Unsupervised algorithms for the automatic classification of EWS maps: a comparison

Federico Di Palma, Giuseppe De Nicolao
University of Pavia
Via Ferrata 1, 27100 Pavia, Italy
{federico.dipalma-giuseppe.denicolao}@unipv.it

Oliver M. Donzelli, Guido Miraglia
STM-Microelectronics
Via Olivetti, Agrate Brianza, Italy
{oliver.donzelli-guido.miraglia}@st.com

Abstract – Recently, it has been shown that the classification of Electical Wafer Sorting failure maps can be performed by means o funsupervised methods. In this work four different unsupervised methods are compared: SOM, K-Means, Neural Gas, and an Expectation Maximization. The algorithms are compared using a benchmark based on a probabilistic model. The performance of the classification is assessed by means of an new index call Index-F based on the knowledge of the real classification. Moreover it is studied the correlation between the proposed index and the following indexes: CH-index, D-index, I-index and average likelihood.

INTRODUCTION

If the devices that fail at the Electrical Wafer Sorting (*EWS*) tests are visualized as black pixels, the spatial distribution of the failures is likely to show characteristic patterns. Different shapes are possible: circular spots, rings, semi-ring, repetitive chessboard-like patterns, to mention the most frequent. These patterns can be used to trace back to the problems that originated the failures either by analyzing their qualitative features or by correlating them with the lot history. Hence, the interest for algorithms that perform the automatic classification of large wafer sets on the basis of their *EWS* maps. Two approaches are possible: the supervised and the unsupervised ones. The supervised classifiers require the preliminary classification of a training set of wafers by a human operator. This approach is time consuming and when the process and/or the product changes the training has to be performed ex-novo. Recently, it has been shown that wafer classification can be performed by means of unsupervised methods: these algorithms create clusters of wafers by identifying their common features and do not use any training set. In [1], the Kohonen's Self Organizing Map (*SOM*) has been used successfully to classify *EWS* wafer maps. Several other unsupervised learning methods may be used to classify wafer maps.

The first aim of this paper is to compare four different unsupervised methods: *SOM*, K-Means, Neural Gas and an Expectation Maximization (*EM*) classifier. The algorithms are compared using a benchmark based on a probabilistic model. The performance of the classification is assessed by means of an index (here denoted as F-index) that measures the misclassification rate.

Another important issue is finding indexes that measure the classification goodness and are appliable to real data. In fact with real data the F-index (which assumes the knowledge of the true classification) is no longer usable. Hence the second aim of this work is to study the correlation between the F-index and the following indexes [2]: CH-index, D-index, I-index and average likelihood.

In the following section the probabilistic model, the F-index and the benchmark are presented. The third section is devoted to description of the clustering algorithms. The fourth one illustrates the results of the comparison. The choice of a clustering index appliable to real data is discussed in Section 5. The conclusion summarizes the main results.

PRELIMINARIES

To make a comparison a goodness criterion is required. The proposed measure is based on the knowledge of the correct classification. Hence the need for a benchmark made of simulated wafer maps whose true classification is known. In order to obtain the simulated data a probabilistic model is adopted.

Probabilistic Model

In [1] a probabilistic model was proposed. According to this model the electrical test of a die has only two results: good (0) or failed (1). Then a binary Bernoulli variable X_d is used describe the outcome of the *EWS* test for the *d*-th die. It was assumed that the electrical failure of a single device occurs independently of the failure of the others with a probability $P(X_d=1)=f(x_d,y_d)$ where x_d and y_d are the planar coordinates of the die *d*. A complete wafer X^w can be created by simulating N_d Bernoulli trials. The origin of the planar coordinates is placed in the center of the wafer and the scales of the cartesian axes are such that the wafer has radius one. This system of coordinates allows one to describes different scales of integration with the same spatial probability.

