ROC Optimisation of Safety Related Systems

Jonathan E. Fieldsend **Richard M. Everson**[1]

Abstract. Many safety related and critical systems warn of potentially dangerous events; for example the Short Term Conflict Alert (STCA) system warns of airspace infractions between aircraft. Although installed with current technology such critical systems may become out of date due to changes in the circumstances in which they function, operational procedures and the regulatory environment. Current practice is to 'tune' by hand the many parameters governing the system in order to optimise the operating point in terms of the true positive and false positive rates, which are frequently associated with highly imbalanced costs.

In this paper we cast the tuning of critical systems as a multi-objective optimisation problem. We show how a region of the optimal receiver operating characteristic (ROC) curve may be obtained, permitting the system operators to select the operating point. We apply this methodology to the STCA system, using a multi-objective $(1 + 1)$-evolution strategy, showing that we can improve upon the current hand-tuned operating point as well as providing the salient ROC curve describing the true-positive versus false-positive trade-off.

1 INTRODUCTION

Many safety related systems can be regarded as two-class classifiers: they classify a particular set of inputs or features into classes that might be labelled *dangerous* and *benign*. Classifications into the *dangerous* class raise an alarm and generally require some sort of human intervention. The specific example with which this paper is concerned is the Short Term Conflict Alert (STCA) system in operation in the United Kingdom and elsewhere. STCA monitors aircraft locations from ground radar and provides advisory alerts to air traffic controllers if a pair of aircraft are likely to become dangerously close. The STCA system is designed to raise a warning to air traffic controllers if there is a developing conflict between aircraft, giving them time to redirect the aircraft.

Taking its input from ground radar, the STCA system is independent of the aircraft, and cannot know the intentions of the pilots or air traffic controllers who may be aware of a potential conflict and already taking measures to avoid it. For this reason, and because STCA must make conservative predictions, there are necessarily *nuisance alerts* as well as *genuine alerts*. There is clearly a trade-off between genuine and nuisance alerts and it is desirable to minimise the number of nuisance alerts in order to maintain the air traffic controllers' confidence in STCA. Regarding STCA as a two-class classifier, which partitions pairs of radar tracks into *dangerous* or *serious* and *benign* classes, allows ROC analysis to be applied in which *genuine alerts* are true positives, while *nuisance alerts* are false positives.

[1] Department of Computer Science, University of Exeter, Exeter, UK email: {J.E.Fieldsend,R.M.Everson}@exeter.ac.uk

The STCA system became operational for part of UK airspace in 1988 [2] and versions capable of coping with complex terminal control airspaces have been in operation since 1994. Since its introduction there have been incremental changes to the software and it is now used across the UK and elsewhere. Importantly, however, there have been changes in the volume and nature of air traffic together with changes to the management of the airspace monitored by STCA. Bringing new software into service involves a lengthy period of testing and scrutiny, even for *advisory* systems such as STCA; consequently, staff at the National Air Traffic Services (the principal civil air traffic control service for the United Kingdom, NATS) undertake parameter reviews in which they adjust (*tune*) the operating parameters of the STCA system in order to reduce the number of nuisance alerts, while maintaining the genuine alerts. This tuning is performed on the basis of a large (170 000) database of track pairs containing historical and recent encounters. The great number of parameters (at least 1500) determining the behaviour of STCA make tuning a highly skilled and labourious business. However, despite a recent step towards automation [2], the optimal receiver operating characteristics of the STCA system have not been known.

In this paper we introduce an approach to resolving these optimisation problems using multi-objective optimisation techniques based on evolutionary algorithms [6, 14, 23]. We cast the true positive and false positive rates obtained by STCA as two opposing objectives to be maximised and minimised respectively. This allows us to obtain the optimal ROC curve from which the operating point can be chosen with a full knowledge of the trade-off between genuine versus nuisance alert rates.

In section 2 we describe the Short Term Conflict Alert system used in the UK; and in section 3 we describe the current optimisation process of STCA within the UK air traffic service, together with previous attempts at the automation of its optimisation. In section 4 we discuss the relation of ROC analysis to the more general theory of Pareto optimality; based on this, in section 5 we describe the multi-objective optimisation technique approach to discovering the ROC curve for the system, and provide results in section 6. The paper concludes with a discussion in section 7.

2 THE SHORT TERM CONFLICT ALERT SYSTEM

Here we focus on the Short Term Conflict Alert system (STCA) which is used widely within Europe by civil aviation authorities, in order to alert air traffic controllers to potential airspace infringements by aircraft pairs (i.e., two aircraft which may become too close). STCA is not strictly a safety critical system—*a system containing computer, electronic or electromechanical components whose failure may cause threat to life and limb or severe damage to prop-*

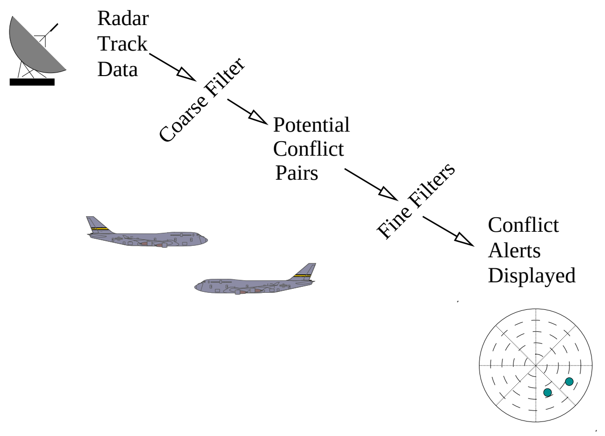

Figure 1. High level view of the STCA model.

