A case study of SVM extension techniques on classification of imbalanced data

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

Abstract

In many classification problems the data is imbalanced, that is the class priors are different. Here, we consider the classification problem of fatigue crack initiation in automotive camshafts, where this imbalance is significant. The extension techniques of Support Vector Machine (SVM) - the Control Sensitivity (CSSVM) and Adaptive Margin (AMSVM) - which offer different ways of dealing with imbalanced data was investigated. Geometric mean was used to evaluate the performance of the model. The CSSVM has outperformed the AMSVM. The use of different kernels did not produce significant changes in the results. The ratio between the misclassification cost and the training size for each class is very similar, indicating a strong relationship between them.

Keywords: ***Imbalanced data, SVM, Geometric Mean***

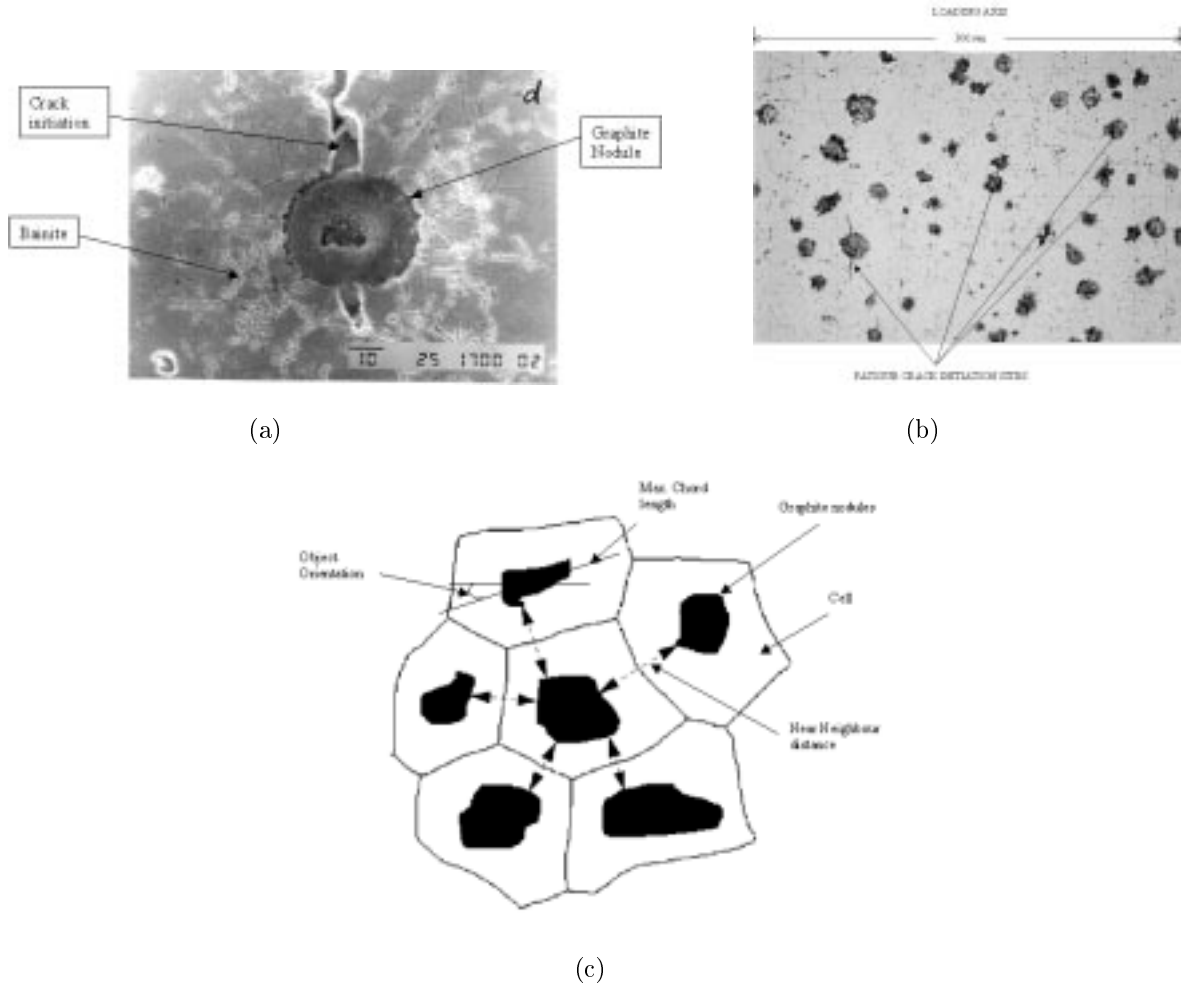
1 Introduction

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 trade-off between high strength and fatigue cracks. As such, it is important to investigate why the crack was initiated from the graphite nodule (see figure 1a). Clearly, the number of "no crack" nodules exceeds those of "crack" nodules and consequently the data is imbalanced (see figure 1b). The graphite nodule size and/or distribution morphology can be obtained from Finite Body Tessellation (FBT) (See figure 1c). These are used as the features for a classifier to learn the characteristics which cause fatigue crack initiation.

The best classification rule is obtained from Bayes rule when the posterior probabilities of the classes are equal. This is only valid when the cost and class size associated with each class is the same. In most real world applications, such as medical diagnosis, the data set is often limited within one class which is under-represented as compared to the other class. Imbalanced data is a well-known phenomenon. Usually, the cost of misclassification associated with the under-represented class are usually more severe than the heavily-represented class. The use of the standard averaging technique for measuring the performance is not applicable in imbalanced data. An unequal cost can be incorporated to distinguish more appropriately between crack and no crack classes. One such performance measure is the Geometric mean, which favours a balanced classification between two class. The structure of this paper is as follows : Sections 2 describes how the problem of imbalanced data can be handled and then proceeds to describe how its performance can be evaluated. Section 3 then describes the SVM and its extensions for imbalanced data. Section 4 provides the model specification and section 5 the results.

