Feature Selection Techniques for Improving Rare Class Classification in Semiconductor Manufacturing Process

Jae Kwon Kim[1], Kyu Cheol Cho[2], Jong Sik Lee[1], and Young Shin Han[3](✉)

[1] Department of Computer Science and Information Engineering,
Inha University, Incheon, South Korea
jaekwonkorea@naver.com, jslee@inha.ac.kr

[2] Department of Computer Science, Inha Technical College, Incheon, South Korea
kccho@ingatc.ac.kr

[3] Department of Computer Engineering, Sungkyul University, Anyang, South Korea
hanys@sungkyul.ac.kr

Abstract. In order to enhance the performance, rare class prediction are to need the feature selection method for target class-related feature. Traditional data mining algorithms fail to predict rare class, as the class imbalanced data models are inherently built in favor of the majority of class-common characteristics among data instances. In the present paper, we propose the Euclidean distance- and standard deviation-based feature selection and over-sampling for the fault detection prediction model. We study applying the semiconductor manufacturing process control in fault detection prediction. First, the features calculate the MAV (Mean Absolute Value) median values. Secondly, the MeanEuSTDEV (the mean of Euclidean distance and standard deviation) are used to select the most appropriate features of the classification model. Third, to address the rare class overfitting problem, oversampling is used. Finally, learning generates the fault detection prediction data-mining model. Furthermore, the prediction model is applied to measure the performance.

Keywords: Semiconductor manufacturing process · Fault detection prediction · Feature selection · Oversampling · MeanEuSTDEV

1 Introduction

Rare class prediction needs the feature selection method, because rare class-related feature in order to enhance the performance. However, conventional data mining methods fail to predict rare class, because most of class- imbalanced data models are inherently built in favor of the majority of class-common characteristics among data instances [1].

The semiconductor manufacturing process is a very complicated process and the structure of the data to be extracted from the process is very complex [2]. Among semiconductor manufacturing processes, some can be detected fault in the FAB process. Therefore, the prediction in the FAB process is important in order to produce the final product. Hence, the pass/fail (regular/irregular) classification technique is necessary in

J.J. Jung and P. Kim (Eds.): BDTA 2016, LNICST 194, pp. 40–47, 2017.
DOI: 10.1007/978-3-319-58967-1_5

the FAB process; the fault detection prediction before final production can improve quality and reliability [3].

Classification method of data-mining can classify pass/fail using semiconductor's various data. In order to generate the classification model, the data preprocessing, including cleaning, feature selection, oversampling, etc., is crucial [1]. Feature selection can increase accuracy of classifying prediction by eliminating unnecessary attributes, while choosing necessary attributes from high-dimension data set [4]. Extract dataset in the FAB process needs the feature selection method, because the extracted dataset from sensor is very complex.

In the present paper, we propose the Euclidean distance- and standard deviation-based feature selection and over-sampling for the fault detection prediction model. We study applying the semiconductor manufacturing process control in fault detection prediction from the SECOM dataset [5]. First, the features of Semiconductor calculate the MAV (Mean Absolute Value) median values. Secondly, the MeanEuSTDEV (Means of Euclidean distance and standard deviation) are used to select the most appropriate features of the classification model. Third, to address the rare class over-fitting problem, oversampling is used. Finally, learning generates the fault detection prediction data-mining model.

2 Methodology

We built a fault detection prediction model using the SECOM dataset [5]. SECOM dataset is the FAB data collected by 590 sensors from the semiconductor manufacturing process. The SECOM dataset consists of the records of 1,567 samples and 590 features. Among the record of 1,567, the fail class is 104 (encoded as 1), the pass class is 1463 (encoded −1). In order to increase accuracy of the rare class prediction model, we have to choose features among the 590 features related to the target class. The imbalance of the pass and fail classes, in addition to the large number of metrology data obtained from 590 sensors, makes this dataset difficult to accurately analyze.

We mainly focused on devising a feature selection method on data-mining techniques to build an accurate model for rare class detection. The framework of our study is shown in Fig. 1.

