C4	P8	S1,b	S2,e	S3,b	S4,c	S5,d	S6,e	S7,a	S8,a	S9,e	S10,d	S11,b	S12,d	S13,b	S14,c	S15,e	S16,c	S17,b	S18,c	S19,b	S20,a
9	C2	P9	S1,a	S2,a	S3,a	S4,c	S5,e	S6,b	S7,d	S8,c	S9,a	S10,d	S11,a	S12,d	S13,b	S14,e	S15,b	S16,b	S17,a	S18,a	S19,a	S20,d
10	C3	P10	S1,d	S2,b	S3,a	S4,a	S5,a	S6,c	S7,d	S8,a	S9,a	S10,d	S11,e	S12,e	S13,b	S14,c	S15,d	S16,b	S17,b	S18,b	S19,c	S20,a
11	C4	P11	S1,d	S2,a	S3,d	S4,a	S5,b	S6,b	S7,a	S8,e	S9,c	S10,a	S11,d	S12,d	S13,b	S14,c	S15,e	S16,c	S17,c	S18,c	S19,c	S20,c
12	C2	P12	S1,a	S2,e	S3,c	S4,e	S5,a	S6,c	S7,a	S8,b	S9,b	S10,b	S11,b	S12,c	S13,c	S14,e	S15,b	S16,b	S17,a	S18,e	S19,a	S20,a
13	C1	P13	S1,a	S2,e	S3,d	S4,a	S5,b	S6,a	S7,b	S8,d	S9,a	S10,e	S11,d	S12,d	S13,d	S14,c	S15,b	S16,b	S17,a	S18,e	S19,c	S20,e
14	C1	P14	S1,a	S2,e	S3,d	S4,a	S5,a	S6,e	S7,a	S8,a	S9,c	S10,b	S11,c	S12,a	S13,a	S14,e	S15,a	S16,c	S17,b	S18,d	S19,d	S20,d
15	C3	P15	S1,e	S2,b	S3,a	S4,a	S5,c	S6,c	S7,a	S8,b	S9,c	S10,d	S11,c	S12,c	S13,e	S14,d	S15,d	S16,b	S17,b	S18,c	S19,d	S20,c
16	C1	P16	S1,a	S2,e	S3,d	S4,a	S5,e	S6,e	S7,e	S8,d	S9,b	S10,d	S11,b	S12,a	S13,c	S14,d	S15,d	S16,b	S17,d	S18,c	S19,d	S20,c
17	C2	P17	S1,a	S2,d	S3,c	S4,d	S5,c	S6,d	S7,b	S8,c	S9,d	S10,a	S11,e	S12,d	S13,e	S14,e	S15,b	S16,b	S17,a	S18,b	S19,d	S20,b
18	C1	P18	S1,a	S2,e	S3,d	S4,a	S5,c	S6,e	S7,a	S8,c	S9,b	S10,e	S11,b	S12,c	S13,d	S14,c	S15,e	S16,b	S17,b	S18,d	S19,e	S20,a
19	C4	P19	S1,c	S2,b	S3,b	S4,b	S5,d	S6,c	S7,a	S8,c	S9,e	S10,d	S11,d	S12,d	S13,b	S14,c	S15,e	S16,c	S17,a	S18,b	S19,c	S20,b
20	C1	P20	S1,a	S2,e	S3,d	S4,a	S5,a	S6,c	S7,a	S8,c	S9,d	S10,a	S11,e	S12,d	S13,b	S14,e	S15,b	S16,e	S17,e	S18,c	S19,d	S20,d
21	C4	P21	S1,a	S2,e	S3,b	S4,b	S5,e	S6,a	S7,b	S8,b	S9,d	S10,e	S11,b	S12,d	S13,b	S14,c	S15,e	S16,c	S17,b	S18,e	S19,b	S20,d
22	C3	P22	S1,c	S2,b	S3,a	S4,a	S5,b	S6,c	S7,c	S8,c	S9,c	S10,c	S11,b	S12,e	S13,a	S14,e	S15,a	S16,e	S17,c	S18,a	S19,e	S20,c
23	C4	P23	S1,a	S2,b	S3,e	S4,e	S5,e	S6,e	S7,e	S8,e	S9,b	S10,d	S11,c	S12,d	S13,b	S14,c	S15,e	S16,c	S17,a	S18,a	S19,c	S20,a
24	C4	P24	S1,e	S2,e	S3,b	S4,c	S5,b	S6,b	S7,e	S8,c	S9,a	S10,e	S11,c	S12,d	S13,b	S14,c	S15,e	S16,c	S17,c	S18,d	S19,a	S20,e
25	C1	P25	S1,a	S2,e	S3,d	S4,a	S5,e	S6,e	S7,c	S8,c	S9,a	S10,a	S11,b	S12,b	S13,a	S14,b	S15,d	S16,d	S17,a	S18,c	S19,c	S20,e
26	C1	P26	S1,a	S2,e	S3,d	S4,a	S5,b	S6,a	S7,d	S8,e	S9,d	S10,e	S11,d	S12,e	S13,a	S14,a	S15,b	S16,e	S17,b	S18,b	S19,d	S20,d
27	C2	P27	S1,e	S2,a	S3,e	S4,a	S5,b	S6,b	S7,d	S8,e	S9,e	S10,a	S11,b	S12,d	S13,c	S14,e	S15,b	S16,b	S17,a	S18,b	S19,e	S20,b
28	C4	P28	S1,b	S2,c	S3,e	S4,e	S5,a	S6,b	S7,e	S8,b	S9,a	S10,e	S11,e	S12,d	S13,b	S14,c	S15,e	S16,c	S17,e	S18,a	S19,d	S20,d
29	C2	P29	S1,a	S2,a	S3,d	S4,d	S5,b	S6,b	S7,d	S8,d	S9,c	S10,d	S11,a	S12,d	S13,a	S14,e	S15,b	S16,b	S17,a	S18,a	S19,d	S20,c