2 Incorporating Misclassification Cost for Imbalanced data

The optimal rule for a two class problem is to compare the posterior probabilities of the different classes, assigning the class label to that class with highest posterior probability. With imbalanced data, the true cost function and size associated with each class is usually different. In order to use the Bayes rule more appropriately in this instance, a different cost can be incorporated into each class and the de-



(a) (b)

(c)

Figure 1: (a) Crack initiation (b) Original microstructural Image (c) FBT Image.

cision 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 $\mathbf{x}$ and the C^- and C^+ denote the misclassification cost associated with each class.

2.1 Performance Criteria for Imbalanced Data

A confusion matrix can be used to describe the performance of classification problems. 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); TP and TN are the correct prediction. In order to represent the true cost function, the standard performance criteria such as the average accuracy of the test set is not applicable in this case.

Receiver Operating Characteristic (ROC) analysis measures the classifier performance over the whole range of thresholds from 0 to 1. This performance is based on measuring the sensitivity (Se) and specificity (Sp). Sensitivity and specificity define the performance in terms of predicted classifications within each of the true classes (FP and FN) :

$$Se = \frac{TP}{TP + FN}; \qquad Sp = \frac{TN}{TN + FP} \quad (2)$$

The average accuracy of the test set is then the summation of Se and Sp. ROC curve describes the tradeoff between sensitivity and specificity with the plot of TP against FP. The ROC curve then allows us to represent simultaneously the classifier performance by two degrees of freedom for a range of possible classification thresholds. Within the ROC curve, a point that has a high value when the TP and TN of the classifier is balanced is known as the geometric mean [1]. This enables a point to be selected on the ROC

curve by maximising :

$$GMean = \sqrt{TN.TP} \tag{3}$$

3 Support Vector Machine

Support Vector Machines (SVM) have gained success in recent years for many classification and regression problems [2], [3]. The SVM was developed based on the idea of *"Structural Risk Minimisation"* (SRM). SRM is used to approximate the function of the learning machine based on its complexity and the finite learning data. Minimising allows a good generalisation to be obtained. Given a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with input $\mathbf{x}_i$ and output $y_i \in \{\pm 1\}$, the optimal hyperplane can be obtained by maximising a margin which is defined as the minimum distance from the hyperplane to the closest point. The optimal hyperplane weight ($\mathbf{w}$) is given by :