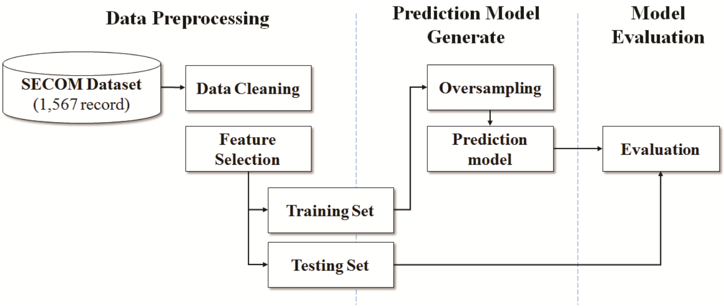


Fig. 1. Frame work

The feature selection method in our study ranges from simply removing a feature with a constant value ('NaN') and missing values (over 60% of values are missing), to statistics-based analysis, such as chi-square, gain ratio, and PCA (Principal Component Analysis). Furthermore, we propose the MeanEuSTDEV to analyze the discrimination of each feature. On the prediction model-building phase, we applied five methods to induce the rare class prediction model, namely LR (Logistics Regression), BPN (Back Propagation Network), SVM (Support Vector Machine), C5.0 (Decision Tree), and KNN (K-nearest neighbor).

Our proposed method for generating a classification model to rare class from the SECOM dataset unfolds in the following steps:

- Data cleaning

First, input the feature data of 590. Remove in the case of a single value. Second, remove the feature in the case of over 60% of NaN (not available) and missing values of record of 1,567.

- Feature selection

We propose the MeanEuSTDEV (Means of Euclidean distance and Standard deviation) based on statistical index. The statistical criterion method can evaluate the distance between two scatter groups (separation index) and directly address the variation of feature in the same group (compactness index). The statistical index should be addressed in both separation and compactness index [6]. ED (Euclidean distance) is the most common use of the distance measure. ED as a separation index. In addition, STDEV is the most robust and widely used measure of variability. STDEV is used as a compactness index. Calculating the MeanEuSTDEV involves the following steps:

(1) Divide data into two class (pass class and fail class)
(2) Calculate the each attribute values using MAV (Mean Absolute Value), median and STDEV (Standard Deviation). *MAV* and *STDEV* is defined as

$$MAV_{class(n)} = \frac{1}{n}\sum_{k=1}^{n}|x_k| \tag{1}$$

$$STDEV_{class(n)} = \sigma = \sqrt{\sum_{k=1}^{n}(x_k - m)^2/n} = \sqrt{(\sum_{k=1}^{n}x_k^2/n) - m^2} \tag{2}$$

(3) Calculate the *ED* using *MAV* and median values. The $ED_{(MAV,\ Median)}$ is define as

$$ED_{(MAV,Median)} = \sqrt{(MAV_{Pass} \text{ - } MAV_{Fail})^2 + (Median_{Pass} \text{ - } Median_{Fail})^2} \tag{3}$$

where, *MAV* and *Median* are the feature mean of two class.

(4) The ratio between ED and STDEV, which we called the MeanEuSTDEV index, is used as a statistic measured index in our study. The MeanEuSTDEV index can be expressed as follows:

$$MeanEuSTDEV = ED_{(MAV,Median)}/STDEV_{(Pass,Fail)} \quad (4)$$

Where $STDEV_{(Pass,Fail)}$ is the average between standard deviation of two classes (pass and fail).

The best performance of classification is obtained when ED is high and the STDEV is low. Hence, MeanEuSTDEV should be large to obtain better performance.

(5) Determine the features that are higher than the average by using descriptive statistics; each value is calculated with MeanEuSTDEV.

- Oversampling

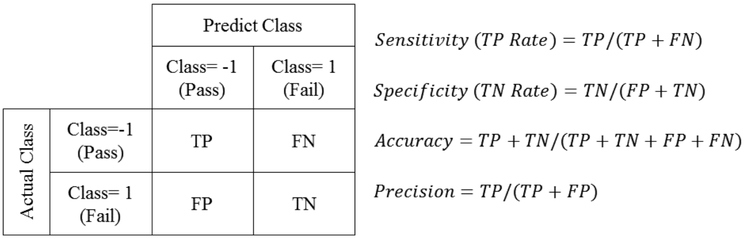
Separate the data into two datasets: the training dataset (70%; total 1,099 records; pass 1,026, fail 73) and the testing dataset (30%; total 468 records; pass 437 fail 31).

Increase the number of records including the fail class in the training data by duplicating the fail class to be same amount as the pass class.

- Prediction model and evaluation

Build a rare class prediction model with LR, ANN, SVM, C5.0, and KNN.

