International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:12, 2012

Clustering Mixed Data Using Non-normal Regression Tree for Process Monitoring

Youngji Yoo, Cheong-Sool Park, Jun Seok Kim, Young-Hak Lee, Sung-Shick Kim, Jun-Geol Baek

Abstract—In the semiconductor manufacturing process, large amounts of data are collected from various sensors of multiple facilities. The collected data from sensors have several different characteristics due to variables such as types of products, former processes and recipes. In general, Statistical Quality Control (SQC) methods assume the normality of the data to detect out-of-control states of processes. Although the collected data have different characteristics, using the data as inputs of SQC will increase variations of data, require wide control limits, and decrease performance to detect out-of-control. Therefore, it is necessary to separate similar data groups from mixed data for more accurate process control. In the paper, we propose a regression tree using split algorithm based on Pearson distribution to handle non-normal distribution in parametric method. The regression tree finds similar properties of data from different variables. The experiments using real semiconductor manufacturing process data show improved performance in fault detecting ability.

Keywords—Semiconductor, non-normal mixed process data, clustering, Statistical Quality Control (SQC), regression tree, Pearson distribution system.

I. Introduction

A Semiconductor manufacturing industry is the most automated and sophisticated technologies industry. Semiconductors are fabricated by similar and complex chemical and physical processes on the surface of wafer repeatedly. There are hundreds of process steps and thousands of facilities in semiconductor manufacturing process [1]. With the rapid development of data collection and storage technologies, large amount of data gathered in the database from hundreds of sensors attached to each facility in real-time. During the manufacturing process, the data are collected from sensors automatically in the form of mixed by different variables such as types of products, former processes, and recipes.

One of main concern in the fab is process monitoring and Statistical Quality Control (SQC) using gathered data. The objective of SQC is defects reduction of the products and maintenance of in control production systems [2]. Due to the complexity of mixed data, most of engineers rely on their own knowledge and experience to identify characteristics of the abnormal product to determine approximate specifications. However, it is hard to analyze numerous data rely on personal knowledge and experience [3]. Therefore, we can use data mining techniques in order to separate mixed data. In general, the data mining techniques are used for extracting meaningful information from large data. When the collected data is used as input variable of SQC, wide control limit is required and decreases performance to detect out-of-control because the mixed data has large variation. Therefore, it is necessary to separate mixed data into similar groups to process control more accurately.

The general purpose of data mining is to find useful information or meaningful patterns from data using multiple hypothesis testing [4]. In particular, decision tree is one of the popular data mining tools. The decision tree classifies data into several subgroups by successive hypothesis testing and predicts results of new input data [3]. Advantages of decision tree are easy to interpret analysis results and to identify the important variables affecting the results and available to make the decision directly. The decision tree performs an iterative binary partitioning that splits a decision node into two sub-division nodes. There are two types of the decision tree, classification tree and regression tree. The classification tree splits node based on input variable when output variable is categorical. If the output variable is continuous, the regression tree is used. In the paper, we use the regression tree to separate semiconductor manufacturing process data into similar groups. Outputs of the regression tree are continuous values of responses which are measured from sensors. Inputs are categorical values such as Process ID, Chamber Step, and Step Sequence, etc. which have explainable characteristics of process.

There have been many studies about splitting criterion in the decision tree such as ID3 [5], C4.5 [6], CART [7], CHAID [8], and AID algorithm [9]. Especially, the AID algorithm based on the theory of ANOVA(Analysis Of Variance) as rules of multiple hypothesis tests is one of the most commonly used. To select splitting criterion, the AID algorithm compares all possible combinations of subset in each decision node and finds point that minimize the sum of squared errors.

To apply ANOVA concept, which is used in the AID algorithm, there are three necessary conditions of normality, independence, and homogeneity of variances on data [10]. However, in reality, distributions estimated from most of semiconductor data do not satisfy normality. The distribution has kurtosis and skewness because many distributions are aggregated by different recipes of processes. Also, the independence and homogeneity of variances are not fulfilled because the various recipes have interaction and different variance at each other.