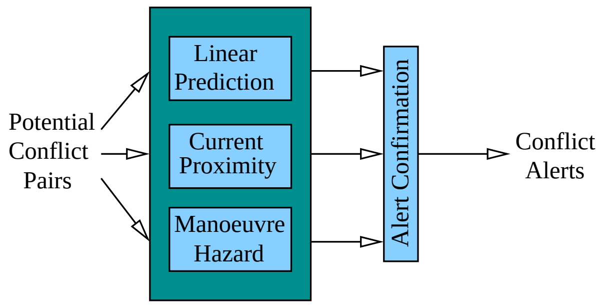

Figure 2. STCA fine filters.

erty[2]—but rather a component of the NATS ‘safety net’, providing *advisory* alerts to air traffic controllers of potential airspace proximity violations. Nonetheless, it exhibits many of the characteristics of a safety critical system: it must be highly reliable, transparent and verifiable. Its importance is highlighted by the fact that it is thought that one of the factors contributing to the midair collision over the border between Germany and Switzerland in July 2002 was that the STCA system in the relevant Swiss control station was switched off for maintenance.[3]

2.1 Overview

Figures 1 and 2 give an overview of the operation of the STCA system, which incorporates a highly complex and proprietary algorithm. Ground radars track the aircraft in a given airspace and those adjoining, and every four seconds (a STCA cycle) create *track pairs* of all possible combinations of aircraft. A *coarse* filter (Figure 1) is used first to remove all those pairs which are simply too far away from each other to be of concern. Potential conflict pairs are then processed in the core of STCA by three *fine* filters: a mixture of three models; a *linear prediction* filter; a *current proximity* filter; and a *manoeuvre hazard* filter (Figure 2). The boolean outputs of these fine filters are combined by the *alert confirmation module*, and aircraft pairs which are in danger of becoming too close are highlighted and alerted on the air traffic controllers’ screens. The STCA is concerned with detecting airspace conflicts that may occur in the near future (around two minutes), so that air traffic controllers may be warned and the situation rectified in sufficient time.

The minimum separation that is counted as an air proximity conflict depends on a number of criteria (for example, the airspace location and available radar cover). Generally in the UK in controlled airspace it ranges between 3, 5 or 10 nautical miles horizontally and 1000ft vertically. The linear prediction filter checks for loss of horizontal or vertical separation assuming that the aircraft continue on in a straight line at their current headings and speeds. The current proximity filter merely checks for a current loss of separation and the manoeuvre hazard filter classifies potential conflicts when either or both of the aircraft are turning. The combination of the binary classifications from the three fine filters by the alert confirmation module (Figure 2) is relatively sophisticated. During the confirmation process alerts from a track pair are checked within a moving time window, and if they are in conflict for a number of successive radar cycles (typically two or three), then an alert is passed onto the controller, although alerts from the current proximity filter are relayed more rapidly.

2.2 Parameterisation

Each portion of the UK airspace is marked as distinct *region types*. For instance *en route* describes the airspace between airports, while regions where aircraft circle until permission is given to land are designated as *stack*. Since aircraft in different region types tend to have different types of flight behaviour, separate parameter sets are used for each one of the region types. The particular parameter set used for classifying a track pair therefore depends upon the region types of the two aircraft; additional rules are used to determine the relevant parameter set if the aircraft have different region types.

The busy airspace above London, with which this study is concerned, is divided into 16 of these different region types. This multiplicity of parameter sets leads to a great number of parameters that can be adjusted to affect the performance of the STCA system. There are 96 parameters pertaining to the three fine filters, which means that the system uses approximately 1550 parameters (the coarse filter using fewer than 20). Note that it is not feasible to adjust the parameters for the filters of each region type independently of the other region types, because track pairs involving pairs of regions lead to significant interactions between the parameters of different region types. On the other hand, as we describe below, only approximately two thirds of the available parameters are routinely adjusted.

The three central components of STCA are readily understood and their operation is capable of verification by practitioners, which is a common feature of the majority of critical systems in use. Regulatory authorities are very uneasy about using black-box techniques, such as artificial neural networks, in which function mappings are not easily described or understood. As we have described, the filter components of the STCA system themselves do, however, possess a large number of user determined parameters, which affect the operation of the system and therefore whether or not the system alerts pairs as being in potential conflict. The STCA program may be thought of as a decision tree, particular branches of which are followed depending

[2] The working definition adopted by an ACM Special Interest Group on Computer-Human Interaction (SIGCHI) workshop [20] and typical of definitions of safety critical systems.

[3] See, for example, `http://news.bbc.co.uk/1/hi/world/europe/2082331.stm` and `http://aviation-safety.net/database/2002/020701-0.htm`.

upon the aircraft track pair being processed and the thresholds which are determined by the operational parameters of STCA. Note that the operational parameters affect the classification produced by altering the thresholds and model parameters; changes in parameter values do not affect the logical routes that may be taken through the decision tree. The logical structure of the program is incrementally altered by NATS Operational Analysis & Support group as new versions of the software are introduced. However, routine tuning of the system does not affect the logical structure.

3 OPTIMISATION OF STCA

The STCA system is in operation in the four UK air traffic control centres and at other air traffic control centres in Europe, so appropriate parameter setting must be chosen for each particular locale. Moreover, changes in the volume of air traffic, changes in local air traffic operational procedures and changes in the regulatory environment mean that the STCA operational parameters must be reviewed and updated in order to prevent the system becoming out of date. In the UK all serious near-miss encounters are reviewed under the auspices of the Airprox Board (see for example [5]). In addition NATS regularly assesses the efficacy of the STCA system by running an off-line version with a database comprised of recent general traffic encounters together with historical serious encounters. The two samples permit the nuisance alert rate for general traffic to be monitored together with the warning time provided for genuine alerts.

Table 1. Encounter categories used by NATS.

