Classification of Imbalanced Data with Transparent Kernels

K K Lee, S R Gunn, C J Harris, P A S Reed*
Department of Electronics and Computer Science,
*Materials Research Group, School of Engineering Sciences
University of Southampton, Highfield Road, Southampton SO17 1BJ
Email : [kkl98r;srg;cjh]@ecs.soton.ac.uk ; *pasr1@soton.ac.uk

Abstract

Two important issues regarding data driven classification are addressed here: Model interpretation and imbalanced data. The aim is to build data driven classifiers that provide good predictive performance for a set of imbalanced data and enhance the understanding of a model by enabling input/output dependencies that exist to be visualised. Here, the classification method is demonstrated on an imbalanced data set that describes fatigue crack initiation in automotive camshafts. To generate interpretable models, the SUpport vector Parismonious ANalysis Of Variance (SUPANOVA) technique is extended to the classification domain. The technique enables an additive decomposition of low dimensional kernel models to be recovered, enhancing model visualization. The standard averaging technique used to assess the performance of the model is inappropriate for imbalanced data. As such, the Geometric mean (Gmean), which is typically maximal on the Receiver Operating Characteristic (ROC) when the true positive and negative classification is balanced between two classes is used. A conventional SVM produced results of generalisation estimate of 55%. However, introducing a class-dependent misclassification cost into the SVM resulted in an improved performance of 74%. The SUPANOVA technique produced a reduced performance of 64%, whilst reducing the model space to just 13 components (10%) out of the possible 512. These resulting components had low dimensions, and consequently can be visualized.

Keywords: ***Imbalanced data, Support Vector Machine, Model Interpretation, Geometric Mean***

1 Introduction

With appropriate heat treatment, Austempered Ductile Iron (ADI) provides good resistance to rolling fatigue, high strength and good wear resistance. This makes it a suitable candidate for camshafts used in automotive industries. However, there is a tradeoff between high strength and the number of fatigue cracks [1]. As such, it is important to investigate why cracks are initiated from the graphite nodules within the microstructure. Clearly, in this example the number of graphite nodules which can be classed as "no crack" exceeds the number of "crack" nodules and consequently the data is imbalanced (see Fig. 1). The graphite nodule size and/or distribution morphology can be obtained from Finite Body Tessellation (FBT). These measurements are used as the features for a classifier to learn the characteristics that cause fatigue crack initiation.

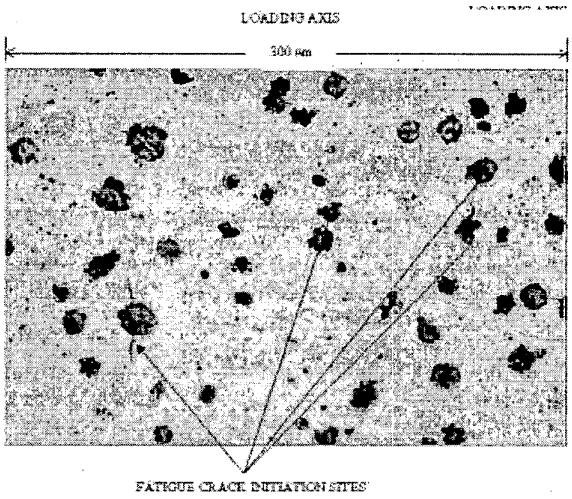


Figure 1: The Original microstructural Images of ADI.

The ability to understand a model's input/output relationships is generally overlooked when classifying e.g. fault detection. In order to optimise the material performance and access the feature validation, it is essential to understand the underlying model, as well as providing good predictive performance for a set of imbalanced data. In accordance with Bayes rule, the best classification rule is obtained when the posterior probabilities of the classes are equal. This is only valid when the cost and class prior are chosen appropriately.