Youngji Yoo, Cheong-Sool Park, and Jun Seok Kim are with the School of Industrial Management Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul, 136-713, Republic of Korea

Young-Hak Lee is with Device Solution Division, Samsung Electronics Co., Ltd., San 16 Banwol-Dong, Hwasung-City, Gyeonggi-Do, 445-701, Republic of Korea

Sung-Shick Kim is a professor emeritus in the School of Industrial Management Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul, 136-713, Republic of Korea

Jun-Geol Baek is an associate professor in the School of Industrial Management Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul, 136-713, Republic of Korea (Corresponding author to provide phone: +82-2-3290-3396; fax: +82-2-929-5888; e-mail: jungeol@korea.ac.kr)

Therefore, the semiconductor manufacturing process data is non-normal and does not satisfy the three necessary conditions to use AID algorithm.

In the study, we propose non-normal regression tree that can apply to the non-normal data. The proposed non-normal regression tree performs hypothesis tests as using split criterion based on Pearson distribution that handle various types of non-normal distributions in parametric method. Generally, the normal distribution has two parameters, mean and variance. But non-normal distributions are handled as non-parametric methods in hypothesis tests such as Mann-Whitney U test [11] and Wilcoxon Rank-sum test [12]. The Pearson distribution system can explain data using four parameters which are first to fourth momentum of moment generating functions of empirical data distribution. The parameters are mean, variance, skewness, and kurtosis. In the experiment, the semiconductor mixed data are separated into cluster by similar process characteristics using the non-normal regression tree. In addition, control chart configured by each cluster to improve performance in detection of out-of-control.

The paper is organized as follows. Section I describes research background, significance, and research aims of the study. Section II describes the fundamental of the research and proposes splitting algorithm based on Pearson distribution for non-normal regression tree. Section III validates the framework with an experiment using real semiconductor manufacturing process data. Section IV concludes with discussion and further study directions.

II. NON-NORMAL REGRESSION TREE USING PEARSON DISTRIBUTION

A. Pearson Distribution

A normal distribution is the most representative probability distribution in statistics. The normal distribution is often used in order to describe continuous random variables that cluster around a single mean value. By central limit theorem, when numerous data of random variables are drawn additively and independently from the same distribution, sample means of the data are distributed approximately normal distribution. The normal distribution is necessary condition in many statistical techniques because large size sampled data have normality by Central Limit Theorem irrespective of the form of the original distribution [13].

However, most of data obtained in reality do not follow the normal distribution for several reasons such as aggregation of data, lack of independence of variables, and interaction between variables. The non-normal data could not be explained by the normal distribution sufficiently. Therefore, many studies such as skew-normal probability distributions [14] and Burr system [15] have been carried out actively to represent non-normal distribution in parametric method.

Pearson distribution system which was introduced by Karl Pearson is one of the most powerful distribution systems which can describe non-normal distribution [16], [17], [18], and [19]. The Pearson distribution system have four parameters, mean, variance, skewness, and kurtosis. So, almost all continuous distributions such as Beta distribution, Chi-square distribution, and Normal distribution can be explained by Pearson distribution system. The Pearson system consists of sub-type distributions which is named as Pearson type I to VII.

Fig.1 is a diagram of existing region of Pearson type distributions using skewness and kurtosis. The x-axis is squared skewness $\beta_1 = \gamma_1^2$, where γ_1 is the skewness or third standardized moment(1). The y-axis is the traditional kurtosis $\beta_2 = \gamma_2 + 3$, where γ_2 is forth standardized moment(2). When standardize variables of probability distribution, relationship of four parameters is expressed as a function of skewness and kurtosis. Parameter κ, which is the function of skewness and kurtosis, is calculated from (3). The Pearson type of distribution is classified by the parameter κ [20].

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{1}$$

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \tag{2}$$

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)} \tag{3}$$

In the Fig.1, regions and lines denote Pearson type with various skewness and kurtosis. The normal distribution, which has skewness and kurtosis of 0 and 3 respectively, is Pearson type 0 and the type is marked by a dot in the Fig.1. If a point (β_1, β_2) is included in the yellow area, the distribution is Pearson type I. If a point is above the dividing line to the pink area and the yellow area, the distribution follows Pearson type III.