C	Alert	Description
1	Necessary	Serious or potentially serious encounter with a significant collision risk for which alerts and additional warning time are considered highly desirable.
2	Desirable	Serious encounters, which involved an actual or potential loss of separation, but little risk of collision, where alerts and additional warning time are considered desirable.
3	Unnecessary	Level off with risk encounters where a standard level off prevented a conflict. The desirability of alert for these encounters is dependent on where (and to some extent when) they occur. In busier airspace, such as stacks, they may be seen as an unnecessary distraction. Whereas in some less busy areas of airspace they may be seen as a valuable safety net (some controllers may reaffirm level off instructions when STCA indicates that a level bust would lead to conflict).
4	Undesirable	No actual or potential conflict. An alert would be considered a nuisance.
5	Bad data	Bad data for which alerts are generally considered a nuisance but are commonly deemed beyond the remit of STCA and therefore not usually taken into account during a parameter review.

3.1 Manual Optimisation

As shown in Table 1, each encounter is categorised by NATS staff into one of five categories of diminishing severity; category 4 encounters are semi-automatically categorised, but all others are manually annotated. Note that without knowledge of a pilot's intentions or the instructions a pilot has received, it is very difficult to predict whether an ascending or descending aircraft will level off at a specified height or 'bust' through the level potentially leading to a conflict. Errors in predicting level off clearly lead to nuisance alerts and as such we count category 3 encounters as false positives.

STCA performance on the database is assessed using the Conflict Alert Management Performance Analysis Package (CAMPAP), which runs the STCA system on the database and analyses the performance for each category in each region [3, 4]. Using CAMPAP, the Operational Analysis & Support group within NATS has over the last 10 years, through manual adjustment of the parameters, tuned STCA to achieve the best balance between genuine and nuisance alerts. In essence this has been achieved by skilled staff running different parameter settings through the CAMPAP simulation, by changing one or more of the values in current use, and assessing the performance on the collated data.

As iterative evolution of the STCA system has occurred, and the airspace in the UK is partitioned into ever more disparate region types, this task clearly becomes more arduous. As an indication of the increasing complexity it may be noted that since the work of Beasley *et al.* [2] in 2002 the increase in the number of fine filter parameters and regions has led to an increase of roughly 500 in the number of STCA parameters.

3.2 Weighted Objective Optimisation

Beasley *et al.* [2] realised that the current approach of tweaking the system variables by hand may be suboptimal, and so applied the tabu search heuristic in an attempt to automate the process. This approach generated a objective term to be maximised, which was a composite weighting of the correct alerts gained and lost and the spurious alerts gained and lost in comparison with the base parameter set. (Warning time gains and losses were also included.)

The problem inherent in optimising composite weightings of objectives is to choose *a priori* that weighting which will provide the desired operating point (model). Assume that we are concerned with two objectives $T(\boldsymbol{\theta})$ (true positive rate) and $F(\boldsymbol{\theta})$ (false positive rate) both of which are dependent on a vector of parameters $\boldsymbol{\theta}$. We assume that $T(\boldsymbol{\theta})$ is to be maximised and $F(\boldsymbol{\theta})$ is to be minimised. If, for instance, we wanted a trade-off between the two we may form the composite objective $c(\boldsymbol{\theta})$:

$$c(\boldsymbol{\theta}) = \beta T(\boldsymbol{\theta}) - \gamma F(\boldsymbol{\theta}). \qquad (1)$$

Here the ratio of β and γ give some preference information, and may be interpreted as the relative costs of making true positive and false positive classifications. The consequence of optimising this type of composite function is that (as shown by the iso-cost contours in Figure 3) models with widely varying T and F values are actually deemed equivalent. The final operating point returned under such an optimisation process is that point on the ROC curve where the gradient is equal to the ratio of β and γ.

Figure 3 illustrates the problem with equal *a priori* weighting ($\beta = \gamma$). Three different underlying ROC curves are shown, which lead to three very different operating points and model properties

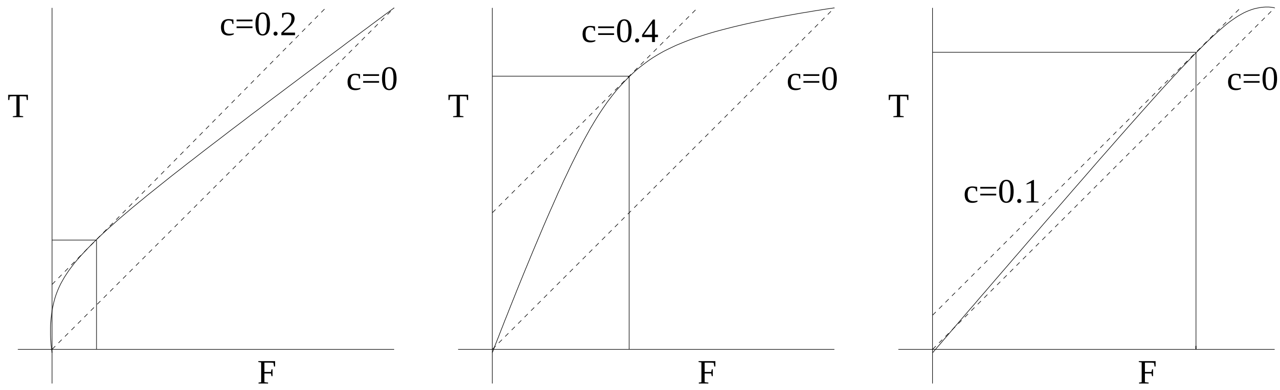

Figure 3. Problems with composite weighting. *Left*: ROC shape means equal weighting of objectives leads to model T three times larger than F. *Middle*: Returned T twice as large as F. *Right*: Returned T only slightly larger than F.

when the composite objective (1) is used. The dashed lines at 45° indicate the points in T–F space where all $c(\boldsymbol{\theta})$ values are equivalent. The higher the dashed line therefore the larger $c(\boldsymbol{\theta})$ and the better the operating points which lie on it. The point where one of these parallel lines is tangential to the ROC curve is the model that maximises $c(\boldsymbol{\theta})$ and is therefore the model returned by the optimisation using composite weighting (assuming the optimisation method is good enough to locate the optimum).

