Management of Semiconductor Manufacture-A Discussion on Multi-class Classification of Imbalanced Structure of IC Package Database

Y.H. Hung[1] / National Chin-Yi University of Technology
Department of Industrial Engineering and Management
57, Lane215, Section 2, Chung-Shan Road,
Taiping, Taichung, 411, *Taiwan, R.O.C*
hys502@ncut.edu.tw

K.C. Yu[2] / National Chin-Yi University of Technology
Department of Refrigeration and Air Conditioning
57, Lane215, Section 2, Chung-Shan Road,
Taiping, Taichung, 411, Taiwan, R.O.C
yu@ncut.edu.tw

C.P. Huang[3] /
115, Chung-Cho Road,
Lungching, Taichung, 434, *Taiwan, R.O.C*
air.2000@msa.hinet.net

***Abstract*—In the past, for the imbalance class distribution, in most cases the representative class data were chosen by sampling, in order to improve the efficacy of the class distribution model in predicting the minority of classes in the imbalanced data set. The research attempts to present a new pre-processing method of data—the Orthogonal Transformation Method (OTM), which, by integrating the conceptions of Taguchi Orthogonal Arrays, without changing the original data structure, improves the Orthogonality of the data structure by adding variables so that the accuracy of the automatic class distribution database of IC products of imbalanced data set is improved, the range of information retrieval is accurately narrowed, the efficiency and the quality of retrieval can be exalted to a great extent and thus the performance of IC design is upgraded. For the first year, the programs to be implemented and expected results are: Orthogonal Transformation Method, programming and performance evaluation.**

Keywords- IC Package Product; Automatic Classifier; Neural Network; OTM, SVM.

I. INTRODUCTION

In the IC product design and development process, the selection of IC encapsulation pattern is as important to IC designers as the decision of construction techniques for the main building structure to architects. In other words, IC designers have to select the proper IC encapsulation pattern to be able to make choice of subsequent IC chip design patterns and manufacturing procedures accordingly. Therefore, provision of most accurate related information regarding IC encapsulation pattern as soon as possible to IC designers can not only correctly determine the IC design method but also dramatically reduce the risks of subsequent operations. The main purpose of the present study is, by integrating currently available relevant information regarding IC encapsulation pattern including size, features, design principles and application principles, to set up IC encapsulation product classification principles by application of data mining and help IC encapsulation product manufacturers to establish a prototype of IC encapsulation classification system, which may render clients, business personnel, customer service staffs or the management personnel able to rapidly and accurately obtain desired information.

II. RESEARCH BACKGROUND

In recently years, the semiconductor industry was hit by the global financial crisis, and the major semiconductor enterprises were affected to different extents. On condition that the global semiconductor growth tended to be conservative, Taiwan's semiconductor industry was also facing unprecedented challenges. In general, the cost of semiconductor parts accounts for more than 60-70% of that of total parts in the system products. Main problems arising from IC design and manufacture include: IC itself or IC package.

In general, an IC design company has more talents, and can solve the problems of the IC itself and Fab process through the cooperation between test engineers and engineers being familiar with Fab process. However, IC design companies face bottlenecks when the problems seem not to be the IC itself because it is generally difficult for IC design companies to recruit engineers with integrated IC package knowledge. In other words, IC package structure and technical capacity of package are the key elements in IC design and production process. Therefore, at present, all IC design engineers are facing the challenge of how to accurately obtain package information of IC products and select the most appropriate package type, so as to reduce the cost and design period.

In the past, part of the classification database prototypes of IC package products were established using the data mining and automatic classification technology, so as to help customer service personnel of package manufacturers to provide IC design engineers with simple product information (IC package type) with the aid of an automatic classifier [1, 2]. However, the problem of imbalanced class distribution exists in actual product data of the currently completed classification model of the product database, thereby affecting the efficiency of the classifier. This study proposed a new data preprocessing algorithm -orthogonal transformation method (OTM), which could enhance the orthogonality of the product data structure through new variables based on the concept of Taguchi orthogonal array and without changing the

978-0-7695-4443-4/11 $26.00 © 2011 IEEE
DOI 10.1109/ISCCS.2011.97

original data structure. The algorithm was proven to improve the classification accuracy in automatic classification database of IC products of imbalanced data sets.

III. STUDY ON THE CLASSIFICATION OF IC PACKAGE

A. IC design and package type

The current IC chip design process is shown in Figure 1. However, according to the vertical division of labor in the semiconductor industry, a company requiring semiconductor assembly capacity does not always understand the aspects of all assembly lines, or lack resources. If a company only considers chip design, regardless of how to integrate with the package, IC chip design process will not be shortened. In other words, one-sided consideration will lead to problems because the size and package restriction may not be able to meet the needs of the final design. Even if in the planning stage of the chip design, a chip company also needs to consider the final design and determine the feasibility of the final form. However, IC design companies often limit their thinking scopes to the initial design, and do not realize that collaborative feasibility plays a key role in successful development of the final design solution. The design process should focus on the IC level, as well as the substrate level and the PCB level.