The confusion matrix is used to compare the sensitivity (TP Rate), specificity (TN Rate), precision and accuracy. Confusion matrix is shown in Fig. 2. (TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative)



		Predict Class	
		Class= -1 (Pass)	Class= 1 (Fail)
Actual Class	Class=-1 (Pass)	TP	FN
	Class= 1 (Fail)	FP	TN

$$Sensitivity\ (TP\ Rate) = TP/(TP + FN)$$

$$Specificity\ (TN\ Rate) = TN/(FP + TN)$$

$$Accuracy = TP + TN/(TP + TN + FP + FN)$$

$$Precision = TP/(TP + FP)$$

Fig. 2. Confusion matrix

3 Experimental

We used JAVA jdk 1.8, Weka, and IBM SPSS Modeler 14.2 for the experiment. To measure the performance of the experiment, we compared the results of chi-square, gain ratio, PCA, and MeanEuSTDEV. Preprocessing results for the fault detection prediction model are shown in Fig. 3.

First, data cleaning is using 309 features by removing 271 features that are over 60% of NaN (not available) and missing values among 590 features. Second, feature selection determines ultimate 117 features among 309 by using the MeanEuSTDEV (average 0.1645). From the feature selection results, feature 104 is the best feature as compared to other features (see Fig. 4). Feature 104 obtains the MeanEuSTDEV of 0.841. Feature 104 is higher than the secondary feature 60 (ca. 0.760). Moreover, as shown in Figs. 5 and 6, feature 162 has the highest ED, but the STDEV of this feature is bad. Hence, the

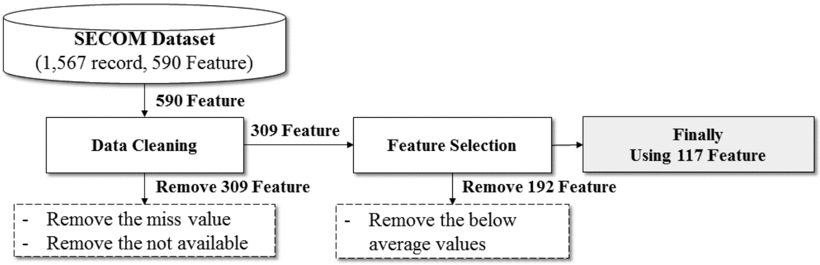


Fig. 3. Preprocessing result

MeanEuSTDEV of feature 162 is not good. However, feature 60 is good group in ED, MeanEuSTDEV. Furthermore, feature 104 is good group in STDEV, MeanEuSTDEV.