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{4}$$

where α_i are known as the *Support Vectors* which can be obtained by solving a quadratic program (QP) by maximising :

$$Q(\alpha) = \sum_{i}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

subject to the constraint $\alpha_i \geq 0$, $\sum_i^n \alpha_i y_i = 0$. For the case where a linear boundary is inappropriate, a Kernel, can be used to transform the input vector to a higher dimensional space via a non-linear transformation. Kernel methods calculate the nonlinear transformation from the input space explicitly; only the dot product of the input vector is required. For the case where the data are non-separable in the feature space, a slack variable, ξ, is introduced to allow for misclassification errors and a capacity control, C, for controlling the tradeoff between complexity of decision boundary and the number of errors allowed. The solution for the QP is similar to that of separable case except that it has an upper limit on α (i.e. $0 \leq \alpha \leq C$).

3.1 Control Sensitivity (CSSVM)

Splitting the C according to the respective classes implies that the misclassification cost associated with each class is different [4]. This leads to a minimisation problem, for the CSSVM,

$$\begin{aligned} \phi(\mathbf{w}, \xi, \xi^*) = \frac{1}{2}\|W\|^2 + C^+ \sum_{i|y_i=+1}^{n} \xi_i \\ + C^- \sum_{i|y_i=-1}^{n} \xi_i^* \end{aligned} \tag{6}$$

subject to the constraint :

$$\begin{aligned} y_i(\mathbf{w}^T \mathbf{x} + b) &\geq 1 - \xi_i - \xi_i^* \\ \xi_i \ , \ \xi_i^* &\geq 0 \end{aligned} \tag{7}$$

where C^+ and C^- are the misclassification costs associated with positive and negative classes. This is suitable for our imbalanced data for fatigue crack prediction in camshafts as the misclassification costs of each class are different. Furthermore, there is a relationship between misclassification cost and class size. Eq. 6 and 7 can be solved with a simple modification to Eq. 5.

So far the L_1-norm loss function has been used and it can be extended to L_2-norm loss function. The α formulation can now be written as :

$$\begin{aligned} \phi(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ - \frac{1}{4C^+} \sum_{i|y_i=+1} \alpha_i^2 - \frac{1}{4C^-} \sum_{i|y_i=+-} \alpha_i^2. \end{aligned} \tag{8}$$

3.2 Adaptive Margin (AMSVM)

In the SVM, the formulation of the margin error and of the support vectors, α_i are independent (see Eq. 6). Making the margin sensitive to each point was proposed by [5] called *Adaptive Margin* (AMSVM). The AMSVM can be thought of as a generalisation of the LOOM technique originated by [6]. The *Leave One Out Method* (LOOM) which minimises the bound of the error directly through slack variables, ξ, and is given as :

$$\text{minimise} \sum_{i=1}^{n} \xi_i \tag{9}$$

subjected to : $y_i f(\mathbf{x}_i) \geq 1 - \xi_i + \lambda \alpha_i k(\mathbf{x}_i, \mathbf{x}_i)$, $\xi_i \geq 0$ and $\alpha_i \geq 0 \ \forall_i = 1.....n$. The significance of the regulariser, λ, at each training point is :

- if λ=0, no adaptation is made on margins.
- if $\lambda \to \infty$, no effort has been made to reduce the emperical risk.
- if λ=1 , this is the LOOM SVM.

LOOM does not reduce the training errors but describes the number of training errors that are classified incorrectly even when they are removed from the linear combination of the decision boundary. This

Classification Performance		TP	TN	GMean
Approaches	Kernels			
SVM	Spline	0.36 $C^+ = 1000$	0.94 $C^- = 1000$	0.58
	RBF (σ=0.5)	0.39 $C^+ = 1000$	0.90 $C^- = 1000$	0.59
CSSVM With L1-norm Error	Spline	0.71 $C^+ = 1$	0.78 $C^- = 0.1$	0.74
	RBF (σ=0.5)	0.76 $C^+ = 1$	0.75 $C^- = 0.1$	0.75
CSSVM With L2-norm Error	Spline	0.40 $C^+ = 1000$	0.91 $C^- = 0.1$	0.60
	RBF (σ=0.5)	0.41 $C^+ = 10000$	0.89 $C^- = 1$	0.60
AMSVM	Spline	0.32	0.91	0.53 $\lambda = 0$
	RBF (0.5)	0.32	0.92	0.53 $\lambda = 0$