0-7803-7044-9/01/$10.00 ©2001 IEEE

However, in most real world applications, such as medical diagnosis and fault detection in materials, the data set is often limited within one class which is statistically under-represented with respect to the other class. There is the need to impose a different misclassification cost associated with the under-represented class and the heavily-represented class. The use of standard averaging techniques for measuring the overall classification performance is not applicable in imbalanced data, as it causes a bias towards the heavily represented class. One appropriate performance measure for dealing with imbalanced data is the Geometric Mean, which favours a balanced classification by measuring the product of the class classification rate. To generate interpretable models, the SUPANOVA technique [2] was extended to the domain of classification problems enabling a predictive model with a high degree of interpretability to be recovered. With this knowledge obtained from modeling, the key production and microstructure features that allow optimised automotive materials performance can be found.

The structure of this paper is as follows : Section 2 describes the problem with learning from imbalanced data and then proceeds to describe how its performance can be evaluated. Section 3 provides detail of the extension of SUPANOVA to classification domains. Section 4 describes the data obtained. Section 5 shows the results obtained from the SUPANOVA approach applied to our imbalanced data classifying the graphite nodule crack initiation and concludes that with the SUPANOVA techniques, the same practical results are obtained.

2 Learning Imbalanced data

Many machine learning algorithms maximise performance criteria which place equal emphasis on each data point, irrespective of class. Clearly this assumption needs to be modified in order to use imbalanced data. One approach to imbalanced data is to alter the size of the training set by either upsampling or downsampling using intelligent sampling techniques [3]. Alternatively, using Bayes rule, a different cost can be incorporated into each class. Then the decision rule in a two class problem becomes :

$$\phi(\mathbf{x}) = \begin{cases} +1 & \text{if } \frac{p(C|\mathbf{x})}{1-P(C|\mathbf{x})} > \frac{C^+}{C^-}, \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

where $p(C|\mathbf{x})$ is the posterior probability of the given data $\mathbf{x}$ (which is usually unknown) and the C^- and C^+ denote the misclassification cost associated with each class. C^+ and C^- are class priors that can be used to compensate imbalanced data or impose heavier cost for the important class prior. One easy way to obtain them logarithmically and the optimal values using an appropriate performance criteria. Alternatively, the C's can be taken as the prior sampling bias from the training set and then incorporate a heavier penalty on the minority class in order to obtain a better classification rate while the prior for the majority remains the same [4]. In this way, the value of C's are more restricted and less combinations have to be made. Imbalanced data is a typical real world application problem in which the class priors are different.

The performance of a classification process can be described by a confusion matrix, it describes the number of points in the data set corresponding to the four categories: False Positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN); where TP and TN are the correct predictions. In order to represent the true cost function, the standard performance criteria such as the average accuracy which is based on the heavy-represented class of the test set, is not applicable in this case. The Receiver Operating Characteristic (ROC) analysis was made popular in the field of knowledge discovery and data mining by [5]. It measures the classifier performance over the whole range of thresholds from 0 to 1 from the plots of Sensitivity (Se) and Specificity (Sp). Se and Sp are defined as TP/(TP+FN) and TN/(TN+FP) respectively. The average accuracy of the test set is then the summation of Se and Sp. The ROC curve then allows us to represent simultaneously the classifier performance by two degrees of freedom for a range of possible classification thresholds. The GMean is defined as :

$$GMean = \sqrt{TN.TP} \quad (2)$$

and is typically maximal on the ROC curve when TN $\approx$ TP enforcing balance in the classification rate between two classes.

3 SUPANOVA for Classification of Imbalanced Data

SUPANOVA is a Support Vector Machine (SVM) extension technique that was initially used in regression problems for providing model visualization [2]. The technique enables an additive decomposition of low dimensional kernel models to be recovered, enhancing model visualisation. It does this by exploiting the good generalization ability of kernel-based algorithms with an additional structural regulariser placed on the additive kernel sub-models. This work was motivated by ANOVA kernels that were used to obtain structural information, by building up all the subsets of the features [6]. In order for an ANOVA kernel to be applicable to a SVM,

it has to satisfy Mercer's conditions. Each of the additive ANOVA kernel are positive definite and hence the full kernel is also positive definite. A multi-dimensional kernels can be represented by the tensor products of univariate kernels. For example, a two dimensional ANOVA kernel (i.e. d=2) can be decomposed as :

$$K_{ANOVA}(u,v) = \prod_{d=1}^{2}\{1+k(u^d,v^d)\} \\ = 1+k(u^1,v^1)+k(u^2,v^2)+k(u^1,v^1)k(u^2,v^2) \quad (3)$$

