How precise has fault detection to be? Answers from an economical point of view

Thomas Boehm[1]

[1] German Aerospace Center (DLR), Institute of Transportation Systems, Germany, Thomas.boehm@dlr.de

Keywords: condition monitoring, accuracy requirements, classifier evaluation, railway switch

Abstract

Condition monitoring as part of an effective condition based maintenance has long been accepted in the academic and industrial world, 25 COMADEM conferences are proof of that. Various methods, algorithms, and tools have been developed to monitor the condition, detect faults, or predict the remaining useful life. Furthermore, there are methods available supporting the engineer to choose the appropriate approach to his problem at hand.

One question that often comes along with the implementation of fault detection and failure prediction is: How accurate has the condition monitoring to be and what is the right balance between false alerts and undetected faults which may lead to failures?

This paper presents a method to calculate the required accuracy in terms of an economical efficient condition monitoring. It gives the parameters necessary to calculate the right balance between undetected failures and false alerts. The diagnosis of a railway point machine is used to illustrate the application of the proposed method. Points are critical to the railway operation and their breakdown has a high impact on delays and hence costs. Moreover, the structure of the railway network makes them a distributed system, sometimes hard to reach for maintenance staff, which also makes the maintenance expensive. This and other facts make the decision complex, but give a good example on how the question for the required accuracy can be answered taking into account the goal of a maximum of cost efficiency.

1. Introduction

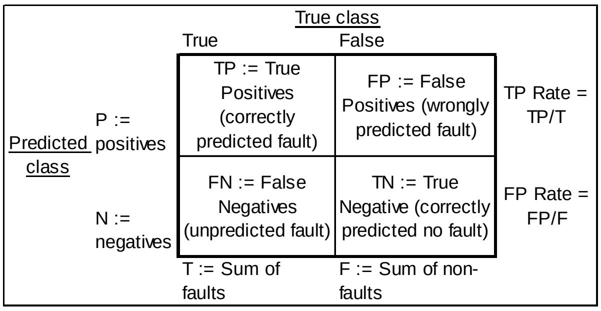
Condition monitoring as part of effective condition based maintenance has long been accepted in the academic and industrial world. One question that often comes along with the implementation of fault detection and failure prediction is: How precise has the condition monitoring to be and what is the right balance between false alerts and undetected faults which may lead to failures? Answers can be given from different points of view. From a technical perspective a perfect solution is wanted, as precise as possible. From the human perspective a reliable solution is wanted, so the maintenance worker trusts the condition monitoring system. While the first perspective might be out of scale, it is hard to tell the precise figure of required accuracy for the second perspective. Especially under the assumption that humans differ in their opinion and their attitude regarding automated diagnosis and predictions. But there is a third perspective, the economical point of view. Naturally, this issue will get much attention from managers having to decide about fault detection and failure prediction systems. Problems with over engineered or ineffective systems should not occur under the economic perspective.

2. Performance Evaluation of Classifiers

The general effectiveness of results of all methods and algorithms to predict (or detect) faults can be described using their resulting scores of:

- Correctly predicted faults or True Positives
- Wrongly predicted faults or False Positives
- Unpredicted faults or False Negatives
- Correctly predicted no faults or True Negatives

This is usually understood as confusion matrix (see **Virhe. Viitteen lähdettä ei löytynyt.**). From the confusion matrix several performance metrics for classifiers are calculated. The accuracy, which is expressing the percentage of correct predictions, is perhaps the most common.



		True class		
		True	False	
Predicted class	P := positives	TP := True Positives (correctly predicted fault)	FP := False Positives (wrongly predicted fault)	TP Rate = TP/T
	N := negatives	FN := False Negatives (unpredicted fault)	TN := True Negative (correctly predicted no fault)	FP Rate = FP/F
		T := Sum of faults	F := Sum of non-faults	

