ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Feature Selection for High-Dimensional Industrial Data

Michael Bensch, Michael Schröder, Martin Bogdan, Wolfgang Rosenstiel

Eberhard-Karls-Universität Tübingen - Dept. of Computer Engineering
Sand 13, 72076 Tübingen - Germany

Abstract. In the semiconductor industry the number of circuits per chip is still drastically increasing. This fact and strong competition lead to the particular importance of quality control and quality assurance. As a result a vast amount of data is recorded during the fabrication process, which is very complex in structure and massively affected by noise. The evaluation of this data is a vital task to support engineers in the analysis of process problems. The current work tackles this problem by identifying the features responsible for success or failure in the manufacturing process (feature selection).

1 Introduction

As part of the project Overall Equipment Efficiency (OEE)[1], the work package Online Tool Controlling (OTC)[2] deals with identifying problems in the chip-production pipeline. Feature selection can guide the engineer and help solve the problem by giving hints as to which features could be responsible when the number of defective chips reaches a specified level.

High data dimensionality, unbalanced classes (low yield values are seldom), and noise complicate the problem. Also, there is no guarantee that the Process Control Monitoring (PCM) data contains all problem relevant information.

However, PCM data seems predestined for feature selection as the electrical measurements contain many linearly dependent features. We do not consider feature extraction methods such as PCA here. Extraction methods are not so useful for engineers due to a loss of semantics of the features.

We therefore concentrate on feature selection (an overview is given in [1], previous work comparing feature selection algorithms can be found in [2, 3]). We study the following goals in the context of industrial production processes:

1. The selection of a very small set of important features may give the engineer insight into a particular production problem.

2. Using only relevant features may lead to a more robust classifier for yield prediction.

[1] Funded by the BMBF, in cooperation with Atmel Germany GmbH, camLine GmbH, Elmos Semiconductor AG, Fraunhofer, Philips SC GmbH, Robert Bosch GmbH, TIP GmbH, X-FAB Semiconductor Foundries AG, ZMD

[2] Funded by the BMBF and supported by the project partners Robert Bosch GmbH and Elmos Semiconductor AG

To attain these goals, a large part of our work will deal with the comparison of wrapper methods and criterion functions for feature selection on PCM data.

The paper is organized as follows. The following section describes the PCM data. In Section 3 we describe the methods used. Results are presented in Section 4 and discussed in Section 5.

2 Data

The datasets contain measurements such as electrical currents, resistances and layer thicknesses. The dataset `pcm1` contains measurements from multiple wafer sites, whereas the mean of multiple wafer site measurements was used for `pcm2`. By choosing a suitable threshold for the wafer yield, a class label for a two-class problem (signifying "high yield" and "low yield") was determined. Both datasets contain some highly correlated features.

Each dataset was kept in chronological order and split into a training set and a test set, in the ratio 3:1 (number of samples). This leads to an uneven distribution of the classes (see Table 1), and complicates the prediction because PCM data is non-stationary. However, chronological ordering leads to more realistic results. Feature selection was performed on the training set. The test set (unseen by the feature selection) was used to test the performance of the feature selection. The need to use an independent test set is shown in [4].

Name	Samples	Samples in class 1	Features
pcm1_train	3000	57%	61
pcm1_test	1000	31%	61
pcm2_train	1000	48%	85
pcm2_test	313	57%	85

Table 1: Datasets used in the experiments.

3 Experiments

We tested the Sequential Forward Selection (SFS) and Sequential Forward Floating Selection (SFFS) using different criterion functions as a measure for feature subset relevance. The SFS is presented in [5] and consists of successively building up a feature subset by adding one feature at a time. A criterion function evaluates feature subsets and chooses the best feature to add at each step. A drawback of SFS is the "nesting effect": once a feature has been added to the subset, it cannot be removed at a later stage.

The SFFS is described in [6, 7] and allows backtracking as long as this increases the criterion function. The SFFS is said to produce results close to the branch and bound method [8]. The SFFS needs far less computational effort than branch and bound (which is optimal given a monotonic criterion function) and the SFFS does not require a monotonic criterion function.

The criterion functions we compared were a k-NN classifier based on the Euclidean distance (with k=5), the mean of the Mahalanobis distances between each sample of one class and the distribution of the second class (M-dist), and a Fuzzy-ARTMAP (FAM) classifier [9].

To test feature selection results, a FAM classifier was used. FAM was found to be a suitable classifier for wafer PCM data [10], requiring little user interaction (the vigilance parameter, which controls the number of neurons of the network, can be fine-tuned). We used FAM as a criterion function in “fast learning” mode and with vigilance set to 0.8. Each FAM training cycle consisted of 15 repetitions with randomly ordered training samples to reduce dependency on the order of the input samples.

The selected feature subsets were validated with FAM to compare the performance of the k-NN, M-dist and FAM criterion functions. By intuition, FAM should be the most suitable criterion, but we were interested in the potential speedup by using a simpler criterion. The validation graphs (Fig. 2) show the relevance of features for the particular production problem under scrutiny. These are of interest for our first goal (see Section 1).