This technique was employed in SUPANOVA to produce a sparse ANOVA kernel that enables an input/output interpretation since it is often possible to remove higher order terms, leaving lower order terms which can be interpreted more easily. This is implemented by introducing a coefficient, a , that influences each of the components to be controlled. This is the method used in the SUPANOVA technique [2]. The two dimensional ANOVA kernel is now modified to :

$$K(u,v) = a_0+a_1k(u^1,v^1)+a_2(u^2,v^2)+a_3k(u^1,v^1)k(u^2,v^2)$$

To enforce sparsity within the ANOVA model, a 1-norm on these coefficients is added to minimise the trade off between the error in approximation in SVM with the sparseness representation and is given as :

$$\min_a L(y,\Phi a)+\lambda \parallel a \parallel_1 \quad (4)$$

where L is the loss in approximation of SVM and is associated with the ANOVA basis, Φa. The λ is the structural regulariser parameter controlling sparsity of the kernel expansion and the weights corresponding to each ANOVA basis.

In dealing with imbalanced data, incorporating a modified class dependent misclassification cost function is required for SVMs. The misclassification cost for each class can be implemented to the capacity control, C (i.e. C^+ and C^- for the respective class). This is known as Control Sensitivity CS SVM [7], the initial approach was to impose a heavy penalty on a skewed class and is extended to imbalanced data. The modified minimisation of the cost function in CS SVM is then given as :

$$\phi(\mathbf{w},\xi,\xi^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C^+\sum_{i|y_i=+1}^{n}\xi_i + C^-\sum_{i|y_i=-1}^{n}\xi_i^* \quad (5)$$

subject to constraint : $y_i(\mathbf{w}^T\mathbf{x}+b) \geq 1-\xi_i-\xi_i^*$ and ξ_i , $\xi_i^* \geq 0$. The GMean performance can be easily obtained using the true classification rate of the positive and negative class of the classifier. Previous work done in this research programme, uses several SVM extension techniques for imbalanced data [8]. This work is extended to providing an interpretable model using the SUPANOVA. A fast way of converting the SUPANOVA from regression to be applicable to classification problems, was described in [9]. This is done by altering the model selection from Mean Square Error to classification rate. This paper however, reformulates the regression task to a classification (i.e changing the quadratic (regression) to a hinge loss function). The SUPANOVA technique enables an additive decomposition of low dimensional kernel models to be recovered, enhancing model visualization. This is a very difficult task and is decomposed to 4 stages similar to that of the regression, except the loss function and the model selection must be changed. Here are the stages involved

1. Model Selection
   a good generalisation estimate from the CS SVM based on Gmean provides the value of the two different capacity controls (i.e. C^+ and C^-) for each class.

2. ANOVA basis selection;
   using the values of C's in model selection, Lagrange multipliers, $\alpha > 0$ are obtained. The decision function (below) is decomposed into all its possible subcomponents assuming all the a's to be 1.

   $$f(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i y_i \sum_{j}^{m} a_j K_j(\mathbf{x}_j.\mathbf{x}) \quad (6)$$

   where $\alpha_i > 0$ are the Lagrange multipliers, y_i are the targets, a_j are the model coefficients, n is the number of training patterns and m is the number of additive kernels used in the model.

3. Sparse ANOVA selection;
   this reduces the number of model coefficients, $a_j \geq 0$ from stage 2 by a 1-norm imposed on the additive model coefficients. The solution to the hinge loss function is then given by :

   $$\phi(\mathbf{a},\xi,\xi^*)=\lambda \parallel \mathbf{a} \parallel_1 + \mathbf{C}^+\sum_{\mathbf{i}|\mathbf{y_i}=+1}^{\mathbf{n}}\xi_\mathbf{i} + \mathbf{C}^-\sum_{\mathbf{i}|\mathbf{y_i}=-1}^{\mathbf{n}}\xi_\mathbf{i}^* \quad (7)$$

   subject to the constraint : $\mathrm{diag}(y)\Phi\mathbf{a} \geq \mathbf{1}-\xi_\mathbf{i}-\xi_\mathbf{i}^*$ and ξ_i , $\xi_i^*, a \geq 0$. Where y_i is the target, Φ is the ANOVA basis obtained, λ is the structural regulariser and ξ's are the slack variable that measures the distance of a point from the optimal hyperplane for the respective classes. Hence, providing interpretability through the additive kernel function.