The left plot of Figure 3 illustrates a situation in which optimising Equation 1 results in a model for which the true positive rate is three times the false positive rate $(T(\boldsymbol{\theta}) = 3F(\boldsymbol{\theta}))$. The middle plot of Figure 3 illustrates a situation where optimising Equation 1 results in a model with $T(\boldsymbol{\theta}) = 2F(\boldsymbol{\theta})$ and the right plot in Figure 3 illustrates a situation in which optimising Equation 1 results in a model with $T(\boldsymbol{\theta}) = 1.1F(\boldsymbol{\theta})$. Extensive examples and mathematical proofs of these issues can be found in [7].

The tabu search optimiser [2] was also found to be susceptible to trapping in local minima and required manual analysis of the parameter space to re-start the search. Perhaps in the light of these considerations, the original iterative person-based adjustment is still in use by NATS.

4 ROC ANALYSIS & PARETO OPTIMALITY

If we wish to satisfy the two opposing objectives of true positive maximisation and false positive minimisation, when the classes are skewed and the costs imbalanced it does not make sense to try and optimise a single objective function as illustrated in the previous section. If the costs of an incorrect classification were known the expected cost for any parameter set could be calculated [10] and used as a single objective function [15]. However, this procedure requires accurate specification of the misclassification costs which are seldom accurately known; indeed it is often desirable to present the user with a ROC curve from which the best operating point can be selected. A common method is to employ the Neyman Pearson criterion: a maximum false-positive rate is specified, which then determines the true-positive rate.

Alternatively, some other summary measure of the ROC curve, such as the area under the ROC curve (AUROC) could be used as a measure of the quality of a set of parameters [9, 16]; this overall measure could then be used as an objective to be optimised with respect to the system parameters.

Of course, all these measures based upon the ROC curve require knowledge of the ROC curve, which hitherto has been unavailable for the STCA system. In this section we show how multi-objective evolutionary algorithms (MOEAs) may be used to derive the ROC curve for the STCA system. However, we take the view that summarising the ROC curve neglects the true value of the curve, namely providing the user with an analysis of the trade-offs inherent in choosing an operating point. In this manner we can entirely circumvent the problematic *a priori* setting of objective weights.

4.1 The ROC curve and Pareto optimality

In general we consider a classifier $g(\mathbf{x};\boldsymbol{\theta})$ which gives an estimate of the probability that a feature vector $\mathbf{x}$ belongs to one of two classes. We assume that the classifier depends upon a vector of adjustable parameters $\boldsymbol{\theta}$, and we denote by $T(\boldsymbol{\theta})$ the classifier's true positive classification rate (measured on a particular dataset of interest), while the false positive rate is denoted by $F(\boldsymbol{\theta})$.

A ROC curve is frequently obtained by varying the probability threshold separating the two classes. As the threshold is varied from zero to one a non-decreasing ROC curve in the (F, T) plane is obtained for any particular fixed set of parameters, and different ROC curves are obtained for different parameters. In this work, we consider the classification threshold to be subsumed in the parameter vector and seek to discover the set of parameters (including threshold) that simultaneously minimise $F(\boldsymbol{\theta})$ and maximise $T(\boldsymbol{\theta})$. In fact, the STCA classifier is a hard classifier, yielding only a binary classification rather than an estimate, however imprecise, of the probability of class membership. Nonetheless, we may still seek the set parameter values that yield the optimal true-positive versus false-positive trade-offs. (See, for example, [12] for extensive discussions of ROC curves for hard and soft classifiers.)

A general multi-objective optimisation problem seeks to simultaneously extremise D objectives:

$$y_i = f_i(\boldsymbol{\theta}), \qquad i = 1, \ldots, D \tag{2}$$

where each objective depends upon a vector $\boldsymbol{\theta}$ of P parameters or decision variables. It is convenient to assume that all the objectives are to be minimised, so for the STCA system we minimise the pair of objectives $(-T(\boldsymbol{\theta}), F(\boldsymbol{\theta}))$. The parameters may also be subject to the J constraints:

$$e_j(\boldsymbol{\theta}) \geq 0, \qquad j = 1, \ldots J \tag{3}$$

so that the multi-objective optimisation problem may be expressed as:

$$\text{minimise} \quad \mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_D(\boldsymbol{\theta})) \tag{4}$$
$$\text{subject to} \quad \mathbf{e}(\boldsymbol{\theta}) = (e_1(\boldsymbol{\theta}), \ldots, e_J(\boldsymbol{\theta})) \geq 0 \tag{5}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ and $\mathbf{y} = (y_1, \ldots, y_D)$.

When faced with only a single objective an optimal solution is one which minimises the objective given the model constraints. However, when there is more than one objective to be minimised solutions may exist for which performance on one objective cannot be improved without sacrificing performance on at least one other. Such solutions are said to be *Pareto optimal* [6, 14, 23] after the 19th century engineer, economist and sociologist Vilfredo Pareto, whose work on the distribution of wealth led to the development of these trade-off surfaces [21]. The set of all Pareto optimal solutions is said to form the Pareto front.