The benchmark

A set of production lots is often affected by several failures with different occurrence rates. In this work we simulated 8 classes each of which characterized by its own spatial distribution, reported in Figure 1. The spatial distributions reflect some common failure patterns: standard production (Class #2), low yield wafers (Class #1) repetitive horizontal and vertical (Classes #5 and #6), spot and ring (Classes #3, #4 and #8), semi-ring (Class #7). The clustering algorithms were tested on several benchmarks with different number of devices (N_d) and wafers (N_w) as shown in Table 1

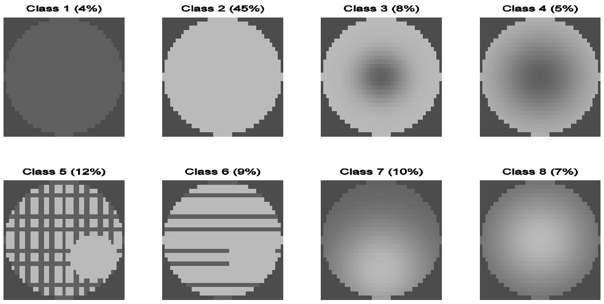
Class 1 (4%)
Class 2 (45%)
Class 3 (8%)
Class 4 (5%)
Class 5 (12%)
Class 6 (9%)
Class 7 (10%)
Class 8 (7%)

Figure 1: Spatial probability of the simulated classes

Benchmark	1	2	3	4	5	6
# Wafers	300	500	800	1200	300	500
# Dies	601	601	601	601	377	377
Benchmark	7	8	9	10	11	12
# Wafers	800	1200	300	500	800	1200
# Dies	601	601	950	950	950	950

Table 1. The twelve data set of the benchmark

Classification assessment

In a classification two kinds of error can be observed: an identified class includes elements coming from different real classes, or a real class is splitted into two or more identified classes. The first kind of misclassification can be disruptive for diagnosis purposes. In fact the reference pattern is explained by the union of different physical problems whose separate detection becomes problematic. Conversely, if a real class is splitted into two or more classes, each identified class is made of wafer characterized by the same physical problem. This misclassification is much less harmful because in the process diagnosis it is possible to merge together class whose reference pattern are similar. For this reason a good classification have to produce homogeneous clusters, each of which made of wafers belonging to the same real class. A simply way to visualize the number and type of misclassified items is given by the scatter matrix [3]. In Table 2 the scatter matrix of a 4 cluster classification of 100 items that belongs to 3 real classes is reported. It can be observed that only clusters #2 and #4 are homogeneous. Clusters #1 and #3 present respectively 4 and 2 elements that are not homogenous. These elements are called "misplaced". For each identified cluster j it is possible obtain the number of misplaced elements as

$$M(j) = \sum_{i=1}^{\tilde{N}_c} T(i,j) - \max_i T(i,j)$$

where with $T(i,j)$ denotes the scatter matrix and $\tilde{N}_c$ is the number of real classes. To measure the degree of homogeneity of the classification we introduce a new index ,called F-index, defined as

$$F = \frac{\tilde{N}_c - 1}{N_w \tilde{N}_c} \sum_{i=1}^{N_c} M(j)$$

It can be shown that the F-Index ranges from 0 (optimal clustering) up to 1 (worst clustering).

	N_1	N_2	N_3	N_4
$\tilde{N}_1$	16	0	14	0
$\tilde{N}_2$	0	45	0	0
$\tilde{N}_3$	4	0	2	19

Table 2. Scatter matrix of a four class clustering of 100 data belonging to three real classes

CLUSTERING ALGORITHMS

In this work are compared three well-known unsupervised clustering algorithms: *SOM*, Neural Gas, K-Mean. This neural network techniques generate a cluster for each neuron. In each neuron c is stored a N_d-dimension vector p^c that represent the centroid of the cluster c.

K-means

This algorithm is the simplest and fastest one. A single iteration is composed by two steps. In the first one each data is associated with the nearest cluster center. In the second step the centers are updated by computing the barycenter (centroid) of each cluster. The two steps are repeated until the centers do not change. The parameters of this algorithm are the number of clusters N_c and the initial values of the cluster centers W_0. It is important to notice that if two cluster centers have been initialized with the same center, the algorithm can encounter difficulties.

SOM

In a *SOM* network the neurons are organized on a $r \times c$ grid. At each iteration t a single wafer X^w (a binary vector) is presented to the network and winner neuron V is defined as

$$V = \arg\min_c \left\| X^w - p^c \right\|$$

where p^c is the characteristic pattern of the c-th class, that is a real vector whose entries belongs to [0,1]. Then all the neurons are updated as follows

$$p^c = p^c + h_T\left(t, \delta\left(c,V\right)\right) \cdot \left(X^w - p^c\right) \quad (1)$$

where *δ(c,V)* is the distance on the grid between neurons *c* and *V* and $h_\Gamma(.,.)$ is a suitable function monotonic decreasing in both its arguments. An interesting feature of the *SOM* is that similar reference patterns are put close to each other on the neuron grid. The function h_Γ in (1) depends on a parameter set *Γ* which include the following parameters [4]: initial η_i and final η_f learning rate, initial σ_i and final σ_f effective width of the topological neighbourhood and total number of iterations t_f.