4. Parameter selection
   using only those coefficients selected in stage 3, re-

construct a new model using.

$$f(\mathbf{x}) = \mathrm{Sgn}(\sum_{i=1}^{n} \alpha_i y_i \sum_{j}^{m} a_j K_j(\mathbf{x}_j.\mathbf{x})) \qquad (8)$$

Prior to proceeding, the form of the univariate kernel must be chosen. There are many kernels that can be employed, such as Radial basis functions, polynomials, splines. However, there are additional parameters within many of these kernels that must be determined. Whilst they provide increased flexibility to the model, a significant additional cost is introduced. A spline kernel has been used here as it does not require any additional parameter to be determined and it does have the ability to approximate any function with arbitrary accuracy [10].

4 Data Description

The ADI materials data set for the automotive camshaft application contains a total of 2923 examples of which 116 samples are crack initiation sites ("Crack" class) while 2807 samples did not act as crack initiation sites ("No Crack" class). These data were obtained from a FBT of ADI [1]. A set of nine measurements relating to the spatial distributions and measures of the object (graphite nodules) were obtained from the tessellation. This set of nine features describes the prior domain knowledge of the microstructural distributions e.g. morphology of secondary particles. FBT involves three stages : binarisation of images, a distance transformation and a watershed transformation [11]. During these stages, noise in the background is attenuated and holes within bodies are filled. Then, the features for each nodule for learning are generated from the following measurements : x_1 Nodule area; x_2 Nodule aspect ratio, x_3 Nodule angle, x_4 Cell area surrounding the nodule, x_5 Local cell area fraction, x_6 Number of near neighbours, x_7 Nearest neighbour distance, x_8 Mean near neighbour distance and x_9 Nearest neighbour angle. Where the near neighbour cells are defined as the cells that are sharing the same boundaries. Prior to using the different approaches to classifying the graphic nodule, the input features are normalised. This will ensure that the input feature is restricted to a unit domain and it provides no bias on the significance of each feature. Upon normalising, the data needs to be partitioned for training and testing. Due to the extensive computation time required for a large data set, a reduced data set was considered which consists of 700("no crack") and 90("crack") for training and the rest of the data being used for testing. This was repeated ten times with random selection of the data each time.

5 SUPANOVA Results

Work done previously by us [8], has shown several SVM extension techniques dealing with imbalanced data. Table 1, shows some of the results obtained using the SVM for classification. In contrast to previous work, based on five random selected testing and training set, we have extended this approach to the use of ten sets. These results were based on setting the capacity control, C (i.e. C^+ - crack and C^- - no crack) sampled logarithmically on [0.01,10000] for each class using the Spline and Radial Basis Function (RBF with σ=0.5). Similar results were obtained using both types of kernels. Using an imbalanced data, the conventional SVM only provides a Gmean of 55%. Further investigation using the CS SVM, shows that the ratio of the C's (i.e. C^+=1 and C^-=0.1) coincide with that of the ratio of the data size (i.e. 90:700 for crack with no crack). This shows that the conventional SVM estimate coincides with the Bayes optimal decision rule, which has been consistent with the proof by [12]. In order to reduce the number of parameters to be determined, spline kernels were used throughout this paper. In the conventional SVM, the C's value is much higher than those of the CS SVM. This indicates that while using the SVM, the minimisation in Eq. 7 emphasises the misclassification error. In conclusion, the CS SVM is more appropriate for the application to imbalanced data as it provide the best Gmean performance of 74% and a lower Gmean variance.