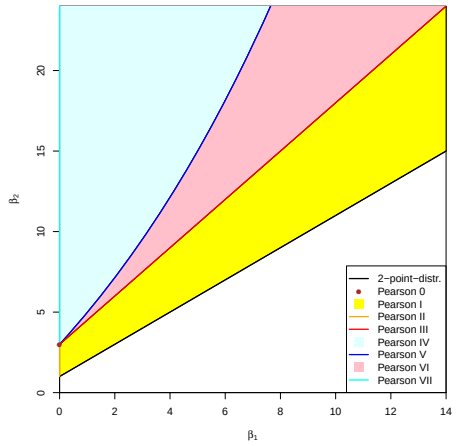

Fig. 1 Pearson Diagram

B. Regression Tree

Regression tree is used when predict continuous output variable based on several input variables. The regression tree separates whole data into several subgroups using recursive binary splitting algorithm. The separated data included in each node can be thought of as a cluster.

Automatic Interaction Detection (AID) algorithm [9] is used in the regression tree to binary split recursively. The AID algorithm can calculate Sum of Squares Between the groups (SSB) in every node by using (4). The SST is within-node

sum of squared errors at parent node. The sum of squared errors of left child node and right child node is SSL and SSR, respectively. Therefore, SSB is difference of sum of squared errors between parent node and two child nodes. In the AID algorithm, the point that maximizes SSB is chosen as the best split criterion.

$$SSB \approx SST - (SSL + SSR),$$
$$SST = \sum_i (Y_i - \bar{Y})^2 \tag{4}$$

However, it is hard to apply to non-normal data because the regression tree has three necessary conditions to use AID algorithm. The non-normal data do not satisfy necessary conditions. Therefore, we propose a new split algorithm based on Pearson distribution to make robust regression tree for non-normal data.

C. Pearson Split Algorithm

Given ν categorical dependent variables $X_1, X_2, \ldots, X_\nu$ and the continuous response variable Y, we can estimate Pearson distribution of Y in every levels of each X variables. The Pearson distribution is estimated using moments which consist of four parameters in (5). In the Pearson split algorithm, the Pearson distribution is used to find best split point that maximize heterogeneity of two subsets of data.

For non-normal regression tree, heterogeneity of subsets of data is measured by approximated p-value determined by (8). The value in (8) could be calculated using (6) and (7). The j in (8) indicates split point that divides into left node and right node in decision node. Y_L and Y_R are subsets of Y that separated into both nodes respectively. n_L and n_R are a number of data of Y_L and Y_R. The function f denotes probability density function of Pearson distribution. The Pearson distribution is estimated by four moments of Y_i which is subsets of Y that corresponding with the ith level of variable X. n_i reflects the size of distribution in the form of weighted sum. The Pearson distributions are estimated as much as a number of levels of X variables.

$$\mathbf{m} = [\mu \ \sigma \ \gamma_1 \ \gamma_2]^T \tag{5}$$

$$P_L = \sum_{i=1}^{j} n_i \int_{\mu_R}^{\infty} f(\mathbf{m}, Y_i) dx \tag{6}$$

$$P_R = \sum_{i=j+1}^{N} n_i \int_{\infty}^{\mu_L} f(\mathbf{m}, Y_i) dx \tag{7}$$

$$f_p(Y_L, Y_R, \mu_L, \mu_R, j) = \frac{P_L + P_R}{n_R + n_L} \tag{8}$$

Fig.2 shows two Pearson distributions of different mean. μ_L and μ_R are means of subsets Y_L and Y_R. The two subsets are separated by split point. The left colored region and right colored region are calculated by P_L and P_R in (6) and (7). The approximated p-value is sum of P_L and P_R. According to increase of distance of two distributions, approximated p-value becomes smaller. The small approximated p-value means that two subsets are heterogeneous. Therefore, the approximated p-value can be used as an indicator to select split point.

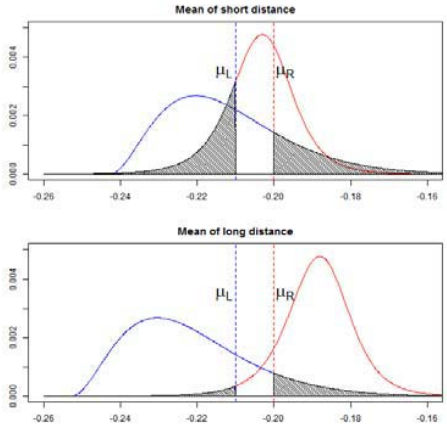

Fig. 2 Approximated p-value of different mean

The Pearson split algorithm is described in the form of Pseudocode in Algorithm 1. The basic idea is finding best split criterion that maximize heterogeneity between the two subgroups. We start at a node that we would like to split and initial approximated p-value is zero.

