Predicting Defects in Disk Drive Manufacturing: A Case Study in High-Dimensional Classification

C. Apte, S.M. Weiss, and G. Grout

IEEE Annual Conference on AI Applications, CAIA-93

March 1993

Predicting Defects in Disk Drive Manufacturing: A Case Study in High-Dimensional Classification

Chidanand Apté
IBM Research Division
Yorktown Heights, NY 10598

Sholom Weiss
IBM Research Division
Yorktown Heights, NY 10598

Gordon Grout
IBM AdStar
Rochester, MN 55901

Abstract

We consider the application of several compute-intensive classification techniques to disk drive manufacturing quality control. This application is characterized by very high dimensions, with hundreds of features, and tens of thousands of cases. Two principal issues are addressed; (a) can a very expensive testing process be eliminated while still maintaining high quality throughput in disk drive manufacturing, and (b) can the manufacturing process be made more efficient by identifying bad disk drives prior to the expensive testing. Preliminary results indicate that although the expensive testing cannot be completely eliminated, a fraction of the disk drives can be determined to be faulty prior to further testing. This detection may improve the throughput of the manufacturing line.

1 Introduction

We consider the application of several compute-intensive classification techniques [14] to solve an important application in manufacturing quality control. Unlike many applications typically reported in the research literature, this application is characterized by very high dimensions, with hundreds of features, and tens of thousands of cases. We examine the efficacy of modern search-based classification techniques to select dynamically the right features for classification when most features are poorly predictive, and the prevalence of one class overwhelmingly dominates the others.

Our efforts in this area supplement an existing expert system that is currently deployed on a disk drive manufacturing line (RAES [6]). Many man-years have been expended to select the right tests and to improve that system to its current expert performance level. In addition to the RAES knowledge base, a wealth of empirical information is available in the form of stored records of manufactured disk drives.

With the increasing computational power of generally available computers, we have seen a resurgence of interest in automated methods that learn from data, such as neural nets or decision trees [16]. Given the current high performance of the RAES system and the engineers' strong knowledge of the area, our efforts were not directed towards replacement of the expert system. Rather, our efforts were a search for knowledge that would assist the engineers in providing another increment of performance to the already highly performing knowledge-based system. Such knowledge could only be obtained by intensively exploring high dimensional data: the huge volumes of records of disk drive testing and performance that are recorded during the manufacturing process.

1.1 The Application

IBM Rochester, Minnesota, manufactures disk drives on an assembly line. At various points in the manufacturing process, three phases of tests are performed on each drive. These tests produce numerous recorded measurements, most of which are recorded as continuous values. A smaller group of measurements can be characterized as categorical values, and they are recorded as true or false.

The first phase of testing is performed before the frame electronics card is added to the drive. Tests in this phase measure the performance and placement of the components in the drive thus far, especially in regards to the reading and writing of information on the disk surfaces.

The next phase of testing is performed with the frame electronics installed. The tests in this phase measure, among other things, the ability of the frame electronics to interface with and control the read/write mechanisms. Each of the tests in these first two phases requires relatively little time to complete.

Once a drive has passed all tests in these first two phases, the third phase begins with a very lengthy test

called RunIn. The purpose of this test is to run the drive vigorously for a long period of time in hopes of catching the majority of early-life failures before shipment to a customer. Though RunIn itself does not have great direct expense per drive, it causes inventory build-up, can be a bottleneck in the test process, requires a large number of testers to handle the volume of drives, and greatly increases the cycle time per drive. Because of these things, RunIn is very expensive and most desirable to eliminate. Figure 1 approximately illustrates the testing scenario at the disk drive manufacturing site.