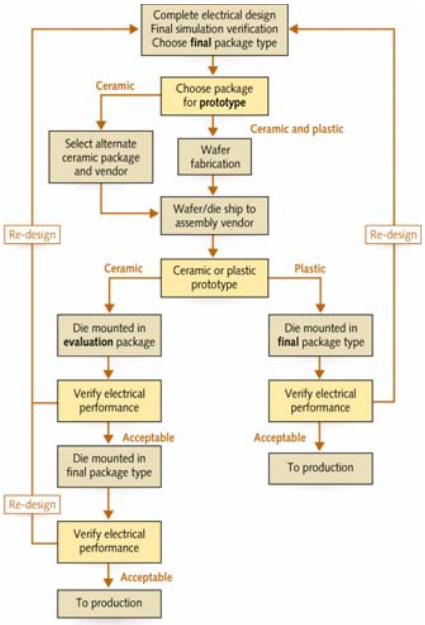

Figure 1. IC chip design process

In other words, only by combining optimized chip design with packages, could higher power density be achieved. Package can improve the ratio of chip area to package area, and reduce thermal resistance. Internal design and material selection must be carefully carried out, so as to ensure the realization of good package reliability. By realizing the package design process and main considerations, companies could then consider package options, and minimize the design time [3].

B. Studies of IC Classification

There are few previous studies on multi-category classification of IC package products. In recent years, classification algorithms such as RST, NN, and SVM, have been widely used in studies of classification of IC package products [1, 2, 4]. SVM was first used to solve the binary classification, but could not be directly used for multi-category classification. Thus, how to effectively promote it to multi-category classification is still a problem. At present, many algorithms have promoted SVM to multi-category classification [5], and these algorithms are generally referred to as "multi-category support vector machines" (M-SVMs). As the market demand changes very rapidly and a number of final product projects are extended, data amount in various categories varies, as shown in TABLE I.

TABLE I. TABLE TYPE STYLES

Package Family	# of Class	Class name (samples)		
TFBGA (632)	16	MCM-VFBGA(16)	6S-STFBGA(16)	VFBGA(32)
		MCM-STFBGA(64)	EDHS-TFBGA(32)	WFBGA(16)
		MCM-TFBGA(64)	3S-SWFBGA(8)	TFBGA(64)
		3S-S2TFBGA(64)	TFTBGA(64)	S2BGA(64)
		4S-STFBGA(32)	SVFBGA(16)	--
		5S-STFBGA(16)	STBFGA(64)	--
QFP (272)	15	EDHS-SQFP(12)	EDHS-LQFP(8)	TQFP(24)
		EDHS-SQFP(12)	E-PADTQFP(24)	S2QFP(12)
		MCM-QFP(12)	DGHS-QFP(12)	LQFP(28)
		MCM-LQFP(28)	DDLQFP(28)	VQFP(4)
		E-PADLQFP(28)	SLQFP(28)	QFP(12)
PBGA (352)	8	EDHS-MCMBGA(56)	MCMBGA(56)	PBGA(56)
		EDHS-PBGA(56)	HS-MPBGA(8)	SPBGA(56)
		MCMS2BGA(56)	MPBGA(8)	--
LGA (424)	17	MCM-S2VTLGA(16)	MCM-LGA(64)	LGA(64)
		MCM-UTLGA(16)	S2VTLGA(8)	SLGA(64)
		MCM-XTLGA(8)	S2-LGA(64)	WTLGA(8)
		MCM-VTLGA(16)	SVTLGA(16)	VTLGA(16)
		MCM-SUTLGA(8)	SUTLGA(8)	UTLGA(16)
		4S-S2VTLGA(8)	WTLGA(16)	--
FCBGA (816)	7	EHS-MCM-FCBGA(96)	EHS2-FCBGA(128)	FCBGA(112)
		EHS-MP-FCBGA(32)	EHS-FCBGA(144)	--
		MCM-FCBGA(144)	DHA-FCBGA(160)	--
Total class	63	--	--	--

Given a large amount of data, including excessive data amount or various attributes of individual data, the frequently used statistical analysis method may fail to meet its basic assumptions. For example, data should be subject to normal distribution, attributes should be independent of each other, and errors should not have self-relevance, etc. However, in practice, processing of imbalanced data set is one of the important subjects in the database with different product demands and fast changes.