Figure 1: Confusion Matrix as Basis for ROC Analysis

Though the accuracy is a single-number metric, which makes it easy for humans to compare and rank different accuracy values, it has some shortcomings. As argued in Hand 1997 and Provost et al 1998 the accuracy metric biggest disadvantage is its sensitivity to class distribution changes. Therefore the Receiver-Operating-Characteristics (ROC) graph as introduced by Egan 1975 should be preferred to analyse classifiers performance (see also Swets 1988). ROC-graphs show the classifiers performance in two dimensional space as False Positive Rate and True Positive Rate (for example see the classifiers A and B in Figure 2). This is much more suitable to evaluate the performance of a prediction, but is still limited to a mere technical reasoning. It does not tell under which circumstances A is preferable over B and vice versa. In fact, both of the examples are equally good from a technical point of view. Also the ROC analysis takes only a few class distributions of True and False into account. These are basically the class distributions at training and testing the classifier. The class distribution at the operational use of the classifier is unknown in most cases and hence not at all part of performance evaluation, but it should be. Additionally, the pure ROC-analysis assumes cost of correct and incorrect classification results to be equal, which will not hold true in many real life classification problems. For example, in medical applications the classification of a patient as healthy while he has a mortal disease is much more costly than the other way around. Another example is the classification for banned substance abuse in sports, in which the False positive result is much more costly for the tested athlete than a False negative result, because the first might lead to disqualification and suspension.

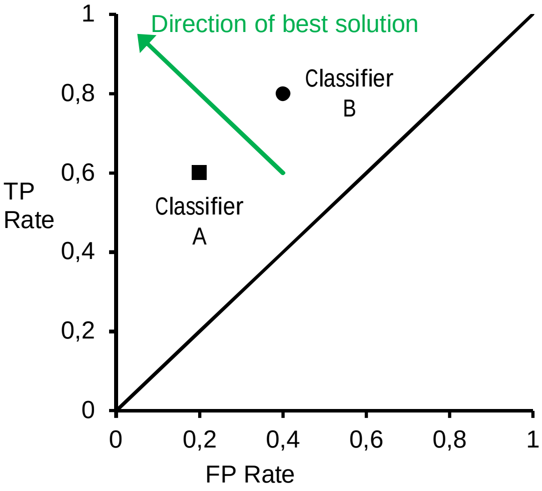


Figure 2: Example of two classifies in a ROC graph

The consideration of different misclassification costs has been addressed in the past. For instance, the performance graphs of Turney 1995, “Regret Graphs” of Hilden and Glasziou 1996, “Loss Difference Plots” of Adams and Hand 1999. The Meta Cost algorithm described by Domingos 1999 integrates costs in the training phase of the classifier. Regret Graphs require exact knowledge of misclassification, which is a problem for many real life classification tasks. Adams and Hand 1999 made progress in this area. They assumed a certain range and distribution of the costs are known to domain experts and can therefore be considered. However, all of the above do not handle the varying class distribution between True and False as well as varying misclassification costs.

This was argued by Drummond and Holte 2006 and lead to the development of cost curves. Cost curves visualise the performance of a classifier (expressed as expected costs) as a function of the misclassification costs and the class distribution in a single number. “Cost curves share many of ROC curves, desirable properties, but also visually support several crucial types of performance assessment that cannot be done easily with ROC curves…” (Drummond and Holte 2006, p. 126). They also show the mathematical association between ROC curves and cost curves, a bidirectional point/line duality. A point in a ROC graph is a line in a cost curve and vice versa.

Allthough cost curves are a very good visualisation and evaluation method for classifier performance there are some shortcomings, too. First, they take the costs of False Positives and False Negatives. In many technical applications of classifiers, like predicting faults, is done to generate a benefit. This means the True Positives generate benefit in terms of preventing or reducing lost money during breakdowns. A manager will invest in fault detection and prediction to save money. False Negative costs occur anyway if there is no classifier. From a mathematical point of view the benefit of True Positives is simply the negative cost of False Negative. But the use of True Positive benefit und False Positive cost (negative benefit) leads to values of positive and negative algebraic sign. The best score no longer is the one with zero costs, but the one with the maximum benefit. This alters the way the costs are normalised and hence the argumentation and interpretation of cost curves.

Second, it is still necessary to view ROC curves and cost curves to evaluate the classifier performance. Thus it is highly desired to analyse and display the performance in a unified framework.