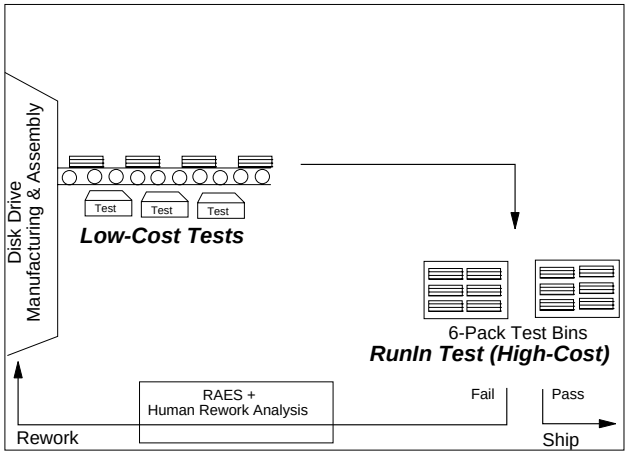

Figure 1: Approximate Current Test Scenario

The RunIn test indicates a pass or fail for a disk drive and also produces several hundred numerical measurements. The results of RunIn are used in determining whether the disk-drive is ready to be shipped. If the RunIn results indicate failure, an expert system, RAES [6], is used in the majority of cases to diagnose the rework that will be required to fix the malfunctioning drive. RAES selects some rework action, usually the replacement of one of a number of components, which it determines will fix the problem. RAES arrives at its conclusion mainly by analysis of the measurements taken at RunIn and relevant previous tests. In some cases, RAES does not have the knowledge required for diagnosis, so a human rework analyzer makes the determination. A bad drive is sent back to be fixed. A reworked drive goes back through part or all of the assembly and test process, including one or more short tests and the long-running RunIn test. A relatively small percentage of the drives fail RunIn. RAES and the rework analyzers together predict the correct rework somewhat less than 100% of the time.

The total number of measurements conducted for each drive (excluding RunIn measurements, but including the outcome of RunIn and the rework recommendations) number well over 600. In terms of data size, thousands of bytes of data is captured for each disk drive that comes off the line. Given the high volume manufacturing capacity of the site, and the fact that these data were being gathered meticulously, there is a huge volume of available data. This paper describes our experiences with applying high-dimensional classification techniques to this data set in an attempt to induce knowledge that could be used for improving the manufacturing process.

The objective of this project is to improve the manufacturing process by predicting the failure of a disk drive prior to the expensive RunIn test. All predictions will be based solely on the results of three of the most relevant tests that occur prior to RunIn. Two types of improvements have been identified:

1. The elimination of the RunIn test for all or some disk drives.

2. The division of the disk drives into two groups, one of which consists of disk drives that are likely to be faulty.

These objectives can be restated in terms of specific classification and prediction problems:

1. Can both the results of the RunIn test and the diagnosis by RAES or a human be predicted for all cases? This task mandates the prediction of the failure of Runin and the determination of one of six classes of failure.

2. Can the results of RunIn be predicted without making a direct determination of the cause of failure and the component that should be replaced?

2 Methods and Procedures

The initial sample consisted of 36,294 records, most of which represent good drives, and the rest drives that needed rework. The result of RunIn is recorded as true (pass) or false (fail). Also recorded in each sample is the rework action recommended for repairing the disk drive. The recommendation is a selection of one of six classes, corresponding to five possible component replacements and one miscellaneous reworks category. While it is desirable to have an objective measure of truth, the only available measure of truth is the classification by RAES and the rework analyzers, which is less than 100% accurate.

The complete elimination of RunIn requires the prediction of the rework diagnosis, a 7 class classification

problem: a good disk drive or one of six bad disk drive types. The prediction of RunIn alone, i.e. the prediction of bad disk drives without identifying the type of problem, is a two class problem. A solution to the RunIn prediction problem may yield efficiencies in performing the RunIn tests. However, the elimination of RunIn requires stronger prediction capabilities than the division into good and bad disk drives.

Two populations were considered. N2 is the true population of 36,294 cases. N1 consists of a smaller number of records divided almost equally between good and bad drives. The N1 population has a proportion of normals that is significantly less than the true proportion. Results on the N1 population are much more easily obtainable, but will be overly optimistic because each error on the normals may actually represent manyfold more errors in a real scenario. However, results on the smaller N1 population will measure the potential for complete success on the true population. Poor performance on the smaller sample set can readily be extrapolated to the true population. For each sample, 564 measurements were obtained and 7 classes were considered. These features were selected from the original 600+ features by simply eliminating all constant value fields.