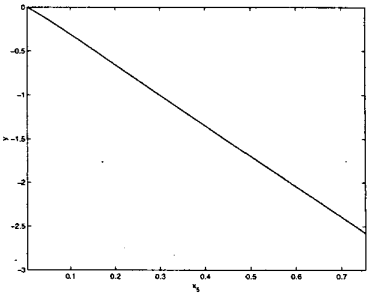
Approaches	TP Crack	TN No Crack	GMean (variance)
Conventional SVM	0.34 C+=1000	0.90 C-=1000	0.55 (0.0315)
CS SVM	0.72 C+=1	0.77 C-=0.1	0.74 (0.0202)
SUPANOVA	0.80 C+=1.0	0.53 C-=0.1	0.64 (0.0275)

Table 1: Summary of test results from SUPANOVA for classification model.

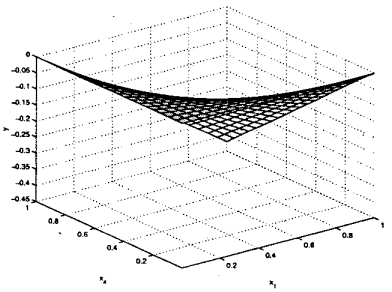
Upon obtaining an acceptable performance value (i.e. with known value of C's), the SUPANOVA was then used to generate model interpretability. The parameter to be determined here is the structural regulariser, λ set in the range in [0.05,1] with an increment of 0.1. The number of components selected was based on its occurrence more than 5 times out of the 10 randomly selected data sets. With λ=0.05, the Gmean for the classification was 64%. A tradeoff of 10% in Gmean leads to 13 significant components being selected among the possible 512. Increasing the value of λ reduces the number of components selected, hence providing a more

easily interpretable classification model. However, the Gmean value is reduced resulting in worse classification model. Table 2 describes the plots obtained from the selected components. Figure 2 shows examples of plots with some of the components selected versus the output from the SUPANOVA classification model. Although, there are different trends observed in component $x_7 \otimes x_8$ and $x_1 \otimes x_5 \otimes x_6$, these differences contribute only a small amount (i.e the output Y, is very small). As such, it can be ignored. There is a inconsistent observation in feature $x_3 \otimes x_6 \otimes x_9$, and further investigation is required. The three most commonly selected components (i.e. based on the univariate) are x_6, x_5 and x_8 respectively. This provides information on the importance of each individual component with respect to other components. Hockley *et al.* [1] uses simple comparisons of means and shows that x_1 and x_5 are important components. Here, we extend our understanding using a predictive and mathematical model. The SUPANOVA technique shows that the model is consistent with those findings with more complex and interesting features selected, which give an improved insight into the possible physical mechanisms occurring.

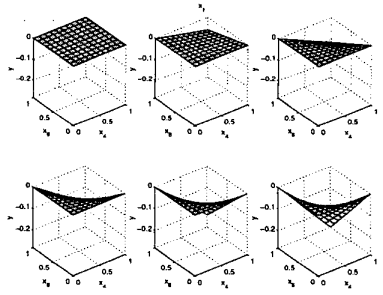
In summary, larger graphite nodules x_1, of high local area fraction x_5, (i.e. local clustering from near neighbour nodules x_6) act as fatigue initiation sites. This may be understood in the following mechanical terms: the graphite nodules have a significantly lower effective Young's modulus than the surrounding matrix, decohere easily and may be considered to act as holes in a mechanical sense. The predominantly spherical nature of the nodules indicates that size increases will not increase the local stress concentration factor, although the larger graphite nodules will give a larger sampling volume of potential initiation points. Local clustering around such larger graphite nodules (as identified by the classifier) may be expected to superimpose local particle stress fields, raising the peak stress levels. The more complex bivariate and trivariate relationships are somewhat harder to assess. The object angle x_3 defines the angle between the loading axis and the major axis of the nodule and if this is high the major axis of the nodule is closer to perpendicular to the tensile axis (which might be expected to promote cracking). However this combined with a relatively far away nearest neighbour might be expected to minimise superimposition of local particle stress fields, and hence make these nodules less likely to act as crack initiation sites. Given the very low aspect ratio of the nodules (they are effectively spherical) correlations with object angle are surprising. Similar reasoning can be applied to the trivariate relationship identified by the classifier, here the situation where the nearest neighbour is aligned either parallel to or perpendicular to the nodule appears to reduce the likelihood of crack initiation, which might be attributed to a local shielding effect. These interpretable classification results allow us to assess the relationships that give rise to crack initiation and hence to identify optimised microstructures with good fatigue resistance for the camshaft application.