30	C4	P30	S1,d	S2,a	S3,a	S4,e	S5,e	S6,e	S7,a	S8,c	S9,b	S10,e	S11,c	S12,d	S13,b	S14,c	S15,e	S16,c	S17,a	S18,c	S19,e	S20,c
31	C4	P31	S1,c	S2,c	S3,d	S4,a	S5,b	S6,d	S7,a	S8,a	S9,a	S10,d	S11,b	S12,d	S13,b	S14,c	S15,e	S16,c	S17,d	S18,a	S19,b	S20,b
32	C1	P32	S1,a	S2,e	S3,d	S4,a	S5,e	S6,a	S7,d	S8,a	S9,d	S10,d	S11,a	S12,d	S13,b	S14,a	S15,b	S16,c	S17,d	S18,a	S19,c	S20,c
33	C4	P33	S1,c	S2,c	S3,a	S4,c	S5,b	S6,c	S7,a	S8,c	S9,d	S10,e	S11,a	S12,d	S13,b	S14,c	S15,e	S16,c	S17,c	S18,d	S19,b	S20,c
34	C3	P34	S1,e	S2,b	S3,a	S4,a	S5,c	S6,a	S7,d	S8,d	S9,d	S10,e	S11,a	S12,e	S13,d	S14,b	S15,d	S16,a	S17,b	S18,e	S19,e	S20,b
35	C4	P35	S1,d	S2,e	S3,b	S4,d	S5,d	S6,e	S7,e	S8,a	S9,e	S10,c	S11,e	S12,d	S13,b	S14,c	S15,e	S16,c	S17,d	S18,d	S19,a	S20,d
36	C2	P36	S1,b	S2,d	S3,b	S4,e	S5,d	S6,a	S7,b	S8,a	S9,c	S10,e	S11,a	S12,a	S13,a	S14,e	S15,b	S16,b	S17,a	S18,b	S19,a	S20,c
37	C3	P37	S1,b	S2,b	S3,a	S4,a	S5,e	S6,c	S7,b	S8,e	S9,a	S10,d	S11,e	S12,c	S13,b	S14,b	S15,a	S16,d	S17,e	S18,c	S19,d	S20,c
38	C2	P38	S1,c	S2,d	S3,c	S4,a	S5,b	S6,c	S7,c	S8,e	S9,d	S10,c	S11,a	S12,e	S13,e	S14,e	S15,b	S16,b	S17,a	S18,c	S19,c	S20,b
39	C4	P39	S1,e	S2,b	S3,c	S4,d	S5,c	S6,c	S7,a	S8,e	S9,e	S10,c	S11,d	S12,d	S13,b	S14,c	S15,e	S16,c	S17,c	S18,c	S19,a	S20,a
40	C1	P40	S1,a	S2,e	S3,d	S4,a	S5,d	S6,d	S7,d	S8,c	S9,c	S10,a	S11,d	S12,e	S13,a	S14,e	S15,d	S16,c	S17,d	S18,a	S19,e	S20,c
41	C3	P41	S1,a	S2,b	S3,a	S4,a	S5,d	S6,a	S7,e	S8,a	S9,b	S10,b	S11,a	S12,e	S13,a	S14,d	S15,b	S16,e	S17,e	S18,e	S19,d	S20,b
42	C3	P42	S1,c	S2,b	S3,a	S4,a	S5,d	S6,a	S7,c	S8,a	S9,e	S10,a	S11,e	S12,d	S13,c	S14,b	S15,d	S16,e	S17,c	S18,d	S19,c	S20,d
43	C3	P43	S1,c	S2,b	S3,a	S4,a	S5,b	S6,a	S7,d	S8,b	S9,b	S10,a	S11,d	S12,a	S13,d	S14,a	S15,e	S16,c	S17,b	S18,d	S19,b	S20,a
44	C4	P44	S1,b	S2,e	S3,e	S4,c	S5,e	S6,d	S7,d	S8,e	S9,c	S10,e	S11,a	S12,d	S13,b	S14,c	S15,e	S16,c	S17,a	S18,b	S19,a	S20,b
45	C2	P45	S1,a	S2,b	S3,b	S4,e	S5,e	S6,b	S7,b	S8,b	S9,d	S10,a	S11,e	S12,d	S13,a	S14,e	S15,b	S16,b	S17,a	S18,a	S19,c	S20,e
46	C4	P46	S1,e	S2,b	S3,b	S4,c	S5,a	S6,a	S7,b	S8,b	S9,e	S10,e	S11,b	S12,d	S13,b	S14,c	S15,e	S16,c	S17,c	S18,e	S19,c	S20,c
47	C2	P47	S1,b	S2,d	S3,d	S4,a	S5,c	S6,c	S7,d	S8,d	S9,a	S10,a	S11,c	S12,e	S13,e	S14,e	S15,b	S16,b	S17,a	S18,d	S19,e	S20,e
48	C1	P48	S1,a	S2,e	S3,d	S4,a	S5,b	S6,e	S7,d	S8,b	S9,c	S10,c	S11,e	S12,e	S13,e	S14,d	S15,b	S16,c	S17,d	S18,c	S19,a	S20,b
49	C2	P49	S1,c	S2,e	S3,a	S4,b	S5,b	S6,a	S7,e	S8,d	S9,c	S10,e	S11,e	S12,c	S13,a	S14,e	S15,b	S16,b	S17,a	S18,c	S19,d	S20,b
50	C3	P50	S1,d	S2,b	S3,a	S4,a	S5,c	S6,a	S7,c	S8,b	S9,e	S10,b	S11,a	S12,d	S13,e	S14,a	S15,d	S16,d	S17,c	S18,c	S19,a	S20,e