3 Results for RunIn Elimination for Faulty Disk Drives

Rather than concentrate on any single method, we favored a balanced approach that applies several well-known learning methods in their standard classical form [13]. If any method demonstrates a clear superiority over the others, further experiments can be performed to elicit the best performance of that method. For our quality control application, if no method is clearly superior, the rule-based solutions are particularly advantageous. They have an inherent explanatory capability that is most suitable for further discussions with the manufacturing engineers. Because the problem is one of very high dimensions, even with high performance workstations, timing considerations make it necessary to emphasize only the most promising directions.

Five classification methods were tried: Linear Discriminant [8], k-Nearest Neighbor [4], Neural Network [9], Tree Classification [2], and Rule Induction [3, 10, 11, 12, 15]. These methods were applied to the smaller N1 population. All error rates were measured by test cases obtained by randomly holding-out 1/3 of the sample cases. No method achieved an error rate better than 38%. The following are the results for the (reduced size) N1 population.

k	Error Rate
1	.52
5	.49
11	.48
25	.48

Table 1: Results for k-Nearest Neighbor

3.1 k-Nearest Neighbor

The feature values were normalized by means and standard deviations, Euclidean distance. After demonstrating poor results on the original set of 564 features, the set was reduced to a smaller set of 223 features that showed showed some significance as measured by standard statistical tests. In Table 1, we list the results for k-nearest neighbor, where k is varied from 1 to 25.

3.2 Linear discriminant (parametric) with feature selection

We could not obtain a discriminant for the full 7 class problem. For class 5, the largest class, versus all other classes the error rate is .40 with 75 features selected.

3.3 Decision trees (CART)

The final tree has only 3 terminal nodes. The estimated error rate is 41%. (A ten fold crossvalidation estimates the error rate at 39%.) If the proportion for each class is adjusted to its correct value, no tree is induced and all samples get classified into the first category (good disk drive, perform RunIn and ship).

3.4 Neural Nets (Standard Backpropagation)

Feed-forward, fully-connected, neural networks were trained. These are the standard backpropagation networks. A single hidden layer was used with hidden units varying from 0 to 20 hidden units. The best generalization was achieved at 0 hidden units with a 38.5% error rate on the test cases.

3.5 Rule Induction

The Swap1 [15] procedure for rule induction was applied to the data. The result was a simple rule of

the form $X > a$ *or* $Y > b$ *or* ..., having an estimated error rate of 39%.

The results of all classification methods were very far from our objective of perfect classification. However, the results did hint at a limited predictive capability for some of the tests. We therefore reconsidered a variation of the rule induction approach to find partitioned subpopulations where good predictive performance was possible.

A set of production rules were induced that were trained on the (smaller) N1 population. These decision rules make no errors on either the N1 or the full (36,294) N2 populations. They select a fairly small portion of the bad disk drives, but do not make a decision on the other disk drives. Thus on a very small group there were hints that it may be possible to avoid RunIn. Further testing on new data was necessary to validate this hypothesis. We obtained an independent set of new data. This new sample was from a more recent snapshot of the assembly line and consisted of 51,047 cases. The original set of rules were based on a very small sample. When tested on the new sample, the rules did not perform perfectly, eliminating their feasibility as a good predictor.

4 Results for Faulty Disk Drive Detection

Based upon the results of the rule induction techniques, we hypothesized that bad disk drives could be identified by concentrating on only a few key fields, from amongst the 600+ features that are present, that have a strong presence in the rule set. Fields 595 through 608 have been identified as the key fields. We also hypothesized an artificial test that is a derivative of other tests: the number of these key fields that have a value of 0.05 or more. Analogous to rule induction techniques for medical diagnostic testing [5], we considered thresholding these tests to determine the predictive value at each numerical cutoff, a process known as referent value analysis. In purely symbolic terms a single best rule has been induced in the following form:

If
5 or more fields in 595-608 are ≥ 0.05
OR
There exists a field in 595-608 that is $\geq X$
Then
The drive is bad with probability PV.