In [1] the authors proposed a choice of *Γ* for the clustering of *EWS* maps. It was observed that the the first 4 parameters affected the speed of learning [5] while the last has to do with the overall learning time. The first four parameter were fixed to general purpose values [4] and a fine tuning of the parameter t_f. was performed. The proposed set is reported in Table 3.

	η_i	η_f	σ_i	σ_f	λ_i	λ_f	t_f
Γ	.5	.005	3	.1	-	-	80 N_w
Δ	.5	.005	-	-	10	.1	50 N_w

Table 3.Parameter Setting for *SOM* and Neural Gas

Neural Gas

The Neural Gas is quite similar to the *SOM* algorithm. The main difference is due to the fact that the neurons topology (e.g. the neuron grid in the *SOM*) is not fixed. At each iteration the neurons are ordered according to the distance from the proposed input (wafer). The nearest neuron has order *0* and the last has order *Nc-1*. The training update of the neurons is done according to

$$p^c = p^c + g_\Delta\left(t, l(c)\right)\cdot\left(X^w - p^c\right) \qquad (2)$$

where *l(c)* is the order of neuron *c* and $g_\Delta(.,.)$is a function monotonic decreasing in both its arguments. The function g_Δ in (2) depends on a parameter set *Δ* [4]:

$$\Delta = \left\{\eta_i, \eta_f, \lambda_i, \lambda_f, t_f\right\}$$

where λ_i and λ_f affect the amount of information to give to the "far" neurons. The consideration made for the *SOM* method still hold for the Neural Gas. In particular the same tuning procedure can be adopted: the first 4 parameter of *Δ* ware fixed to general purpose value and a fine tuning was made on t_f. Several classifications were performed on the benchmarks described in Section 2 and the F-index was used to assess the performance. The optimal parameter values are reported in Table 3.

Expectation Maximization (EM)

Given the observations *X* and a probabilistic model characterized by a parameter vector *θ,* the likelihood *L* is given by *L=P(X|θ)*. According to our probabilistic model the parameters are the N_c probability vectors p^c and the data are the N_w binary vectors X^w. The maximum likelihood estimate θ^{ML} is the vector that maximize the likelihood. The EM is just an efficient algorithm to perform such a maximization in classification problem. A notable property of the method is that it guarantees an increase of the likelihood at each iteration.

	SOM	*K-mean*	*Gas*	EM
Random	**.02816**	.04262	.03958	.20953
Centre	.02931	**.04036**	.04308	**.19346**
Prototypes	.02871	.07915	**.03789**	.61586
Data mean	.02929	.08779	.04332	.21315
Min-Max	.02943	.08182	.04310	.60589

Table 4. Value of the F-Index in 60 experiments using different initialization techniques

COMPARISON

All the algorithm considered have two types of parameters: the number of clusters N_c and the initial values of the references patterns of the classes.

The number of clusters is not a critical choice for this application. In fact as already observed, in the classification of *EWS* maps the main target is to find homogenous clusters. This is possible only if the number of identified classes N_c is greater than the true one $\tilde{N}_c$. If we have some a priori knowledge on $\tilde{N}_c$ it makes sense overestimate N_c. In the benchmark, we let $N_c=12$.

Conversely, the initial reference pattern may affect the performance of the classifier. In order to choose the best methods, different initializations were considered [6]:

Random In this initialization the patterns are chosen randomly from the valid set.

Center This initialization requires that the valid starting points define a limited region; then the technique chooses the center of the validity set as starting point. In order to obtain different values for each neuron a small random perturbation is added.

Random Prototypes N_c randomly chosen data are used as initialization.

Data Mean This initialization chooses as starting point the mean of the data set. In order to obtain different values for each neuron a small random perturbation is added.

Min Max This techniques use an algorithm to select N_c well distanced data as starting point.

For each benchmark 60 different initialization were generated for each technique. For each classification the F-Index was evaluated. In Table 4 the average over the 12 benchmarks of the median of the F-index over the 60 experiments are reported.