Finally, we forecast yield values for the test data. For validation and testing, FAM vigilance was set to 0.9 and kept constant to avoid fitting this parameter to the test data. The test results (Fig. 3) show how well the selected feature subsets generalize to new data, which is important for yield prediction, our second goal. They reflect the combined performance of the feature selection algorithm and FAM classifier. The test was repeated 15 times, as for the validation.

There is an important point to note concerning error estimation. A simple 10-fold cross-validation (CV) of PCM data is bound to underestimate the error: While wafer-to-wafer variation within a lot is normally small, lot-to-lot variation can be high. As soon as some samples from a particular lot are included in the training set, forecasting samples from the same lot in the validation set is easier.

Thus, we use a sliding window validation (SWV) method to validate the chosen feature subsets for the k-NN and FAM criterions: Samples are kept in chronological order and a contiguous subset (window) of samples is viewed at a time. Training samples are taken from the front of the window and validation samples from the rear. The window slides through the data in a predefined number of steps. This way we achieve a more realistic validation result.

4 Results

Result graphs are structured as follows: thin lines are SFS results, thick lines are SFFS. Lines are dotted for k-NN, dashed for M-dist, and solid for FAM. SWV results show the appropriateness of the selected features for the closed-world problem (i.e. features relevant for a particular production problem). Test set results show the appropriateness of the selected features for predicting future yield values.

The criterion curves (Fig. 1) peak between 10 and 20 features and imply that not much is to be gained by increasing the subset size further. This shows that our first goal, namely that of finding a small set of relevant features for the training set, can be reached for both datasets.

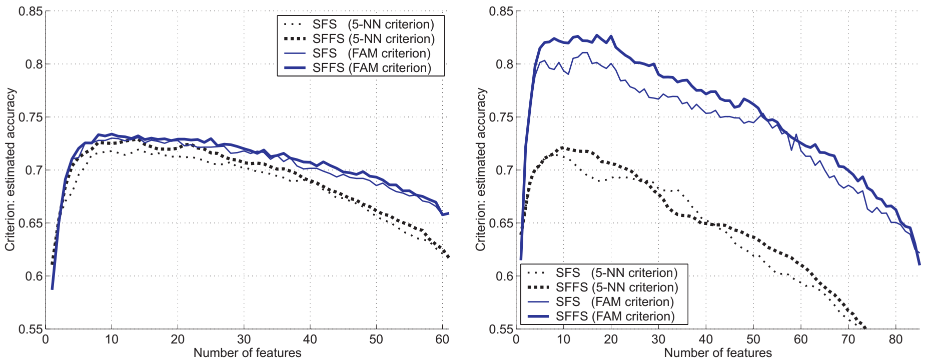

Fig. 1: Left: pcm1 training set. Right: pcm2 training set. Comparison of two criterion functions for SFS and SFFS.

Figure 2 displays the SWV results which are similar to the criterion curves. As expected, the FAM criterion outperformed the other methods. It is evident that the choice of criterion function is much more important than deciding on SFS or SFFS, at least when determining features for the closed-world problem.

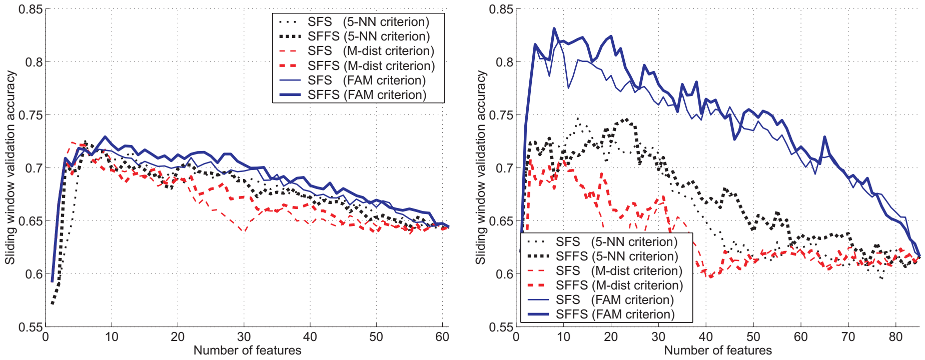

Fig. 2: Left: pcm1 training set. Right: pcm2 training set. FAM estimated classification accuracy (SWV) for feature sets determined by SFS and SFFS.

Figure 3 shows the test results. These reveal that using less than 10 features leads to the best prediction accuracy. This confirms the usefulness of feature selection for reaching our second goal. For `pcm1` data, the same result was obtained by repeating the test on a second `pcm1` test set. This underlines the robustness of the selected feature subsets. All methods show similar performance. The fastest method (M-dist SFS) can therefore be used for this type of data.