C. Related methods to solve classification-based prediction of imbalanced data

In the classification model of imbalanced data type, it is assumed that data in the data set belong to two target categories and, where is the data category with a small amount of data, while is the data category with a large amount of data; if and have greatly different amounts of data, the classification-based prediction model tends to predict all the data as T2 category. At present, the method to solve imbalanced data set is mainly divided into two types: balanced imbalances and imbalanced imbalances. The frequently used balanced imbalances include (1) cost sensitive learning and (2) sampling. For the former, users must define a cost penalty function, and give different classification costs to misclassification results of different target categories. Its disadvantages are that it is difficult to establish the cost penalty function, and classification results of different cost functions are greatly different. The

frequently used sampling methods include: (a) decrease of the majority (b) increase of the minority and (c) multiple expert classifier.

IV. RESEARCH METHOD

A. Data preprocessing-orthogonal transformation method

This study proposed a new data preprocessing algorithm- orthogonal transformation method (OTM), which could enhance the orthogonality of the product data structure through new variables based on Taguchi orthogonal array and without changing the original data structure. The algorithm could improve the classification accuracy in automatic classification database of IC products of imbalanced data sets. When the orthogonality between variables in the data set is increased, distinguishability between the data is also increased. In other words, if the original data are saved, the orthogonal / distinguishable effect of the data set can be increased in the following steps:

1. Assuming that the data set $IS = (U, A)$ contains T_2 category with a large amount of data and T_1 category with a small amount of data.
2. U ={ x_1 , x_2 … , x_n } is a data set, A={ $a_1, a_2 \ldots, a_p$ } is attribute variables, $a \to V_a$, V_a the value set of a.
3. U is divided into two separate subsets D_1 and D_2, D_1={ $(x, x') \in U^2 \mid \forall a \in A$
4. Let B be an orthogonal dummy variable
5. Based on data items [x_i b_i] of D_1', a dummy variable K of D_2 is established, K∈B, and D_2'=[D_2 K] is completed.
6. The converted data set is $U' = \begin{bmatrix} D_1' \\ D_2' \end{bmatrix}$, $IS' = (U', A)$.

B. Evaluation of classification models

In general, training data account for 80% of the original data, while the remaining 20% of the data is regarded as the test data. The main advantage of this method is easy data processing, but it is not applicable to the situation with a small amount of data. In addition, cross-validation method divides the data into m parts with equivalent size using k-fold-cross-validation, and k groups of the data are mutually exclusive. For example, the datum R is divided into m parts R_1 , R_2 , ..., R_m . m training and testing stages must be repeated using this method. Each time one group of R_i there of is regarded as the test data, the remaining data are regarded as the training data to validate its accuracy. After repeating m times, final accuracy of this model is obtained through averaging the accuracies in the m times. This method takes more time than the holdout method, and is less applicable to the situation with a large amount of data. In order to validate the feasibility of OTM, before this study, the IC package family (FCLGA Package Family, FCCSP Package Family, QFN Package Family and SOP Package Family) is established in advance. Classification-based prediction results of these families are compared under the same conditions respectively based on the original dataset and 4 groups of different OTM dummy variable groups. According to the preliminary results of the experiments, as shown in TABLE II, classification accuracy of the training data set can be effectively improved using the OTM of different dummy variables, but that of the test data set varies greatly. The results indicated that the combination of dummy variables has significant effect on the classification results.

TABLE II. PERFORMANCE OF THE CLASSIFICATION ACCURACY OF THE OTM ALGORITHM

		OTM		Sampling		
%	Data sets	OTM1	OTM2	PCA	減少多數	增加少數
90%-10%	IC package dataset	91.5%	95.5%	76.4%	76.4706%	85.2941%
	UCI-Zoo	100%	100%	90.9091%	100%	100%
	UCI-Skin	97.2222%	97.2222%	94.4444%	100%	98.4848%
	Accuracy	96.2407%	97.5741%	87.2512%	92.1569%	94.5930%
80%-20%	IC package dataset	85.7143%	88.4712%	74.5491%	82.7206%	86.6422%
	UCI-Zoo	91.6667%	91.6667%	90.4702%	83.3333%	100%
	UCI-Skin	98.6111%	98.6111%	97.2222%	95.8333%	99.2424%
	Accuracy	91.9974%	87.4158%	87.4158%	87.2957%	95.2949%
50%-50%	IC package dataset	85.0402%	87.249%	75.641%	85.2941%	86.2255%
	UCI-Zoo	86.6667%	86.6667%	92.1569%	85.7143%	100%
	UCI-Skin	98.324%	98.324%	95.5307%	91.6667%	98.7879%
	Accuracy	90.0103%	90.7466%	87.7762%	86.0878%	95.0045%