Table 1. 50 products with their traceability and tags (ClassID)

ClassID	# of Product	Class Marker	Length
C1	12	S1,a S2,e S3,d S4,a	4
C2	13	S14,e S15,b S16,b S17,a	4
C3	10	S2,b S3,a S4,a	3
C4	15	S12,d S13,b S14,c S15,e S16,c	5
Total	50		

Table 2. The statistics of four classes with markers hidden in their corresponding traceability

contains the MR_j ; let $tf(MR_j, Ci)$ be the number of the MR_j appearing in the product traceability of class Ci . Note that it might happen that the MR_j appears more than twice in one product traceability if there exists repeat processes within the product line.

Let $WC(MR_j, Ci) = \frac{df(MR_j, Ci)}{n_{Ci}}$, the coverage of one maximal repeat MR_j within class Ci , be the percentage of the number of traceability within the Ci that contain the MR_j . Let $DF(MR_j) = \sum_{i=1}^{i=k} df(MR_j, Ci)$ and $TF(MR_j) = \sum_{i=1}^{i=k} tf(MR_j, Ci)$. Intuitively, the MR_j is representative for class Ci when the value of $CF(MR_j)$ is "1" and the value of $WC(MR_j, Ci)$ is as high as possible. That is, the MR_j appears only in one class Ci and the majority of product traceability in Ci contain the MR_j . In this study, for simplicity, one maximal repeat MR_j is estimated as one "marker" of class Ci if $CF(MR_j) = 1$ and $WC(MR_j, Ci) = 100\%$. That is, the MR_j appears only in class Ci and all of the sequences of traceability in Ci contains the MR_j .

2.2.2 Maximal Repeat Extraction

To identify these markers from a huge amount of the sequences of product traceability attached with ClassID, this paper adapts the previous work in [5] that is a scalable approach based on Hadoop MapReduce programming model [7]. This approach can extract the maximal repeats [8] from a huge amount of tagged sequential data, and meanwhile compute the class frequency distribution of these repeats within these sequential data, where the types of classes are derived from the tags specified by the users or observers for comparison on purpose. To mine for specific subsequences, called markers in this study, that can be used to distinguish one class from another, this study filter out the maximal repeats that only appear in one class and meanwhile happen in the majority of product traceability of that class.