Algorithm 1 Pearson split algorithm

```
bestsplit = NULL
p = 0
for i ← 1, ν do
    κ = number of levels[X_i]
    for l ← 1, κ do
        μ_l = mean(Y_l)
    end for
    κ̃ = sortlevels(μ_l)
    n = number of κ̃
    j = 1
    while j < n do
        Y_L = Y[κ̃ ≤ j]
        Y_R = Y[κ̃ > j]
        μ_L = mean(Y_L)
        μ_R = mean(Y_R)
        p' = 1 − f(Y_L, Y_R, μ_L, μ_R, j)
        if p' > p then
            p ← p'
            bestsplit = currentsplit
        end if
        j ← j + 1
    end while
end for
return bestsplit
```

There are ν categorical input variables, and loop from X_1 to X_v to find the best split point. κ is a number of levels of ith input variable. We can obtain averages of corresponding

Y values for each levels and then sort the levels in ascending order of the averages. j, which denotes split point, is equal to number of levels allocated in the left node. If any input variable has N levels, Y variables corresponding to levels sorted from the 1st to the jth are assigned to the left node. Also Y variables corresponding to remaining $N - j$ levels are assigned to the right node. We can estimate Pearson distributions from subset Y of each level in both nodes. Using the Pearson distributions, splitting value p is calculated as subtracting approximated p-value from 1. The Pearson split algorithm find split point that maximize splitting value instead of minimizing approximated p-value for convenience of the calculation in the Regression tree. Looping through the possible binary splits, return the best split criterion maximize splitting value.

III. Experiments

The purpose of experiments is that gather semiconductor data by similar conditions using non-normal regression tree and design control chart for process monitoring. In order to evaluate the performance of proposed method, four encoded dataset is used. The dataset is collected chronologically from different facilities in real semiconductor manufacturing process during two month. Due to properties of the semiconductor manufacturing process, distributions of the datasets follow non-normal distribution. The datasets have five categorical input variables and one continuous output variable. The input variables denote process recipes and the output variable denotes responses measured from sensor.
Training set is 70% of data chronological order and remaining 30% of data is test set. Regression tree fits to training set using AID algorithm and Pearson split algorithm, respectively. The two trees are employed to make predictions on the test sets. The β value measures a performance of prediction from described in (9). The out-of-control indicates the abnormal data when control limits are considered.

$$\beta = \frac{\text{number of out-of-control}}{\text{number of test data}} \times 100 \qquad (9)$$

In order to configure control chart, the AID algorithm uses traditional control limits and the Pearson algorithm uses empirical control limits reflected non-normality. The α is fixed to 0.0027.
Table I shows result of the experiment. Footnote indicates the best result.

TABLE I
Results of Experiment

Dataset	β(%)		Improvement(%)
	AID	Proposed	
1	0.767*	0.842	-9.845
2	2.933	2.321*	20.867
3	3.242	2.901*	10.526
4	8.664	4.259*	50.848

According to the results of the above table, the proposed algorithm has lower beta value in dataset 2, 3, and 4. The table also represents improvement of prediction. For three datasets, the proposed algorithm improves predictive power more than 10% and the dataset 4 improves predictive power by 50.8% comparing with the regression tree using AID algorithm. The performance of the proposed method using the dataset 1 is reduced about 10% than AID algorithm. That's because distribution of the dataset 1 is similar with normal distribution and the AID algorithm is robust under normal conditions than proposed algorithm. Therefore, the proposed algorithm yields better performance than AID algorithm on the non-normal semiconductor data.
Fig.3 and Fig.4 show experiment result visually of Dataset 2. Fig.3 is regression tree using proposed algorithm. The regression tree has three terminal nodes and each node represents each cluster, which is separated by similar conditions. In the Fig.4, data of each cluster is distributed right skewed. Control limit has different lower and upper bound interval because the limit is considered empirically by non-normality of the data.