The notion of *dominance* may be used to make Pareto optimality clearer. A decision vector $\boldsymbol{\theta}$ is said to *strictly dominate* another $\boldsymbol{\phi}$ (denoted $\boldsymbol{\theta} \prec \boldsymbol{\phi}$) iff

$$\begin{aligned} f_i(\boldsymbol{\theta}) \leq f_i(\boldsymbol{\phi}) \quad \forall i = 1, \ldots, D \quad &\text{and} \\ f_i(\boldsymbol{\theta}) < f_i(\boldsymbol{\phi}) \quad \text{for some } i&. \end{aligned} \tag{6}$$

Less stringently, $\boldsymbol{\theta}$ *weakly dominates* $\boldsymbol{\phi}$ (denoted $\boldsymbol{\theta} \preceq \boldsymbol{\phi}$) iff

$$f_i(\boldsymbol{\theta}) \leq f_i(\boldsymbol{\phi}) \quad \forall i = 1, \ldots, D. \tag{7}$$

A set of M decision vectors $\{\boldsymbol{\theta}_i\}$ is said to be a *non-dominated set* if no member of the set is dominated by any other member:

$$\boldsymbol{\theta}_i \not\prec \boldsymbol{\theta}_j \quad \forall i, j = 1, \ldots, M. \tag{8}$$

A solution to the minimisation problem (4) is thus *Pareto optimal* if it is not dominated by any other feasible solution, and the non-dominated set of all Pareto optimal solutions is the Pareto front. Recent years have seen the development of a number of evolutionary techniques based on dominance measures for locating the Pareto front; see [6, 8, 23] for recent reviews.

5 OPTIMISATION USING MULTI-OBJECTIVE EVOLUTIONARY ALGORITHMS

Anastasio, Kupinski & Nishikawa introduced the use of multi-objective evolutionary algorithms (MOEAs) to optimise ROC curves, illustrating the method on a synthetic data [19] and for medical imaging problems [1]. Here we used a similar methodology, albeit with improved convergence properties.

The multi-objective evolutionary algorithm used in this study is a stochastic search algorithm, based on a simple $(1 + 1)$-evolution strategy (ES), similar to that introduced in [17]. In outline, the procedure for locating the Pareto front/ROC curve, operates by maintaining an archive, A, of mutually non-dominating solutions, $\boldsymbol{\theta}$, which is the current approximation to the Pareto front/ROC curve. At each stage of the algorithm some solutions in A are copied and perturbed. Those perturbed solutions that are dominated by members of A are discarded, while the others are added to A and any dominated solutions in A are removed. In this way the estimated Pareto front A can only advance towards the true Pareto front. This algorithm, unlike earlier versions [17], maintains an archive which is unconstrained in size, permitting better convergence properties [13].

Algorithm 1 describes in more detail the algorithm as applied to the optimisation of the STCA system. Following the current operating practice of NATS and [2], we choose to optimise only 912 of the > 1500 parameters affecting the STCA system; these parameters are those parameters which have different values in different regions after tuning by NATS. Furthermore we restrict these parameters to the ranges over which they are adjusted by NATS.

Algorithm 1 A MO $(1 + 1)$-ES for STCA optimisation.

Inputs:
N Number of ES generations

1: $A :=$ initialise()
2: $n := 0$
3: `while` $n < N$:
4: $\quad \boldsymbol{\theta} :=$ select(A)
5: $\quad \boldsymbol{\theta}' :=$ perturb$(\boldsymbol{\theta})$
6: $\quad (T(\boldsymbol{\theta}'), F(\boldsymbol{\theta}') :=$ STCA$(\boldsymbol{\theta}')$
7: $\quad$ `if` $\boldsymbol{\theta}' \not\preceq \boldsymbol{\phi} \; \forall \boldsymbol{\phi} \in A$:
8: $\qquad A := \{\boldsymbol{\phi} \in A \,|\, \boldsymbol{\phi} \not\prec \boldsymbol{\theta}'\}$
9: $\qquad A := A \cup \boldsymbol{\theta}'$
10: $\quad$ `end`
11: $\quad n := n + 1$
12: `end`

The archive or frontal set A is initialised by drawing parameters for the STCA system uniformly from their feasible ranges; in addition the current 'best' parameter set from manual tuning $\boldsymbol{\theta}^\star$ is added to A. Of course many of these randomly selected parameter vectors are dominated by other parameter vectors and these dominated parameters are deleted from A so that A is a non-dominated set (8). In fact, in the work reported here, we found that of 100 randomly initialised parameters only $\boldsymbol{\theta}^\star$ and one other parameter vector remained in A after dominated parameters were removed.

Following initialisation, the loop on lines 4–11 of Algorithm 1 is repeated for the desired number of iterations. At each iteration a single parameter vector $\boldsymbol{\theta}$ is selected from A; selection may be uniformly random, but partitioned quasi-random selection [13] was used here to promote exploration of the front. The selected parent vector is perturbed to generate a single *child* (line 5). Each individual parameter in the parent vector is perturbed with equal probability (0.2 here); the perturbations themselves are made by adding a random number to the parent parameter value. Yao *et al.* [24] have shown that perturbations drawn from heavy-tailed distributions facilitate convergence by promoting exploration and we draw perturbations from a Laplacian density, $p(x) \propto e^{-|x/w|}$, whose width is set equal to one tenth the feasible range of the parameter being perturbed; perturbations that lie outside the feasible range are resampled.

The true $T(\boldsymbol{\theta}')$ and false $F(\boldsymbol{\theta}')$ positive rates for the perturbed vector are evaluated by running the STCA/CAMPAP system with parameters $\boldsymbol{\theta}'$ on the test database of track pairs (Table 1). Following NATS practise, we consider category 1 and 2 alerts to be true positives, while category 3 and 4 alerts are treated as false positives. Category 5 alerts are ignored. If the child $\boldsymbol{\theta}'$ is not dominated by any of the parameter vectors in A, any parameter vectors in A that $\boldsymbol{\theta}'$ dominates are deleted from the archive (line 8) and $\boldsymbol{\theta}'$ is added to A (line 9). These two steps ensure that A is always a non-dominated set whose members dominate any other solution encountered thus far in the search.