X was varied over a range and the corresponding PV computed for this rule over all 36,294 samples.

X	Predictive Value	
	Original Sample	New Sample
0.50	0.859	0.831
0.75	0.913	0.860
1.00	0.950	0.862
1.25	0.945	0.856
1.50	0.956	0.877
1.75	0.944	0.893

Table 2: Predictive Value Analysis

Because predictions are made with an extremely small feature subset on a large population of 36,294 cases, it was conjectured to be highly likely that this result would hold for future predictions. To continue the verification of this hypothesis, it was tested against the second sample that was obtained from the disk drive manufacturing facility. This data set contained 51,047 samples, of which again a small portion were failures. The ratio of failed disk drives as a proportion of the total output is approximately the same as that of the original sample that we worked with. We then applied referent value analysis using our most promising rule on this new data.

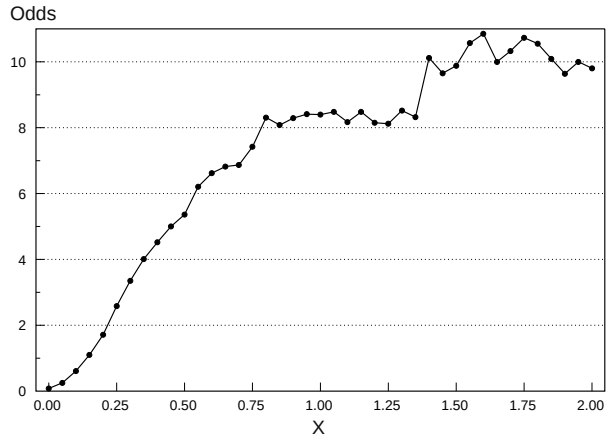

Figure 2: Odds of Detecting Bad Disk Drive in Combined Sample Set

Comparing the results of our predictive value analysis on the original sample set and the new sample set (Table 2), we can observe that the strength of the predictive values seem to remain consistent, with some very minor deviation. We combined both the new and old sample sets and applied the rule to the combined set for a range of X values, to compare the odds that a detected disk drive is bad and the corresponding fraction of total number of bad disk drives that are detected. These results are plotted in Figures 2 and