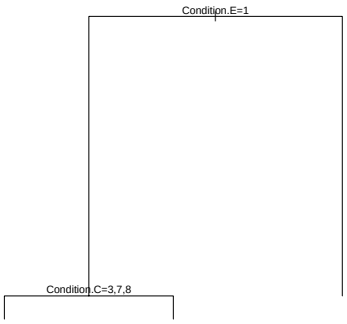

Fig. 3 Non-normal regression tree of Dataset 2

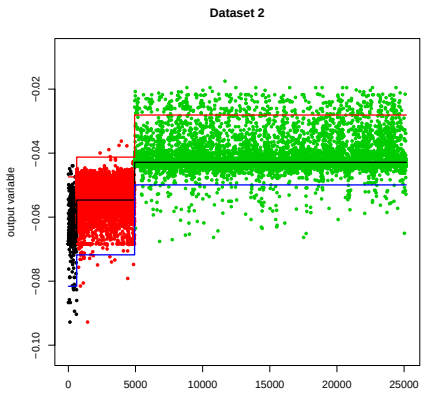

Fig. 4 Clusterd control chart of Dataset 2

IV. Conclusion

In the research, we propose a non-normal regression tree to cluster non-normal mixed data generated in semiconductor

manufacturing processes. The data collected in real semiconductor process does not follow normal distribution. Therefore, non-normal regression tree finds better split criterion than existing regression tree in non-normal mixed data. The proposed method reduced the number of out of control of test set when the data much more does not follow a normal distribution.

In the paper, sample size is weighted to each distribution in order to reflect distribution size. As the future study, it is needed to consider sample Pearson distribution. From sample Pearson distribution, we can determine robust non-normal regression tree and control limits for accurate semiconductor manufacturing process monitoring.

ACKNOWLEDGMENT

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the IT R&D Infrastructure Program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-(B1100-1101-0002)).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0008332)

REFERENCES

[1] H. Yoon, J.-G. Baek, C.-S. Park, and Y.-H. Lee, "A constrained clustering method for mixed process data using recursive partitioning and regression trees," in *Proceedings of the IIE Asian Conference*, 2012.

[2] S. Bagchi, R. J. Baseman, A. Davenport, R. Natarajan, N. Slonim, and S. Weiss, "Data analytics and stochastic modeling in a semiconductor fab," *Applied Stochastic Models in Business and Industry*, vol. 26, no. 1, pp. 1–27, 2010.

[3] C. F. Chien, W. C. Wang, and J. C. Cheng, "Data mining for yield enhancement in semiconductor manufacturing and an empirical study," *Expert Systems with Applications*, vol. 33, no. 1, pp. 192–198, 2007.

[4] D. J. Hand, "Principles of data mining," *Drug Safety*, vol. 30, no. 7, pp. 621–622, 2007.

[5] J. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, p. 81, 1986.

[6] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[7] L. Breiman, *Classification and regression trees*. Wadsworth International Group, 1984.

[8] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 29, no. 2, pp. 119–127, 1980.

[9] J. N. Morgan, J. A. Sonquist, J. N. Morgan, and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American Statistical Association*, vol. 58, no. 302, p. 415, 1963.

[10] R. E. Walpole and R. H. Myers, *Probability and Statistics for Engineers and Scientists*, 5th ed. Macmillan Coll Div, 1993.

[11] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, p. 50, 1947.

[12] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, p. 80, 1945.

[13] G. Casella and R. Berger, *Statistical Inference*. Duxbury Press, 2001.

[14] N. Henze, "A probabilistic representation of the skew-normal distribution," *Scandinavian Journal of Statistics*, vol. 13, no. 4, pp. 271–275, 1986.

[15] I. W. Burr, "Cumulative frequency functions," *Annals of Mathematical Statistics*, vol. 13, pp. 215–232, 1942.

[16] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[17] K. Pearson, "Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material," *Philosophical Transactions of the Royal Society of London. A*, vol. 186, pp. 343–414, 1895.

[18] K. Pearson, "Mathematical contributions to the theory of evolution. x. supplement to a memoir on skew variation." *Philosophical Transactions of the Royal Society of London Series a-Containing Papers of a Mathematical or Physical Character*, vol. 197, pp. 443–459, 1901.

[19] K. Pearson, "Mathematical contributions to the theory of evolution. - xix. second supplement to a memoir on skew variation." *Philosophical Transactions of the Royal Society of London Series a-Containing Papers of a Mathematical or Physical Character*, vol. 216, pp. 429–457, 1916.

[20] Y. Nagahara, "Non-gaussian filter and smoother based on the pearson distribution system," *Journal of Time Series Analysis*, vol. 24, no. 6, pp. 721–738, 2003.