3. Cost Curve using Benefit of True Negatives and False Positives

Equation (1) defines the expected benefit as the sum of the True Positive Rate (*TP$_r$*) multiplied by the number of true faults (*T*) multiplied by the benefit of a True Positive classification (*B$_{TP}$*) and the False Positive Rate (*FP$_r$*) multiplied by the number of False (*F*) multiplied by the benefit of a False Positive classification (*B$_{FP}$*), which are the negative costs of a False Positive classification (*C$_{FP}$*).

$$E[B] = TP_r * T * B_{TP} + FP_r * F * B_{FP} \tag{1}$$

where $B_{FP} \overset{!}{<} 0$ and $B_{FP} = -C_{FP}$

In a first step the distribution of True and False class members is normalised to get equation (2) with the expected benefit rate (*E[B$_r$]*) and the share of False (*F$_r$*) and True (*T$_r$*).

$$\begin{aligned} E[B_r] &= \frac{E[B]}{(T+F)} \\ &= TP_r * T_r * B_{TP} + FP_r * F_r * B_{FP} \end{aligned} \tag{2}$$

where $T_r = \frac{T}{(T+F)}$ and $F_r = \frac{F}{(T+F)}$

It is now possible to express the relative share of True and False class members in the benefit evaluation. To normalise (in an interval from 0 to 1) the expected benefit rate regarding the possible costs it is necessary to define the minimum of the expected benefit rate (*minE[B$_r$]*) and the maximum (*maxE[B$_r$]*). The minimum is reached in case the classifier detects or predicts all faults while producing zero false alerts. Note, that this is also the optimal solution from a technical or a user perspective, though it may never be found in reality.

$$\min E[B_r] = 0 * T_r * B_{TP} + 1 * F_r * B_{FP} \tag{3}$$

$$\max E[B_r] = 1 * T_r * B_{TP} + 0 * F_r * B_{FP} \tag{4}$$

That leads to the normalised expected benefit rate (*Ê[B$_r$]*) using equation (3) and (4) to get equation (5).

$$\begin{aligned} \hat{E}[B_r] &= E[B_r[\min..\max]] \xrightarrow{lin} E[B_r[0..1]] \\ &= \frac{E[B_r] - \min E[B_r]}{\max E[B_r] - \min E[B_r]} \\ &= \frac{TP_r * T_r * B_{TP} + FP_r * F_r * B_{FP} - F_r * B_{FP}}{T_r * B_{TP} - F_r * B_{FP}} \end{aligned} \tag{5}$$

Two additional variables are introduced to simplify the resulting term of equation (5), the normalised benefit of the True rate (*B$_r$(T$_r$)*) and the normalised benefit rate of the False rate (*B$_r$(F$_r$)*).

$$B_r(T_r) = \frac{T_r * B_{TP} - F_r * B_{FP}}{T_r * B_{TP} - F_r * B_{FP}} \tag{6}$$

$$B_r(F_r) = \frac{F_r * B_{FP} - F_r * B_{FP}}{T_r * B_{TP} - F_r * B_{FP}} \tag{7}$$

where $B_r(T_r) + B_r(F_r) \overset{!}{=} 1$

In any situation the addition of (6) and (7) must give the normalised expected benefit rate of 1. This leads to equation (8), which is the common algebraic term for a first degree polynomial.

$$\begin{aligned} \hat{E}[B_r(T_r)] &= TP_r * B_r(T_r) + (1 - FP_r) * (1 - B_r(T_r)) \\ &= TP_r * B_r(T_r) + 1 - B_r(T_r) - FP_r + FP_r * B_r(T_r) \\ &= (TP_r + FP_r - 1) * B_r(T_r) - FP_r + 1 \end{aligned} \tag{8}$$

Similar to cost curves, this enables the visualisation of the normalised expected benefit across all possible distributions of True and False class members. If there are no faults the normalised expected benefit is 1-*FP$_r$*. If there are only faults the normalised expected benefit is the *TP$_r$*, resulting from the number of prevented faults. Of course, a machine which has only faults will not be operated at all.