In a $(\mu + \lambda)-$ES, μ parameter vectors are perturbed to generate λ new vectors. That is, μ parameter vectors are selected (whose

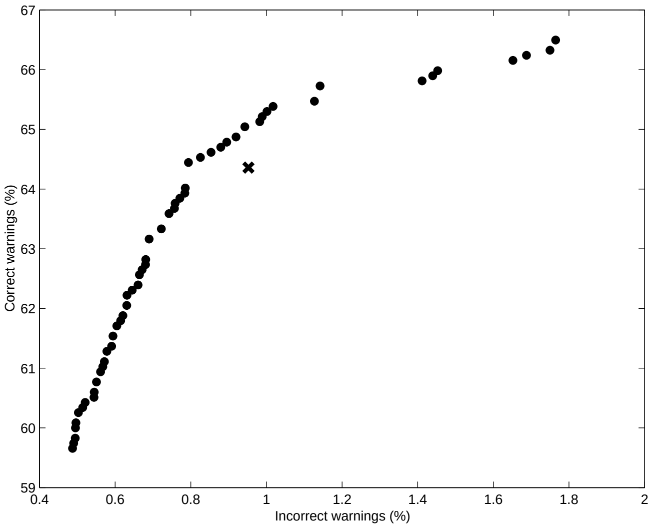

Figure 4. Dots show estimates of the Pareto optimal ROC curve for STCA obtained after 5000 evaluations of the $(1+1)$-ES multi-objective optimiser. The cross indicates the manually tuned operating point $\boldsymbol{\theta}^{\star}$.

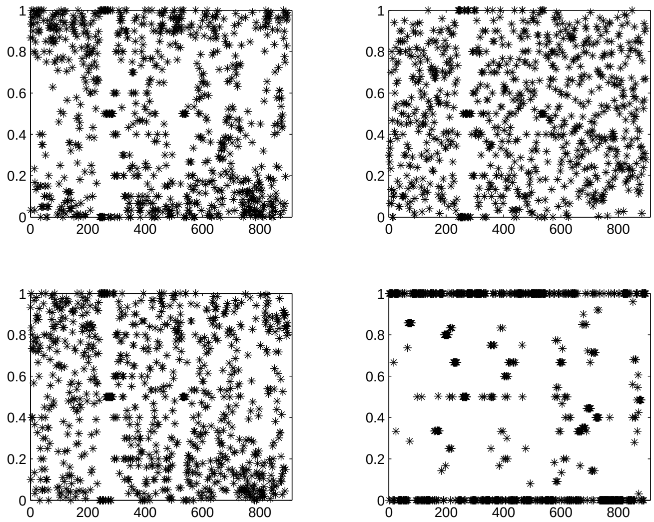

Figure 5. Normalised parameter values for operating points on the Pareto optimal ROC curve shown in Figure 4. Panels correspond to parameters for: low F, low T (bottom-left); medium F, medium T (top-left); high F, high T (top-right); and the manually tuned operating point $\boldsymbol{\theta}^{\star}$ (bottom-right).

performances have already been evaluated); these *parents* are copied and have their parameter values perturbed in order to generate λ *children*. Optimisation schemes with $\lambda > 1$ are attractive because the evaluation of the children may be performed in parallel. The computational cost of evaluating a single set of STCA parameters within CAMPAP is fairly high, at approximately 5 minutes. However, the system is written in a proprietary variant of PASCAL, which necessitates it be run on a Compaq Alpha machine. Since only a single Alpha was available to us, we used a $(1+1)$-ES, which has been shown to perform well compared to $(\mu+\lambda)$ MOEA implementations [18].

6 RESULTS

In this paper we present an initial conservative application of the MOEA method to STCA optimisation. It is conservative in that the ranges of parameters to be varied are defined by the current ranges of that parameter across the 16 region types within the currently applied STCA parameterisation of NATS. This means effectively we are only concerned with adjusting $2/3$ of the model parameters (still a significant number!), and the parameters are confined to regions of decision space with which personnel at NATS have considerable experience.

We optimised the true and false positive rates for a database comprised of manually and semi-automatically categorised encounters. The database included historical track pairs leading to serious or potentially serious encounters together with general traffic track pairs from two weeks in 2001.

Even this conservative optimisation approach produces some striking results. Figure 4 shows the estimates of the Pareto optimal ROC curve obtained using the multi-objective optimiser after $N = 5000$ evaluations (approximately 10 days computation). The current NATS operating point is also plotted as a cross. The optimisation has located an ROC curve consisting of 58 points ranging from 59.5% to 66.5% true positive and 0.5% to 1.8% false positive. In addition the manually tuned STCA operating point $\boldsymbol{\theta}^{\star}$ lies *behind* (is dominated by) several operating points on the estimated ROC curve. Although the improvement over $\boldsymbol{\theta}^{\star}$ is relatively small in percentage terms, the quantity of track pairs processed by the STCA system means that a significant reduction in the *number* of false alerts could be achieved while maintaining the current genuine alert rate. We regard as more important, however, the production of the ROC curve itself, because it reveals the true-positive versus false-positive trade-off permitting the operating point to be chosen. In fact it may be observed that the current operating point $\boldsymbol{\theta}^{\star}$ is close to the corner of the Pareto optimal curve. Choosing an operating point to the left of the corner would result in a rapidly diminishing genuine alert rate for little gain in the nuisance alert rate; whereas operating points to the right of the corner provide small increases in the true positive rate at the expense of relatively large increases in the false positive rate.