Table 1: Summary of the best test result by averaging the the set of five random selection sample. TP and TN are the true classification rate for "crack" and "no crack".

implies that the significance of α for a particular training point is dependent of its contribution to making the decision boundary. The AMSVM can be implemented by adding a regulariser term to the loss of margin in the constraint.

4 Model Specification

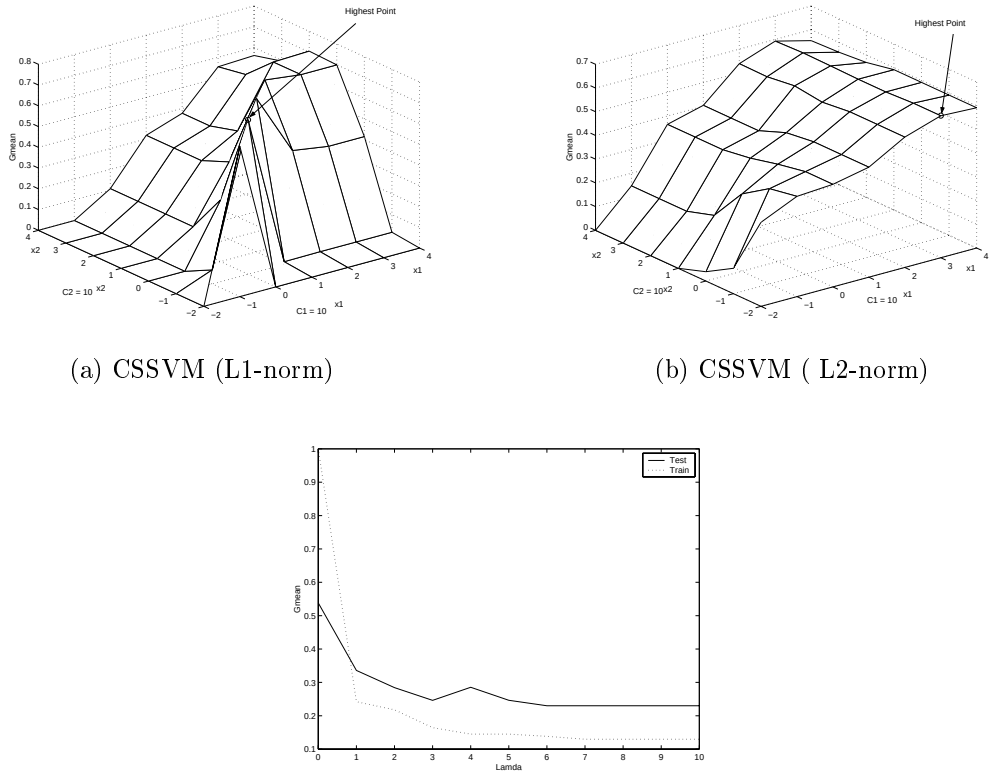
The ADI materials data set for the automotive camshaft application contains a total of 2923 examples of which 116 samples are crack initiation sites ("Crack" class) while 2807 samples did not act as crack initiation sites ("No Crack" class). These data were obtained from a FBT of ADI [7]. A set of nine measurements relating to the spatial distributions and measures of the object (graphite nodules) were obtained from the tessellation. This set of nine features describe 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 [8]. During these stages, noise in the background is attenuated and holes within bodies are filed. Then the features for each nodule for learning are generated from the following measurements : 1. Nodule area; 2. Nodule aspect ratio, 3. Nodule angle, 4. Cell area surrounding the nodule, 5. Local cell area fraction, 6. Number of near neighbours, 7. Nearest neighbour distance, 8. Mean near neighbour distance and Nearest neighbour angle. Note: that 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 significant for each feature. Upon normalising, the data needs to be partitioned for training and testing. Due to the extensive computation time required for large data, we consider a reduced data set which consisted of 700("no crack") and 90("crack") for training and the rest of the data for testing. This was repeated five times with random selection of the data each time.