$$\hat{E}[B_r(T_r)] = \begin{cases} 1-FP_r, when\ B_r(T_r)=0 \\ TP_r, when\ B_r(T_r)=1 \end{cases} \qquad (9)$$

From equation (9) it is easy to draw a line in a two dimensional graph by connecting the two values. Table 1 shows some examples. Therein the classifier A is shown in three variants A1 to A3. They differ in their True and False class distribution and the benefit. Also the values of classifier B are listed. B has a better True Positive Rate and a better False Positive Rate, but for demonstration neither the True Positives nor the False Positives produce a benefit. The values are negative. From the visualisation of the classifiers performance as cost curves (see Figure 3) it seems that all variants of A perform equally. Moreover, B appears to be superior to A over all possible class distributions. A more detailed discussion about cost curve interpretation can be found in Drummond and Holte 2006. However, comparing the cost curves with the values of the expected benefit (*E[B$_r$]*) in Table 1, it becomes clear that cost curves do not reveal the Break Even Situation.

Classifier	A1	A2	A3	B
Value				
FP Rate	0.13	0.13	0.13	0.1
TP Rate	0.38	0.38	0.38	0.8
F	1234	123	1234	1234
T	567	4567	567	567
B FP	-10	-10	2	-10
B TP	20	20	10	-20
maxE[Br]	6.30	19.48	3.15	-6.30
minE[Br]	-6.85	-0.26	1.37	-6.85
Br(0)	0.87	0.87	0.87	0.9
Br(real)	0.64	0.39	0.00	2.03
Br(1)	0.38	0.38	0.38	0.8
E[Br]	2.705	34.549	2.475	-10.306

Table 1: Example of classifiers and their corresponding values for visualisation as cost curves

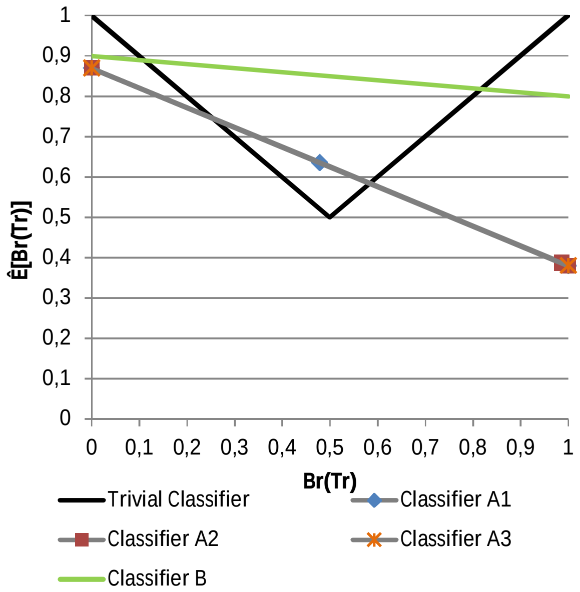


Figure 3: Visualisation of classifier performance as cost curves

4. Break Even Situation

The idea behind a performance analysis of fault detection and prediction from an economical point of view is to define and visualise the Break Even Situation. The economic science defines the Break Even point as balance between profit and loss. In this paper, the Break Even Situation is understood as the situation in which the benefit from correctly prevented faults is equal to the loss from reacting to false alerts. In the special case of the Break Even Situation the expected benefit (*E[B]*) of using a fault detection or prediction technique is zero. Since a positive economic benefit is wanted (*E[B]*) should be greater.

Therefore, the condition $E[B_r] \overset{!}{>} 0$ is applied to equation (1) and equation (2), respectively. The latter is changed as follows:

$$\begin{aligned} E[B_r] &= TP_r * T_r * B_{TP} + FP_r * (1-T_r) * B_{FP} \\ &= (TP_r * B_{TP} - FP_r * B_{FP}) * T_r + FP_r * B_{FP} \end{aligned} \qquad (10)$$

Additionally, a point of *FP$_r$* and *TP$_r$* is defined in which the Break Even situation is reached (*P$_0$*) while *FP$_r$* is greater than zero (see expression (11)).