The `pcm2` test shows a slight superiority of FAM over the other methods. The difference between SFS and SFFS is negligible considering the high variation in the results. This once again shows that the criterion curve, which displays better values for SFFS, can be misleading. Although higher accuracies are reached than for `pcm1`, finding the correct number of features for prediction is hard.

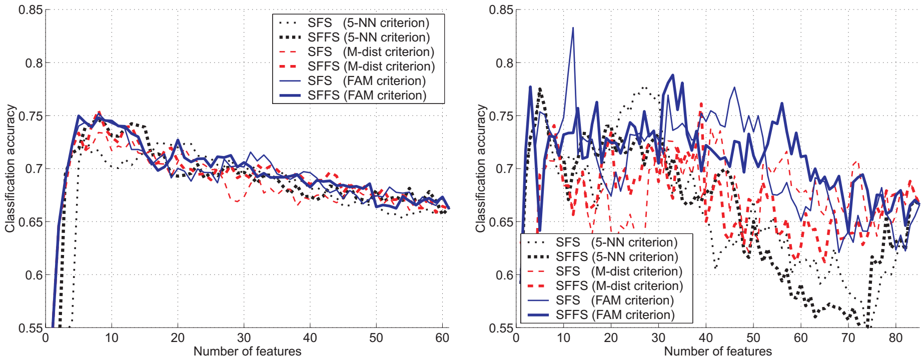

Fig. 3: Left: pcm1 test set. Right: pcm2 test set. FAM prediction accuracy for feature subsets determined by SFS and SFFS.

5 Discussion

Even though the SFFS achieved marginally better criterion results than the SFS, this improvement is not visible in the test results. Given the much shorter running time (by a factor of 3–5 in our experiments), sequential feature selection on PCM data should preferably be done by SFS.

The SWV results undermine the intuition that the classifier and criterion function should be the same in closed-world problems (FAM in our case). For yield prediction however, test results for both data sets show no preference for any of the three criterion functions in the interesting range of 5–30 features. Above 40 features, FAM and M-dist feature subsets clearly outperform k-NN on the `pcm2` data, which could be attributed to the fact that no outlier removal was done on `pcm2` data.

An empirical evaluation of feature subsets was undertaken by engineers to confirm the presence of problem relevant features among the first 5 features selected. In the first case, the subset attained with the M-dist criterion was found to contain the most relevant features for the closed-world production problem under inspection. In the second case, all three subsets were found to contain important features at the first position. The subset attained with the k-NN criterion was viewed as the most appropriate. We conclude that to reach goal 1 (obtain problem insight by selecting a small number of features), no single method was found to be the best. However, the SWV graphs assert that very small feature subsets can already contain most of the information describing the problem.

Note that we did not optimize the test results. To ensure a fair comparison of feature selection techniques, we abstained from adjusting the vigilance and sliding window parameters. Results could be boosted even further by omitting samples in the medium yield range or by introducing the "voting strategy" for FAM presented in [9]. For automated yield prediction in a wafer production environment, cross-validated feature selection should be used [4], taking the highest point on the averaged FAM criterion curve as the number of features to train. In addition, retraining of the FAM classifier should be repeated regularly to shorten the time span between training and yield prediction.

To summarize, we have proposed a procedure whereby an engineer looking for a feature subset describing a particular problem should select a very small feature subset by SFS. We confirmed that a few relevant features model the problem very well, although the choice of the best criterion function was found to be data dependent.

`pcm1` test results suggest that yield prediction can be done well, with an optimal number of 5–15 features in our case. The trend in the `pcm2` graph is not as clear. Thus, we plan to analyze more datasets to confirm our positive yield prediction result.

6 Acknowledgements

The authors thank Elmos Semiconductor AG, Dortmund, and Robert Bosch GmbH, Reutlingen, for providing the data. This project was funded by the German Bundesministerium für Bildung und Forschung (BMBF).

References

[1] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003.

[2] Mineichi Kudo and Jack Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1):25–41, 2000.

[3] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[4] Juha Reunanen. A pitfall in determining the optimal feature subset size. In *Proceedings of the Fourth International Workshop on Pattern Recognition in Information Systems*, pages 176–185, Apr 2004.

[5] A.W. Whitney. A direct method of nonparametric measurement selection. *IEEE Trans. Comput.*, 20:1100 – 1103, 1971.

[6] P. Pudil, J. Novovicova, and J.V. Kittler. Floating search methods in feature selection. *PRL*, 15(11):1119–1125, November 1994.

[7] P. Pudil, F.J. Ferri, J. Novovicova, and J.V. Kittler. Floating search methods for feature selection with nonmonotonic criterion functions. In *ICPR94*, pages 279–283, 1994.

[8] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions in Computers*, 1977.

[9] G.A. Carpenter and S. Grossberg. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 1992.

[10] Lothar Ludwig. *Automatisierte neuronale Netze zur Analyse technischer Daten mit dem Ziel der Qualitätssicherung.* PhD thesis, Eberhard-Karls Universität Tübingen, 2001.