If the bold values in Table 4 are considered, it can be notice that *SOM* gives the best performances.

MEASURING CLASSIFICATION PERFORMANCES ON REAL DATA

The evaluation of the F-Index requires the knowledge of the true classification, that is obviously not available in a real problem. To assess the performance of a classification some other clustering index must be used. Generally these indexes evaluate the within-class and the between-class distances. The first factor is an index of the wafer closeness to the reference pattern while the second one evaluates the distance between the clusters. The most widespread ones are [2]: Davies-Bouldin (*DB*), Dunn (*D*), Calinski Harabasz (*CH*) and *I* indexes. All of them regard a classification with compact and well separated cluster as a good one. However, for our specific application the distance between classes is not the main concern, especially if two or more real classes are close to each other. A further index is the likelihood introduced in Section 3. For computational reason we use the "average device likelihood" (*AL*) defined as:

$$AL = \sqrt[N_c N_w]{L}$$

The goal of our study is to find the index that best approximates the F-Index. For this purpose, from each of the 12 benchmarks 30 different clusterings were randomly created. For each classification all the indexes was evaluated. In Figure 2 some significant scatter plots are reported. Then for each index and benchmark the correlation coefficient was computed. Table 5 shows the average correlation coefficient with the F-Index. It can be noticed that the *AL* and *CH* indexes gives the best correlation.

DB	*CH*	*D*	*I*	*AL*
.9885	**-.9912**	-.5649	.8939	**-.9914**

Table 5. Correlation Coefficient between F-Index and other clustering indexes.

CONCLUSION

In this work a comparison was made between some clustering techniques applied to *EWS* maps classification: k-means, *SOM*, Neural gas and *EM*. The last method and the simulated benchmark are based on a simple probabilistic model. In order to evaluate the performances of a classification, a new index, called F-Index, was introduced. This index was created considering the possible use for fault detection in a semiconductor manufacturing environment. The classifiers were tested simulating different numbers of devices and wafers. From this analysis it turns out that the *SOM* is most suitable algorithms. The last result presented in this work is the choice of a performance index to be used the classification of real data. The objective was to find an index that correlates well with the F-Index, which is proportional to the number of misclassified wafer but requires the knowledge of the correct classification. It turns out to be the Average Likelihood and the *CH* Index.

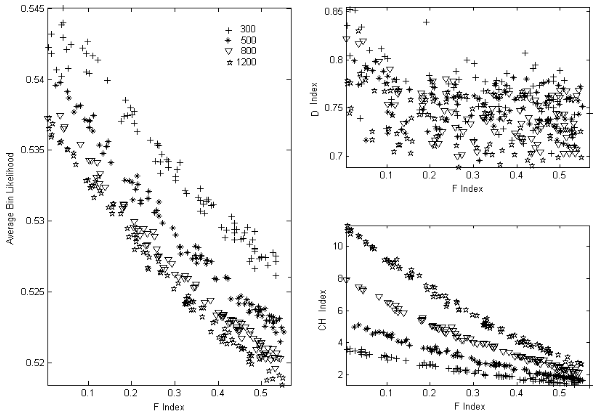

Figure 2. Comparison of indexes measuring the classification goodness. The *AL* and the *CH* index correlate well with the F-index that is proportional to the number of misclassified items

REFERENCES

[1] F DiPalma, G DeNicolao, E Pasquinetti, G Miraglia, F Piccinini, "Unsupervised spatial pattern classification of electrical failures in semiconductor manufacturing, *Pattern Recognition Letters*, 2005 (to appear).

[2] U. Maulik and S. Bandyopadyay, "Performance eavaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 12, pp. 1650–1654, 2002.

[3] A. D. Gordon, *On Classification*. London: Chapman & Hall/CRC, 1999.

[4] B. Fritzke, Some competitive learning methods, 1997, Ruhr-Universität Bochum. ftp://ftp.neuroinformatik.ruhr-uni-bochum.de/pub/software/NN/DemoGNG/sclm.ps.gz

[5] S. Haykin, *Neural Networks, a comprehensive foundation*. New York: MacMillan College Publishing Company, 1994.

[6] A. Juan, J. Garc´ýa-Hern´andez, and E. Vidal, "Em initialisation for bernoulli mixture learning." in *SSPR/SPR workshop*, 2004, pp. 635–643.