Figure 5 gives an indication of how the 912 parameters which could be altered during the optimisation vary as the Pareto front is traversed. Each of the four panels in figure 5 shows the 912 variable parameters, each normalised to the interval $[0, 1]$, so that 0 represents the minimum value it was permitted to assume during optimisation and 1 represents the maximum. The bottom-right panel shows the parameters at the manually tuned operating point $\boldsymbol{\theta}^{\star}$; many of the parameters are at their extreme values because we choose the allowable ranges to be defined by the extremal values located by NATS manual optimisation. There is a resemblance between these parameters and the parameters corresponding to the middle of the Pareto front ($F = 0.79\%$, $T = 64.44\%$) shown in the top-left panel. The bottom-left and top-right panels show $\boldsymbol{\theta}$ corresponding to the extreme ends of the front. Although there is a resemblance between the parameters for the bottom-left end of the front ($F = 0.49\%$, $T = 59.66\%$) and the middle, the solutions at the top-right end ($F = 1.76\%$, $T = 66.50\%$) appear to have a qualitatively different character. Indeed, we observed that solutions near the left and middle of the front resulted from perturbations to $\boldsymbol{\theta}^{\star}$, whereas solutions on the right of the front emanated from the other non-dominated solution discovered during the archive initialisation.

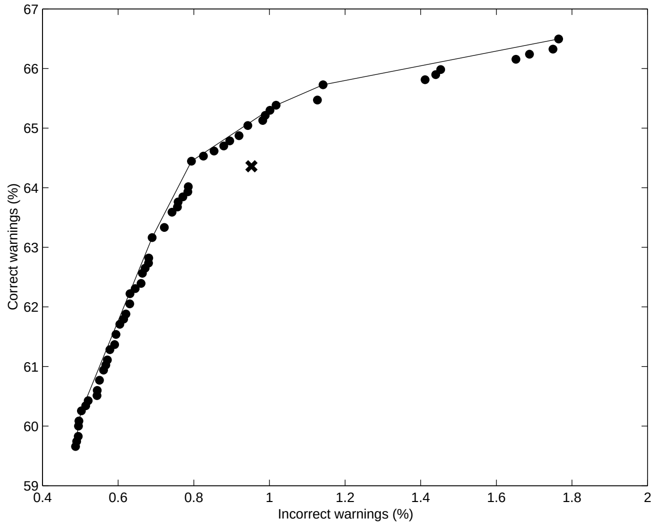

Figure 6. Convex hull of the Pareto optimal ROC curve shown in Figure 4.

7 DISCUSSION

We have presented a straightforward multi-objective optimisation scheme for locating the optimal ROC curve for the Short Term Conflict Alert system employed to give warning of potential breaches in air proximity by aircraft. The results show that parameters yielding a range of genuine and nuisance alert rates are located by the MOEA, thus revealing the genuine versus nuisance alert trade-off and permitting the operating point to be set with explicit knowledge of the trade-off. The idea of dominance is essential to the simultaneous optimisation of both true and false positive alert rates and it is interesting to note that the manually tuned operating point is dominated by several of the solutions found by multi-objective optimisation.

The production of the front took approximately ten days of computer time. However, we emphasise that this was *unattended* computer time, in contrast to the labour-intensive and skilled process by which STCA systems are currently optimised. We anticipate that once an optimised ROC curve has been located for a particular STCA system and database, the subsequent optimisation following incremental incorporation of new cases into the database will be much faster. More rapid optimisation schemes are readily implemented via $(\mu + \lambda)$-ES, which are amenable to coarse parallelisation.

It should be emphasised that the true and false positive alert rates were evaluated on a database of over 170 000 track pairs, consisting of historical alerts deemed to be serious and two weeks worth of relatively current data, this comprises the same database that is currently used for manual tuning of operational STCA systems for the London sector airspace. It is important current work for skilled staff to inspect the parameter values obtained.

It will also be important to analyse the robustness of the optimised solutions, especially in respect to changes in the data. We remark that bootstrap resampling methods [11] provide a principled way of evaluating the robustness which is readily implemented in an automated system such as this.

The optimisations reported here were conservative in that they optimised only the 900 or so parameters that are routinely adapted by NATS, and these parameters were restricted to the ranges used by NATS. Although, as Figure 5 shows, solutions on the front are obtained for parameter values lying between the extremes used by NATS, we look forward to optimising a larger number of parameters and to permitting the parameters to vary over broader ranges.

Although we have explored the genuine alert versus nuisance alert trade-off here, we remark that the multi-objective optimisation methodology is readily applied to trade-offs in three or more objectives, perhaps providing a weapon with which to attack the open question of ROC analysis for classification into more than two classes. In the STCA context a relevant third variable to be optimised is the warning time given by the STCA system of a potential air proximity breach. It is clearly important to provide as much advance warning as possible, without sacrificing the false positive rate; indeed it is possible that the improvements in false positive rate for portions of the front shown in Figure 4 are at the expense of reduced warning times for the genuine alerts. Multi-objective algorithms provide a technique for exploring this fully three-dimensional trade-off surface. It will also be interesting to explore the trade-off surface between the classification rates for the four categories of alert shown in Table 1, rather than combining categories 1 & 2 and 3 & 4 as we have done here.

The Pareto front located by the MOEA is comprised of a discrete set of parameter vectors at which the STCA system could be operated. However, we point out that the work of Scott *et al.* [22] shows that by randomly combining classifiers any operating point on the convex hull of the ROC curve may be obtained (see Figure 6). Indeed it is apparent that if the objectives to be optimised are statistical expectations, then Scott *et al's* work may be readily extended to three or more objectives to obtain an operating point on the convex hull of optimised solutions in many dimensions. It should be noted, however, that although the probabilistic combination of classifiers may lead to provably better average operating points, there are potential legal and ethical ramifications.