$$P_0(FP_r, TP_r) = \left\{ \begin{aligned} &(FP_r, TP_r) \in [0..1] \ \mid FP_r > 0, \\ &TP_r * T_r * B_{TP} + FP_r * F_r * B_{FP} > 0 \end{aligned} \right\} \qquad (11)$$

By combining equations (10) and (11) and transpose them, it is now possible to determine the required *TP*TP_r in any given combination of FP_r and T_r with the use of equation (12).

$$f(FP_r,T_r)=\frac{FP_r*T_r*B_{FP}+FP_r*B_{FP}}{T_r*B_{TP}} \qquad (12)$$

$$=TP_r$$

It also defines a plane in the three-dimensional space of FP_r, TP_r and T_r. Obviously, the plane separates the space in two parts (see Figure 4). One contains the classifier generating a positive overall benefit and the other contains the classifier costing money if used in this situation.

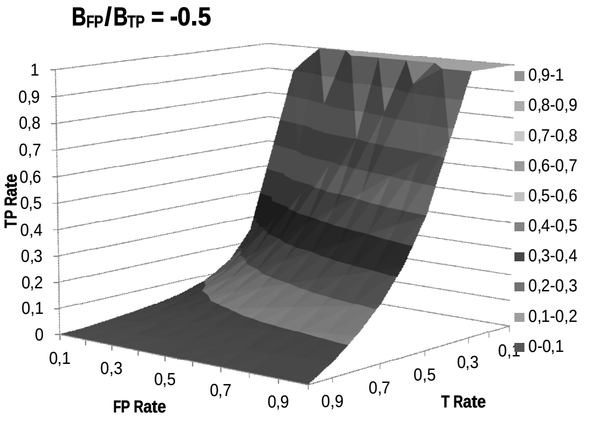


Figure 4: Relation of FP_r, TP_r and T_r for a benefit > 0 as a result of the benefit values from classifier A1 in Table 1

As shown in Figure 4 it is now easy to see if a classifier performs in a state of positive benefit or not. Using the benefit of True Positives and False Positives from classifier A1 in Table 1, all situations in which results are allocated on the left upper side will generate a positive benefit. In this case Classifier A1 is sufficient enough as long 263.1 or more of the 1801 class members are faults, or the T_r is greater 0.1461.

As stated previously the benefit values are unknown in many situations. Thus, in order to integrate this uncertainty in the analysis the cost benefit ratio (*CBR*) is introduced and defined in equation (13).

$$CBR=\frac{B_{FP}}{B_{TP}} \qquad (13)$$

With the help of the cost benefit ratio it is possible to combine and transpose the equations (10), (11) and (13) to equation (14). It determines the required TP_r in any given combination of FP_r , T_r and *CBR*.

$$f(FP_r,T_rCBR)=\frac{(T_r-1)*FP_r*CBR}{T_r} \qquad (14)$$

$$=TP_r$$

By analogy to equation (12), it defines a hyperplane in four dimensional space separating this space in two parts. One contains the classifier generating a positive overall benefit and the other contains the classifier costing money if used. The visualisation of the hyperplane is relinquished here, because of the difficulties to illustrate graphs of four dimensions.

However, the analysis of one or more classifiers detecting or predicting faults can now be evaluated in a way similar to the analysis done with the original ROC graphs. With its help the question for the required precision of classifiers can be answered properly from an economical point of view. A classifier is precise enough, if it is on the top side of the hyperplane, meaning it exceeds the Break Even Situation.

5. *Example: Evaluating the Failure Detection and Prediction of Railway Point Machines*

Railway Points are critical to the railway operation and their breakdowns have a high impact on delays of trains and costs, respectively. Moreover, the structure of the railway network makes them a distributed system, sometimes hard to reach for maintenance staff, which also makes the maintenance expensive. This and other facts make the decision complex, but give a good example on how the question for the required accuracy can be answered taking into account economic factors.

Point engine diagnosis systems are used to detect and predict failures of railway points. As shown in Boehm 2012a the reliability of such systems is questionable. Thought, improvements of a system is shown exemplarily in Boehm 2012b, the argumentation how precise the fault detection or prediction should be is not presented.