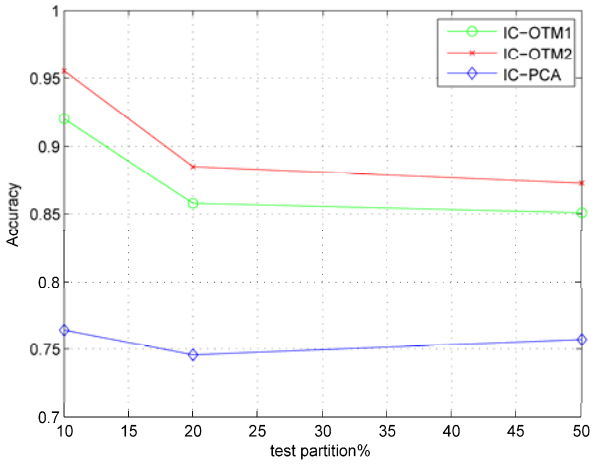

Figure 2. Test performance of the IC package family

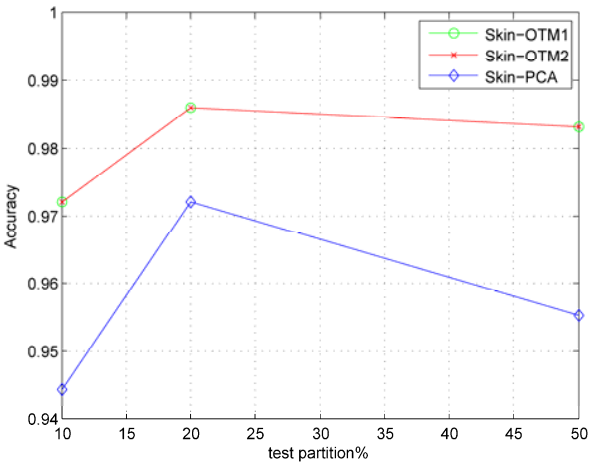

Figure 3. Test performance of the Skin database

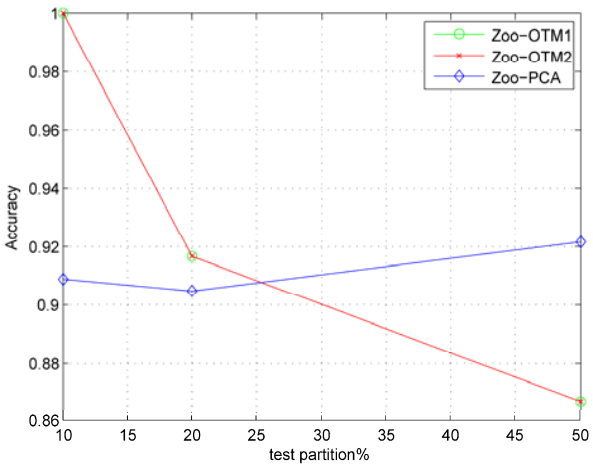

Figure 4. Test performance of the Skin database

V. CONCLUSIONS

According to the experimental results shown in TABLE II and Fig. 2-4, OTM of amplified dummy variables has better performance than traditional PCA and sampling method. Compared to Skin and Zoo databases of UCI, OTM has improved the performance by 9% and 3%, and to 100% and 97.22%. In addition, its test accuracy of the old package family is improved from 76.4% (PCA) to 95.2% (OTM2) with the improvement degree reaching as high as 19%. In general, if a suitable B value can be discovered, OTM will significantly contribute to improve the test performance of existing databases.

ACKNOWLEDGMENT

The authors thank the anonymous referees for their careful reading of the paper and for making several suggestions that improved it. The authors also thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 98-2221-E-167 -010 -MY2.

REFERENCES

[1] Y.H. Hung, "A Neural Network Classifier with Rough Set-based Feature Selection to Classify Multiclass IC Package Products", Advanced Engineering Informatics, Vol. 23(3), 2009, pp.348-357. (NSC-96-2221-E-167-022).

[2] Y.H. Hung, "Improving Classification Accuracy of IC Packaging Products Database Based on Variable Precision Rough Sets", Information Technology Journal, Vol. 7(3), 2008, pp. 440-449. NSC-96-2221-E-167-022.

[3] Phipps G. Selecting the best package for your design, [http://www.ecnasiamag.com/article-12997-selectingthebestpackageforyourdesign-Asia.html], Advanced Interconnect Technologies, 2007.

[4] C.H. Chang, Y.H. Hung and S.C. Liu, "A Classifier to Select the best IC Package type Based on BPNN Algorithm", Semiconductor Equipment and Materials International, technology Report, available at http://www.semi.org.cn/technology/news_show.aspx?ID=594&classid=1, 2009, April. (NSC-97-2221-E-167-014)

[5] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems", Machine Learning, Vol.47(2-3), 2002, pp. 201-233. DOI: 10.1023/A:1013637720281.