In this paper we have focused on the STCA system as an example safety related system; however, the STCA/CAMPAP system is treated purely as a subroutine of our evolutionary algorithm. Indeed in our implementation, the STCA/CAMPAP programs run on a separate computer. This 'wrapping' of the system to be optimised is important for two reasons. First, it shows that the technique is applicable to any critical system whose operating point is dependent on parameters that must be tuned and whose performance can be automatically evaluated. Second, and more importantly for safety-related systems, the wrapped system has not been modified in any way, thus preserving its integrity and the integrity of any safety case constructed for it.

Finally we remark that the majority of the parameters in the STCA filters have direct physical or mechanical interpretation, and that the transparency of the classification process is an important component in assuring the safety case for STCA. However, whether tuned by hand or optimised by a machine algorithm, the operational parameters are inferred from data and we look forward to the construction of safety cases for purely statistical classifiers whose operational parameters are inferred from data and have no ready physical interpretation.

ACKNOWLEDGEMENTS

We would like to express our thanks to Rod Bacon, Hellen Finney and Katherine Marren from the National Air Traffic Service Operational Analysis & Support group. We thank Trevor Bailey, Adolfo Hernandez, Wojtek Krzanowski, Derek Partridge, Vitaly Schetnin and Jufen Zhang for their helpful comments. We gratefully acknowledge the support of EPSRC grant GR/R24357/01.

References

[1] M. Anastasio, M. Kupinski, and R. Nishikawa:, 'Optimization and FROC analysis of rule-based detection schemes using a multiobjective approach', *EEE Transactions on Medical Imaging*, **17**, 1089–1093, (1998).

[2] J.E. Beasley, H. Howells, and J. Sonander, 'Improving short-term conflct alert via tabu search', *Journal of the Operational Research Society*, **53**, 593–602, (2002).

[3] G. Beeston, H. Howells, and M. Richards, *PAP User Guide*, National Air Traffi c Services Ltd., Research and Development Group, Department of Technical Research and Development, 1.0 edn., August 2000.

[4] G. Beeston, H. Howells, and M. Richards, *Software Description for CAMPAP (Issue 1.03) Module*, National Air Traffi c Services Ltd., Research and Development Group, Department of Technical Research and Development, 1.0 edn., August 2000.

[5] UK Airprox Board, 'Analysis of Airprox in UK Airspace (July 2002 to December 2002)', Technical Report 9, United Kingdom Airprox Board, UK Airprox Board, Hillingdon House, Uxbridge, Middlesex, UB10 0RU, (2003). Available from `http://www.caa.co.uk/ukab`.

[6] C.A.C Coello, 'A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques', *Knowledge and Information Systems. An International Journal*, **1**(3), 269–308, (1999).

[7] I. Das and J.Dennis, 'A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems', *Structural Optimization*, **14**(1), 63–69, (1997).

[8] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, Wiley, Chichester, 2001.

[9] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson, 'Comparing the areas under two or more correlated receiver operating charteristic curves: a non parameteric approach', *Biometrics*, **44**(11), 837–845, (1988).

[10] R.O. Duda and P.E. Hart, *Pattern Classifi cation and Scene Analysis*, Wiley, New York, 1973.

[11] B. Efron and R.J. Tibshirani, *An introduction to the Bootstrap*, number 57 in Monographs on Statistics and Probability, Chapman & Hall, New York, 1993.

[12] T. Fawcett, 'ROC Graphs: Notes and Practical Considerations for Researchers', *Machine Learning*, (2004). (Submitted).

[13] J.E. Fieldsend, R.M. Everson, and S. Singh, 'Using Unconstrained Elite Archives for Multi-Objective Optimisation', *IEEE Transactions on Evolutionary Computation*, **7**(3), 305–323, (2003).

[14] C.M. Fonseca and P.J. Fleming, 'An Overview of Evolutionary Algorithms in Multiobjective Optimization', *Evolutionary Computation*, **3**(1), 1–16, (1995).

[15] D.J. Hand, *Construction and assessment of classifi cation rules*, Wiley, 1997.

[16] J.A. Hanley and B.J. McNeil, 'The mean and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, **82**(143), 29–36, (1982).

[17] J. Knowles and D. Corne, 'The Pareto Archived Evolution Strategy: A new baseline algorithm for Pareto multiobjective optimisation', in *Proceedings of the 1999 Congress on Evolutionary Computation*, pp. 98–105, Piscataway, NJ, (1999). IEEE Service Center.

[18] J.D. Knowles and D. Corne, 'Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy', *Evolutionary Computation*, **8**(2), 149–172, (2000).

[19] M.A. Kupinski and M.A. Anastasio, 'Multiobjective Genetic Optimization of Diagnostic Classifi ers with Implications for Generating Receiver Operating Characterisitic Curves', *IEEE Transactions on Medical Imaging*, **18**(8), 675–685, (1999).

[20] P. Palanque, F. Paternó, and R. Fields, 'Designing user interfaces for safety critical systems', in *Bulletin of ACM Speical Interest Group on Computer Human Interaction*, volume 30, Association of Computing Machinery, (1998).

[21] V. Pareto, *Manuel D'Économie Politique*, Marcel Giard, Paris, 2nd edn., 1927.

[22] M.J.J. Scott, M. Niranjan, and R.W. Prager, 'Parcel: feature subset selection in variable cost domains', Technical Report CUED/F-INFENG/TR.323, Cambridge University Engineering Department, (1998).

[23] D. Van Veldhuizen and G. Lamont, 'Multiobjective Evolutionary Algorithms: Analyzing the State-of-the-Art', *Evolutionary Computation*, **8**(2), 125–147, (2000).

[24] X. Yao, Y. Liu, and G. Lin, 'Evolutionary Programming Made Faster', *IEEE Transactions on Evolutionary Computation*, **3**(2), 82–102, (1999).