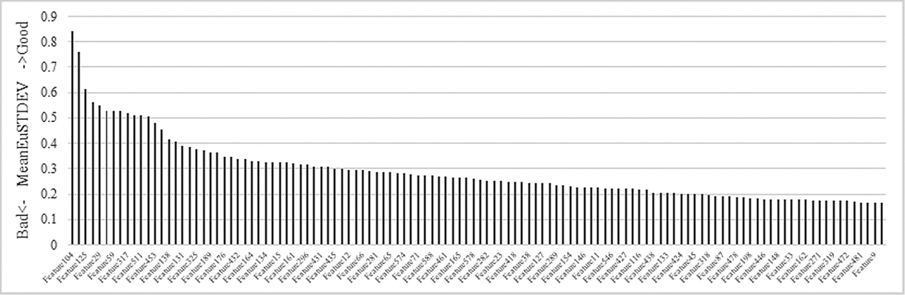


Fig. 4. Bar plot of MeanEuSTDEV of 117 feature

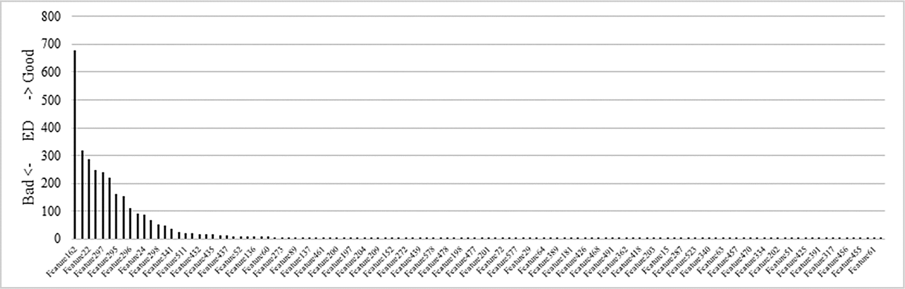


Fig. 5. Bar plot of ED of 117 feature

Next, for generation of the prediction model and experiment, we separated the dataset into the training set, 70% (total 1,099 record; pass 1,026, fail 73) and the testing set, 30% (total 468 record; pass 437 fail 31). Also, the number of training set's fail was set to 953 using oversampling. Finally, the fault detection prediction model was generated using LR, ANN, SVM, C5.0, KNN. The fault detection prediction model's confusion matrix is shown in Table 1. Each model's performance measure is shown in Figs. 7, 8, 9 and 10.

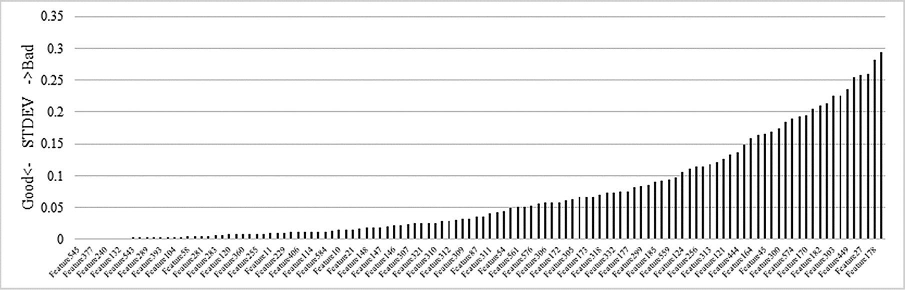


Fig. 6. Bar plot of STDEV of 117 feature

Table 1. Confusion matrix

Model	Feature selection	TP	FN	FP	TN
Chi-Square	LR	317	12	120	19
	ANN	377	23	60	8
	SVM	327	12	110	19
	C 5.0	402	20	35	11
	KNN	387	50	21	10
Gain ratio	LR	319	11	118	20
	ANN	342	14	92	17
	SVM	328	10	109	21
	C 5.0	412	22	25	9
	KNN	391	46	24	7
PCA	LR	303	13	134	18
	ANN	335	14	102	17
	SVM	336	13	101	18
	C 5.0	412	22	25	9
	KNN	391	46	24	7
Non (only cleaning)	LR	357	22	80	9
	ANN	367	23	70	8
	SVM	415	28	22	3
	C 5.0	414	28	23	3
	KNN	390	26	47	5
eanEuSTDEV	LR	353	17	84	14
	ANN	392	26	45	5
	SVM	409	28	28	3
	C 5.0	418	27	19	4
	KNN	386	24	51	7

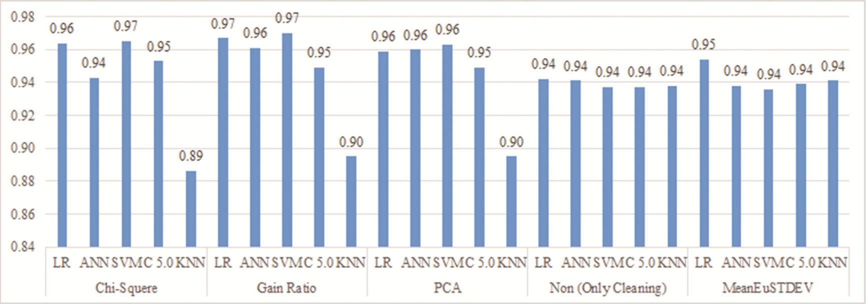


Fig. 7. Performance of sensitivity (TP Rate)

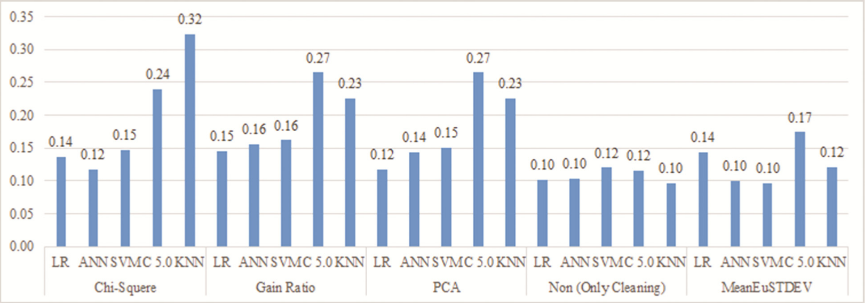


Fig. 8. Performance of specificity (TN Rate)