To have this study more self-contained, the approach of maximal repeat extraction are described briefly in the following. Based on Hadoop MapReduce programming, the input of tagged sequences are split into several fixed-size partitions and are distributed to mappers by Hadoop system evenly and automatically. First of all, each of mapper generated all of suffixes of each sequences and have each of these suffixes attached the same tag as that sequence it derived from. Secondly, the mappers output the pair of "key-value" to the reducers, where the "key" is the fixed-length prefix of each suffix and the "value" is the tagged suffix. According to the feature of MapReduce programming, each of reducers will receives the values, some tagged suffixes, with the same key while these suffixes are sorted in lexicographical order. Finally, the reducers scan the sorted suffixes and extract candidate maximal repeats with "right" boundary validation; meanwhile compute the class frequency distribution of these repeats via a stack with push/pop operations[5]. To have maximal repeats with "left" boundary validation, it is similar with the same processes as described above, but with the reversed sequences as the input. The maximal repeats are those candidate maximal repeats pass both of "right" and "left" boundary validation.

3 Results

In this study, there are 50 sequences of product traceability generated and are tagged with specific markers manually for experimental resources. Table 3 gives the distribution of 413 maximal repeats extracted from these 50 tagged sequences in Table 1 according to the values of "CF" and the length of maximal repeats. It is observed that there are 99 maximal repeats whose "CF" values are 1 and the length of these repeats ranges from 1 to 8.

To further inspect the WC values of these 99 maximal repeats ("CF=1") for one class they appear, Table 3 sorts the WC values of and selects the top 18 maximal repeats whose WC values of one class

CF	The Length of Maximal Repeat	The Number of Distinct Maximal Repeat
1	1	3
1	2	33
1	3	14
1	4	8
1	5	17
1	6	13
1	7	9
1	8	2
2	1	16
2	2	127
2	3	39
2	4	9
2	5	3
3	1	33
3	2	31
3	3	4
4	1	46
4	2	6
Total		413

Table 3. The statistics of maximal repeats extracted from the tagged sequences in Table 1

	Maximal Repeat	DF	TF	CF	Length	Class Frequency Distribution (ClassID#df#tf)	Class Frequency Distribution (Coverage within one class)			
							C1	C2	C3	C4
1	S1,a S2,e S3,d S4,a	12	12	1	4	(C1#12#12)	100.0%			
2	S14,e S15,b S16,b S17,a	13	13	1	4	(C2#13#13)		100.0%		
3	S2,b S3,a S4,a	10	10	1	3	(C3#10#10)			100.0%	
4	S12,d S13,b S14,c S15,e S16,c	15	15	1	5	(C4#15#15)				100.0%
5	S2,d	7	7	1	1	(C2#7#7)		53.8%		
6	S1,a S2,e S3,d S4,a S5,b	4	4	1	5	(C1#4#4)	33.3%			
7	S11,b S12,d S13,b S14,c S15,e S16,c	5	5	1	6	(C4#5#5)				33.3%
8	S12,d S13,b S14,c S15,e S16,c S17,c	5	5	1	6	(C4#5#5)				33.3%
9	S13,a S14,e S15,b S16,b S17,a	4	4	1	5	(C2#4#4)		30.8%		
10	S13,e S14,e S15,b S16,b S17,a	4	4	1	5	(C2#4#4)		30.8%		
11	S14,e S15,b S16,b S17,a S18,a	4	4	1	5	(C2#4#4)		30.8%		
12	S1,c S2,b S3,a S4,a	3	3	1	4	(C3#3#3)			30.0%	
13	S1,e S2,b S3,a S4,a	3	3	1	4	(C3#3#3)			30.0%	
14	S2,b S3,a S4,a S5,b	3	3	1	4	(C3#3#3)			30.0%	
15	S2,b S3,a S4,a S5,c	3	3	1	4	(C3#3#3)			30.0%	
16	S2,e S3,b	4	4	1	2	(C4#4#4)				26.7%
17	S2,c	4	4	1	1	(C4#4#4)				26.7%
18	S1,a S2,e S3,d S4,a S5,e	3	3	1	5	(C1#3#3)	25.0%			

Table 4. Examples: the statistics of 18 maximal repeats whose CF values are equal to 1 and WC values of one class are $\geq 25\%$

are greater than or equal to 25%. It is observed that the four class markers as shown in Table 2 are extracted and are the same as the first four repeats whose WC values are 100%. That is, experimental results as described above shows that the previous approach[5] did extract and identify these class markers that are hidden within the sequences of product traceability.