(a) Local Area Fraction, x_5



(b) Graphite Area, x_1 Vs Cell Area, x_4



(c) Graphite Area, x_1, Cell Area, x_4 and No. of near neighbours x_8

Figure 2: A example of plots with those components selected versus the output with SUPANOVA for the classification task. Where a negative trend in Y denotes a crack initiation.

6 Conclusions

By utilising the CS SVM and Geometric mean performance criteria, the conventional SVM can be extended in its application to imbalanced data. The SUPANOVA has been extended to the classification domain for im-

Components	Freq.	Description
x_5	7	As local area fraction increases, cracks likely to initiate
x_6	9	As number of near neighbours increases, cracks likely to initiate
x_8	8	Mean near neighbour distance has a threshold of 0.2-0.3 before cracks initiate
x_9	7	As Nearest neighbour angle increases, cracks unlikely to initiate
$x_1 \otimes x_4$	6	Both object and cell area increase, cracks likely to initiate
$x_2 \otimes x_5$	8	Both object aspect ratio and local cell area increase, cracks likely to initiate
$x_3 \otimes x_5$	9	Both object angle and local cell area increase, cracks likely to initiate
$x_3 \otimes x_8$	7	Both object angle and mean near neighbour distance increase, cracks unlikely to initiate
$x_7 \otimes x_8$	8	Same observation as above. However, 2 data sets show different trend.
$x_1 \otimes x_4 \otimes x_6$	7	As all three features increase, cracks likely to initiate
$x_1 \otimes x_5 \otimes x_6$	7	Same observation as above. However, 1 data set show different trenches.
$x_3 \otimes x_6 \otimes x_9$	6	A very inconsistent observation is obtained
$x_4 \otimes x_6 \otimes x_8$	5	As all three features increase, cracks likely to initiate

Table 2: These results are based on considering the 10 randomly sampled sets and the selected component terms occur more than 5 times during the 10 runs.

balanced data, and shown to provide a comparable performance to an SVM whilst providing much greater insight into the model, enabling input significance, and input interactions to be visualized. With this knowledge obtained from the modeling, the key production and microstructure features that will optimise automotive materials performance can be found, hence producing a camshaft which is more resistant to fatigue cracks.

References

[1] R. Hockley, D. Thakar, J. Boselli, I. Sinclair, and P. Reed, "Effect of graphite nodule distribution on 'crack' initiation and early growth in austempered ductile iron," *Small Cracks Mechanics and Mechanisms*, 1999.

[2] S. Gunn, "Supanova - a sparse, transparent modelling approach," *In Proc. IEEE Int. Workshop on Neural Networks for Signal Processing*, 1999.

[3] F. Provost, D. Jensen, and T. Oates, "Efficient progressive sampling," *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, 1999.

[4] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[5] F. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," *In Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Press*, pp. 43–48, 1997.

[6] M. Stitson and J. Weston, "Implementational issues of support vector machines," Tech. Rep. CSD-TR 96-18, Computational Intelligence Group, Royal Hollway, University of London, 1996.

[7] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support machines," *Proceedings of the Int. Joint Conf. on Artifical Intelligence (IJCAI99), Sweden*, 1999.

[8] K. Lee, C. Harris, S. Gunn, and P. Reed, "Approaches to imbalanced data for classification : A case study," *International ICSC Congress on Computational Intelligence Methods and Applications (CIMA) - Advances in Intelligent Data Analysis*, 2001.

[9] K. Lee, C. Harris, S. Gunn, and P. Reed, "Regression models for classification to enhance interpretability," *Intelligent Processing and Manufacturing of Materials*, 2001. Submitted.

[10] G. Wahba, *Spline Models for Observational Data*, vol. 59 of *Series in applied Mathematics.* Philadelphia: SIAM, 1990.

[11] J. Boselli, P. Pitcher, P. Gregson, and I. Sinclair, "Secondary phase distribution analysis via finite body tessellation," *Journal of Microscopy*, vol. TM 140, 1998.

[12] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in nonstandard situations," Tech. Rep. 1016, University of Wisconsin, March 2000.