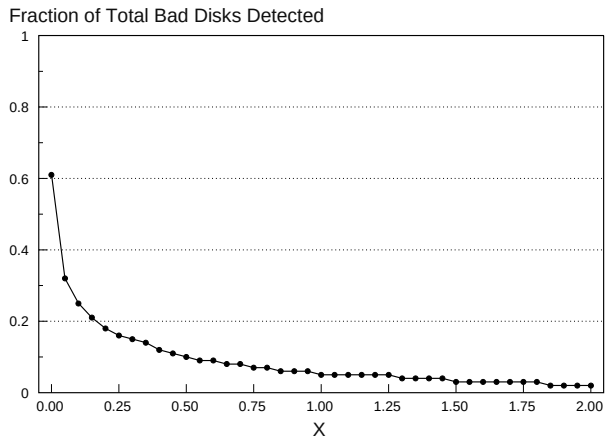

Figure 3: Bad Drives Detected as a Fraction of Actual Total Bad Disk Drives

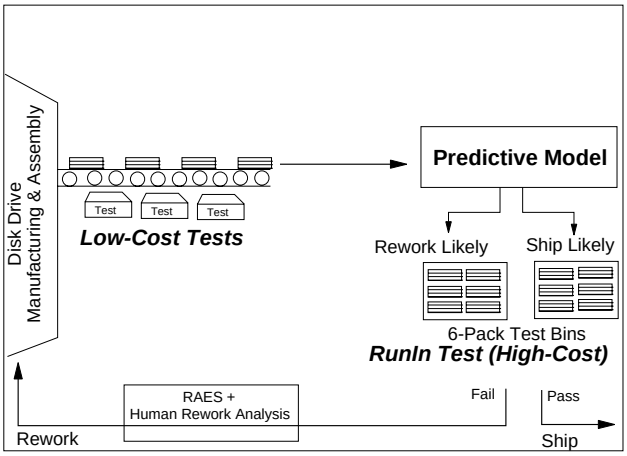

Figure 4: Proposed Test Scenario

3. These plots illustrate, for example, that if X is chosen to be 0.6, then the odds of detecting a bad disk drive are approximately 6:1, and that about 10% of the total bad disk drives will be detected.

5 Improving the Throughput of RunIn

In this section, we review one possible approach to improve manufacturing efficiency using the rule for faulty disk drive detection. The RunIn test is a particular bottleneck in the assembly process. When a disk drive is ready for RunIn, it is packed with others in a testing bin, six to a bin. The 6-pack bin performs RunIn on all six disk drives and is unloaded when all drives in the bin have completed their tests. A bad disk drive will finish sooner than a normal disk drive. Although the six disk drives are tested in parallel, the bin can not be unloaded until disk drive testing has been completed for all six disk drives. If the 6-pack testers can be loaded with bad disk drives, the test process may terminate more quickly, reducing the test queue bottleneck.

Predicting the pass or failure of RunIn corresponds to the determination of a bad disk drive without identifying the failing component. Time may be saved by packing only bad disk drives in the 6-pack bin. Because a bad drive finishes RunIn much earlier than a good drive, a 6-pack loaded with only bad drives will terminate earlier than one which has a combination of good and bad drives. Currently, the probability of loading a six-pack with all bad drives is essentially 0.

Assuming that the desired predictive value is to be at least 90% (in the original sample), the proper value for X can be determined and the number of bad disk drives that may be detected can be calculated. It is likely that by using a rule of this type, the throughput of the 6-pack RunIn process can be improved. Currently, the probability of having a 6-pack bin loaded with all bad drives is essentially 0. If a 6-pack is loaded with disk drives satisfying this rule, the probability is > 0.5 that all the disk drives are bad. About 8% of the bad disk drives satisfy this rule. The actual economies of disk drive screening would depend on numerous factors including the time savings, scheduling, and the number of available 6-packs. The proposed modification to the testing scenario is illustrated in Figure 4.

6 Discussion

A massive amount of data was available for performing our experiments. The two samples that we examined were about 200MB in size. They represent a small fraction of the total data. The high dimensionality of the data (over 600 features per sample) adds a degree of complexity to the analysis. The bulk of these features represent electrical, mechanical, and digital measurements. As is the case with any recordings of this nature, they are potentially prone to noise as well as calibration problems. Our current attempt has focused on combining some preliminary feature analysis with the more widely-used classification methods.

Five classification methods were tried; Linear Discriminant, k-Nearest Neighbor, Neural Network, Tree Classification, and Rule Induction. Using these methods and the sample data and feature set, we performed numerous simulation experiments that suggest the following conclusions:

1. The elimination of the RunIn test for all or some

disk drives is not feasible.

2. About 8% to 9% of the bad disk drives can be detected with greater than 80% certainty.

The first result was fully consistent with our expectations. We did not expect to find that the manufacturing engineers had needlessly been using expensive tests when in fact the inexpensive tests could provide the same information. On the other hand , the amount of human effort effort expended in the data analysis is minimal, and the cost of computer processing pales in comparison with typical knowledge acquisition efforts. Thus, exploring the possibility of eliminating RunIn was a reasonable investigation.