5 Results

The type of Kernel was restricted to splines, radial basis functions (RBF), polynomials of order 3. The capacity control values were sample logarithmically on $[0.01, 10000]^2$. Better results may be possible with a finer and more intelligent sampling method. Several widths of RBF was used and the width of 0.5 was selected as it provided the best results. The Geometric mean was used to select a point on the ROC curve where the TP and TN are balanced.

Results obtained show that the SVM has high classification rate on the "correct crack" class. When the CSSVM was used, the TN was reduced by about 10% but the TP sensitivity increased by 50% (see table 1). This increased the overall performance of our model based on the geometric mean by 15% with both TN and TP being about 71%. C has a larger value in the SVM than in the CSSVM. This indicates that while using the SVM, the minimisation emphasises misclassification rather than maximising the margin. Notice that the misclassification cost between the crack and no crack were 0.1 and 1 respectively. This suggests that the ratio between the misclassification cost may be due to the imbalanced data that we have used for training. An RBF with width 0.5 and spline were investigated for the subsequent approaches. The results obtained for the SVM with L_1-norm and CSSVM with L_2-norm are very



(a) CSSVM (L1-norm)

(b) CSSVM (L2-norm)

(c) AMSVM

Figure 2: Summary of the result obtained from Spline kernel with CSSVM and AMSVM techniques based on the geometric mean performance criteria. a) & b) were obtained from CSSVM with L1-norm and L2-norm error, where C1 and C2 denote the "crack" and "no crack" misclassification cost respectively and its scale are in logarithmic increment . c) the training and testing result obtained from AMSVM. Note: the trend between the training and testing is similar.

similar. However, the ratio of the misclassifications is approximately 4. This could be due to the quadratic penalty used for the slack variable. Fig. 2 a and b show that surface plot of L_1-norm and L_2-norm are of the concave and hyperplane.

Rather than trying to control the capacity control more locally and imposing smoothness, the adaptive margin automatically adapts its margin for each data. What is left to tune in this approach is the tradeoff between smoothness (i.e. $\lambda > 0$) and the empirical risk imposed (i.e. λ=0). If λ=1, the data is equivalent to Leave-One-Out Method (LOOM) SVM. Figure 2c, shows the training and testing result obtained using different regularisers. Results show that for the case of the spline kernel, the training result obtained shows similar trends to that of the test result. As for the case of the RBF, the training performance became 0 at λ=4. This is very similar to the case of the spline kernel which becomes rather constant when it reaches λ=4. This shows that the decision boundary is at its flattest. However, the best test results were obtained when no adaptation of the margin was imposed on it (λ=0). This indicates that the empirical risk alone can be used to represent the true distribution of the data. As such, this technique of automatic adaptation of margins may not be appropriate when dealing with imbalanced data.

6 Conclusions

CSSVM and AMSVM, SVM extension techniques, have been used to explore the classification of an imbalanced real world data set. Results show that the CSSVM outperforms AMSVM. The use of different kernels shows no significant changes in results. The ratio between the misclassification cost seem to coincide with that of the training size used for each class. The future work will emphasis on using balanced data to investigate its effect and also as a means of tuning the parameters.

7 Acknowledgments

I would like to acknowledge the financial support from Federal Mogul Technology. I would like to thank Tony Dodd for valuable discussions.

References

[1] M. Kubat, R. Holte, and S. Matwin, "Learning when negative examples abound," *Machine Learning*, vol. 30, pp. 195–215, 1998.

[2] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[3] A. Smola, *Learning with Kernels.* PhD thesis, Technical University of Berlin, 1998.

[4] 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.

[5] R. Herbrich, *Learning Linear Classifiers - Theory and Algorithm.* Draft thesis, Technical University of Berlin, 2000.

[6] J. Weston, "Leave one out support vector machines," *Proccedings of the International Joint Conference on Artifical Intelligence, Sweden*, 1999.

[7] 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.

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