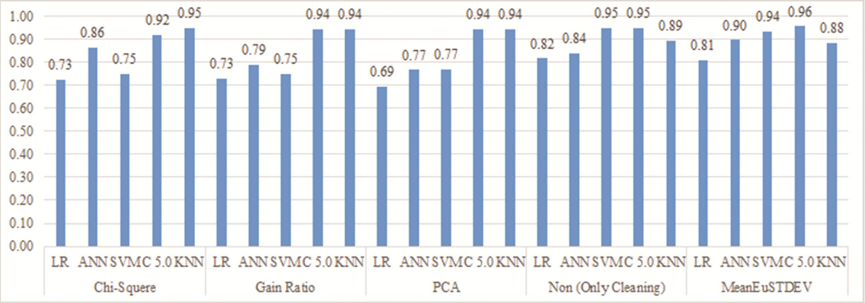


Fig. 9. Performance of precision

The MeanEuSTDEV for C5.0 is highest in precision and accuracy. In other words, the MeanEuSTDEV for C5.0 is useful in feature selection of the semiconductor manufacturing process. Gain ratio is the highest in average of sensitivity (94.8%). PCA is highest in specificity (18.1%) The reason of all model's specificity is lower than 18%, the distribution of pass/fail is imbalance. Therefore, a solution of the data unbalance problem is necessary.

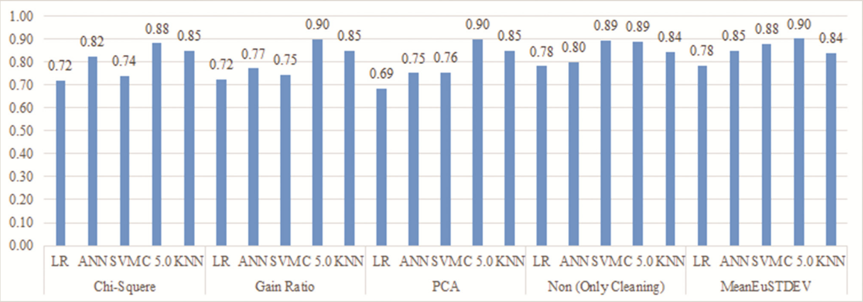


Fig. 10. Performance of accuracy

4 Conclusion

The problem of rare class prediction is important in many real world applications, including the FAB process of semiconductor manufacturing. For a higher accuracy of the rare class prediction model, we need feature selection related to rare class. In this study, we proposed the the Euclidean distance- and standard deviation- based feature selection for the fault detection prediction model. The proposed MeanEuSTDEV calculates the MAV and median value of each feature and extracts the feature using the Euclidean distance and the combination between ED and STDEV. This generates the prediction model using LR, ANN, SVM, C5.0, and KNN. The MeanEuSTDEV for C5.0 is demonstrated to have a better performance than any other feature selection technique.

Acknowledgement. This work was funded by the Ministry of Science, ICT and Future Planning (NRF-2015R1C1A2A01051452).

References

1. Chomboon, K., Kerdprasop, K., Kerdprasop, N.: Rare class discovery techniques for highly imbalance data. In Proceeding International Multi Conference of Engineers and Computer Scientists, vol. 1 (2013)
2. May, G.S., Spanos, C.J.: Fundamentals of Semiconductor Manufacturing and Process Control. Wiley, New York (2006)
3. Purnomo, M.R.A., Dewi, I.H.S.: A manufacturing quality assessment model based-on two stages interval type-2 fuzzy logic. In: IOP Conference Series: Materials Science and Engineering, vol. 105, no. 1, pp. 012044. IOP Publishing (2016)
4. Arif, F., Suryana, N., Hussin, B.: Cascade quality prediction method using multiple PCA+ID3 for multi-stage manufacturing system. IERI Procedia **4**, 201–207 (2013)
5. SEmi COnductor Manufacturing (2010). http://www.causality.inf.ethz.ch/repository.php
6. Phinyomark, A., Hirunviriya, S., Limsakul, C., Phukpattaranont, P.: Evaluation of EMG feature extraction for hand movement recognition based on Euclidean distance and standard deviation. In: International Conference on IEEE (ECTI-CON), pp. 856–860 (2010)