The complete elimination of the RunIn test requires perfect classification. Bad disk drives cannot be shipped so all expected normal disk drives must be tested by RunIn. Good disk drives cannot be sent back for rework because rework is even more expensive than RunIn. The experiments strongly suggest that perfect accuracy in classification cannot be achieved using solely the measurements from the three inexpensive groups of tests. If however, one wishes to pursue further experiments, future directions might include the application of disk drive manufacturing knowledge in conjunction with statistical and information theoretic methods to come up with a more meaningful feature subset. Variations on the distance metric in the k-nearest neighbor technique, or in neural network topologies are also possible alternate experiments to undertake. However, the consistent results that we obtained across an entire family of well known and powerful methods suggest that not much more may be gained without stronger features, i.e. more relevant test measurements of the disk drives during the manufacturing process.

The results of our rule induction experiments provided us with a good starting point for developing a metric for detecting a subset of the bad disk drives. This was made possible by the fact that the outputs of a rule induction system are parsimonious and interpretable. This permitted us to perform further analyses on the induced rule set to extract a promising subcomponent. The fact that such analyses are possible make rule induction methods an excellent choice when explanations of decisions are critical.

We view it as very promising that our final results allow for a subset of the bad disk drives to be detected with high likelihood, prior to the expensive RunIn test. As has been pointed out earlier in this paper, the nature of the RunIn test and the fact that this is a high volume operation, suggest that even the application of this simple test may result in a sizable reduction of the total manufacturing costs, in absolute terms. Manufacturing specialists at the site are currently reviewing the impact of these results.

Increasingly, high volume domain specific data is being made available for knowledge engineering activities across a broad range of applications [1, 7]. There is great potential for using this corpus of data for analysis and induction of hidden knowledge. The high dimensionality of these large data sets will make the application of many classification techniques more complex than usual. For these applications, with many weak or even useless features, machine learning techniques may provide a useful means for *finding the needle in a haystack.*

References

[1] T. Anand and G. Kahn. SPOTLIGHT: A Data Explanation System. In *Proceedings of the Eighth IEEE CAIA*, pages 2–8, 1992.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Monterrey, Ca., 1984.

[3] P. Clark and T. Niblett. The CN2 Induction Algorithm. *Machine Learning*, 3:261–283, 1989.

[4] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

[5] R. Galen and S. Gambino. *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses.* John Wiley & Sons, New York, 1975.

[6] G. Grout and S. Tamminga. RAES - Improving the DASD Rework Process. In *Proceedings of the IBM ITL Conference on Expert Systems*, pages 331–337, 1990.

[7] P.J. Hayes, P.M. Andersen, I.B. Nirenburg, and L.M. Schmandt. TCS: A Shell for Content-Based Text Categorization. In *Proceedings of the Sixth IEEE CAIA*, pages 320–326, 1990.

[8] M. James. *Classification Algorithms.* John Wiley & Sons, 1985.

[9] J.L. McClelland and D.E. Rumelhart. *Explorations in Parallel Distributed Processing.* The MIT Press, 1989.

[10] R. Michalski, I. Mozetic, J. Hong, and N. Lavrac. The Multi-Purpose Incremental Learning System

AQ15 and its Testing Application to Three Medical Domains. In *Proceedings of the AAAI-86*, pages 1041–1045, 1986.

[11] G. Pagallo. Learning DNF by Decision Trees. In *Proceedings of the Eleventh IJCAI*, pages 639–644, 1989.

[12] J.R. Quinlan. Generating Production Rules From Decision Trees. In *Proceedings of the Tenth IJCAI*, pages 304–307, 1987.

[13] B.D. Ripley. Statistical Aspects of Neural Networks. In *Proceedings of SemStat (Séminaire Européen de Statistique)*. Chapman & Hall, 1992. To Appear.

[14] C. Stanfill and D. Waltz. Statistical Methods, Artificial Intelligence, and Information Retrieval. In P. Jacobs, editor, *Text-Based Intelligent Systems*. The MIT Press. To Appear.

[15] S. Weiss and N. Indurkhya. Reduced Complexity Rule Induction. In *Proceedings of the Twelfth IJCAI*, pages 678–684, 1991.

[16] S.M. Weiss and C.A. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann, 1991.