4 Discussion

There are still room to make this study more robust and practical in the future. First of all, in this study, one class marker is defined as a consecutive (continuous) subsequences. Indeed, this assumption may be too restricted in real cases because the false points may be not consecutive to form only one complete segment but several short segments. It needs more efforts or another strategy to overcome this problem. On the other hand, the *WC* condition of one marker that appears in only one class in which all of sequences contain that marker can be relaxed by selecting the markers whose values of the entropy[13] of class frequency distribution is lower than a given threshold.

Secondly, the values generated in each of steps in the sequence of product traceability are set to be one of five symbols as "a", "b", "c", "d" and "e". When dealing with real practical data, these values may be numeric but not just several symbols. It is necessary to learn how to partition these numeric values into symbols precisely [14] because different transformation might have totally different results.

Thirdly, the experimental resource used in this study only contains 50 sequences of product traceability and the markers are created and inserted manually. From the practical point of view, it is expected to have these tagged sequences for experiments from the factory in real world. Furthermore, the approach proposed in this study do improve the quality control to reduce the searching ranges of the candidate false points that results in abnormal product in this future.

5 Conclusions

This study provides a novel approach to speed up the processes of locating the false points within the product line via comparing the tagged sequences of product traceability. These class markers extracted can provide clues to locate false point in the product line more efficiently and furthermore can be stored in database as false patterns for quality assurance analysis in the future. As the time of industry 4.0 combined with internet of things (IOT) and artificial intelligent (AI) is coming, it is worth of embedding this approach into quality control services as a package or tool such that monitoring the product line and searching the false points can be done automatically in the future. It is expected that this approach proposed in this study can provide modern industry with new research or direction about quality control (assurance), especially in drug and semiconductor manufacturing.

Acknowledgments

This study is supported by Ministry of Science and Technology, Taiwan under project *MOST 106-2632-E-468-002*.

References

- [1] A. Gilchrist, *Industry 4.0: The Industrial Internet of Things*, 1st edn. (Apress, Berkely, CA, USA, 2016), ISBN 1484220463, 9781484220467
- [2] P. Waher, *Learning Internet of Things*, Community experience distilled (Packt Publishing, 2015), ISBN 9781783553549, <https://books.google.com.tw/books?id=auRrBgAAQBAJ>
- [3] J. Jain, *Quality Control and Total Quality Management* (McGraw-Hill Education (India) Pvt Limited, 2001), ISBN 9780070402140, <https://books.google.com.tw/books?id=8Q3sRWAAP74C>
- [4] B. Marr, *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results* (Wiley, 2016), ISBN 9781119231387, <https://books.google.com.tw/books?id=wA0mCgAAQBAJ>

- [5] J.-D. Wang, *Extracting significant pattern histories from timestamped texts using mapreduce*, *The Journal of Supercomputing*, pp. 1–25, 2016. [Online]. Available: <http://dx.doi.org/10.1007/s11227-016-1713-z>
- [6] R. McCreadie, C. Macdonald, I. Ounis, *MapReduce indexing strategies: Studying scalability and efficiency* *Information Processing & Management* **48**, 873 (2012), large-Scale and Distributed Systems for Information Retrieval
- [7] F. Li, B.C. Ooi, M.T. Özsu, S. Wu, *Distributed Data Management Using MapReduce* *ACM Comput. Surv.* **46**, 31:1 (2014)
- [8] D. Gusfield, *Algorithms on Strings, Trees, and Sequences : computer science and computational biology* (Cambridge University Press, 1997), ISBN 0-521-58519-8
- [9] J.D. Wang, M.C. Hwang, *Extracting significant pattern histories from timestamped texts using mapreduce*, *Applied Sciences*, **7** (2017). [Online]. Available: <http://www.mdpi.com/2076-3417/7/9/878>
- [10] J.D. Wang, Y.C. Wang, R.M. Hu, J. Tsai, *Extracting the Co-occurrences of DNA Maximal Repeats in both Human and Viruses*, in *The 17th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE2017)* (2017)
- [11] C.T. Wang, *Method for extracting maximal repeat patterns and computing frequency distribution tables* (2017), US Patent App. 15/208,994, <https://www.google.com/patents/US20170255634>
- [12] J.D. Wang, W.L. Chan, C.C. Wang, J.G. Chang, J.J. Tsai, *Mining Distinctive DNA Patterns from the Upstream of Human Coding&Non-Coding Genes via Class Frequency Distribution*, in *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2016)* (2016)
- [13] E. Alpaydin, *Introduction to Machine Learning* (The MIT Press, 2004)
- [14] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)* (Elsevier, 2011), ISBN 0120884070