The search for the minimum precision of the diagnosis systems starts with the required economical parameters. Figure 1 shows a confusion matrix. The case of the True False has no costs or benefit, because it is correct to do no maintenance. The cost for False Negatives are also set to zero, because failures would occur without a diagnosis system, hence no additional cost can be claimed. As mentioned above, True

Positives can save money, meaning generate benefit. False Positives lead to unnecessary maintenance activities resulting in costs, meaning negative benefit.

The benefit of False Positives (B_{FP}) contains the following parameters:

- Costs for driving to the point location, including the travel time, fuel consumption, etc.
- Labour costs for the maintainers
- Material costs is set to zero, because no components are replaced
- Lost income from unawarded train paths during maintenance

The parameters are summed and multiplied by -1 to calculate the B_{FP}.

The benefit of True Positives (B_{TP}) contains the following parameters:

- Average minutes of delays caused by a point
- Costs per minute of delay

The two parameters are multiplied to calculate B_{TP}.

A particular point from the railway network of Deutsche Bahn is selected as an example. The B_{FP} can be determined exactly at 50.17 Euro. The value of B_{TP} is much harder to determine, because there is no exact figure for the costs per minute of delay available. Although, different figures from different railway operators can be found in international literature, only the figures published in Schilling and Lücking 2003 are used here, because they had been derived from analysis in Germany. They mention costs of 60 to 130 Euro per delay minute in the German network. The railway point has an average of 334.02 delay minutes per failure. The figures give a range for the *CBR* from -0.002504 to -0.001155. The range for the T_r can be derived from the results of the analysis in Boehm 2012b and from discussions with maintainers. Approximately, a point is failing ones every 2000 to 20000 turns. Hence the range for the T_r is 0.0005 to 0.00005.

Assuming these figures are true, the resulting space of positive benefit is quiet small (see Figure 5 and Figure 6). The plane representing the performance of the diagnosis system is never intersecting the plane of the Break Even Situation. Hence, the current system is not precise enough in order to generate a positive economic benefit.

Additionally, Table 2 shows some results of the required TP_r for a discrete set of values for FP_r. Most cells contain values above 1. They exceed the boundaries of all possible TP_r values. Of course, no system will ever detect or predict more faults correctly, than faults exist.

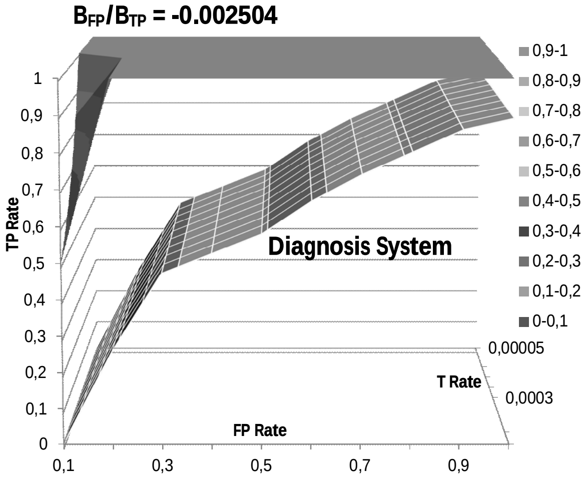


Figure 5: Plane of the Break Even Situation in case of 60 Euro costs per minute of delay and the performance of the diagnosis system

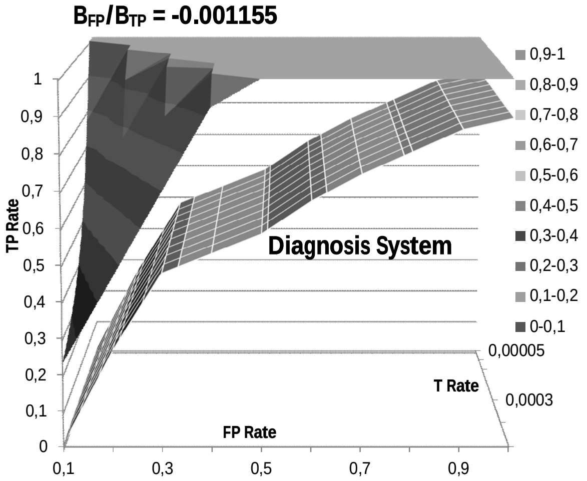


Figure 6: Plane of the Break Even Situation in case of 130 Euro costs per minute of delay and the performance of the diagnosis system

-	CBR = -0.002503		CBR = -0.001155	
-	T Rate	T Rate	T Rate	T Rate
-	0.00005	0.0005	0.00005	0.0005
FP Rate	-	-	-	-
0.1	5.006	0.500	2.311	0.231
0.2	10.013	1.001	4.621	0.462
0.3	15.019	1.501	6.932	0.693
0.4	20.025	2.002	9.243	0.924
0.5	25.032	2.502	11.553	1.155
0.6	30.038	3.002	13.864	1.386
0.7	35.044	3.503	16.175	1.617

0.8	40.051	4.003	18.485	1.848
0.9	45.057	4.504	20.796	2.079
1.0	50.063	5.004	23.107	2.310

Table 2: Some values of the TP_r under given FP_r, T_r and CBR for a railway point if benefit >0

However, the *CBR* or the T_r requires a rather low rate of False Positives in order to generate a positive economic benefit. Even if only the 160 Euro per minute of delay is considered the original diagnosis system does not operate economical efficient. It raises too many false alerts. The proposed improvement of the accuracy in Boehm 2012b does not generate a positive benefit either (see Figure 6). Hence, more sophisticated methods to detect and to predict failures of railway points need to be developed as long as the *CBR* or the *Tr* are not changing in favour of an easier to reach positive benefit.

6. Conclusions

A method to determine and analyse the precision of fault detection or prediction techniques was presented. This method allows analysing classifiers from an economical point of view. Therefore, ROC-graphs are extended into four dimensions in order to enable a user to analyse classifiers performance independent from the class distribution and the costs of misclassifications. The visualisation of the Break Even Situation as a plane in three dimensional space allows a simple view on weather a positive benefit is generated or not. The failure detection of railway points was used as an example how to analyse the performance of the corresponding diagnosis system. The results show that more sophisticated methods are necessary. The question if the point diagnosis system is precise enough from an economical point of view must be negated in this case.

7. References

Adams, Niall M.; Hand, David J. (1999): Comparing classifiers when the misallocation costs are uncertain. In: Pattern Recognition, Vol. 32, no. 7, pp. 1139-1147.

Boehm, Thomas (2012a): Zustandsorientierte Instandhaltung bei Eisenbahnweichen. In: Next Generation Railway System. Ausgewählte Projektergebnisse. Braunschweig: DLR-Institut für Verkehrssystemtechnik (Berichte aus dem DLR-Institut für Verkehrssystemtechnik, no. 18, pp. 51-62.

Boehm, Thomas (2012b): Accuracy Improvement of Condition Diagnosis of Railway Switches via External Data Integration. In: Structural Health Monitoring 2012. Proceedings of the Sixth European Workshop on Structural Health Monitoring. Germany, pp. 1550-1558.

Domingos, Pedro (1999): MetaCost: A General Method for Making Classifiers Cost-Sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM (KDD '99), pp. 155-164.

Drummond, Chris; Holte, Robert C. (2006): Cost curves: An improved method for visualizing classifier performance. In: Machine Learning, Vol. 65, pp. 95-130.

Egan, James P. (1975): Signal Detection Theory and ROC Analysis. New York: Academic Press (Series in Cognitition and Perception).

Hand, David J. (1997): Construction and assessment of classification rules. Chichester: Wiley (Wiley series in probability and mathematical statistics).

Hilden, Jørgen; Glasziou, Paul (1996): Regret Graphs, Diagnostic Uncertainty and Youden's Index. In: Statistics in Medicine, Vol. 15, no. 10, pp. 969-986.

Provost, F.; Fawcett, T.; Kohavi, R. (1998): The Case Against Accuracy Estimation for Comparing Inductive Algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 445-453.

Schilling, Rosemarie; Lücking, Lars (2003): Senkung der Lebenszykluskosten. In: EI - Eisenbahningenieur, Vol. 54, no. 5, pp. 58-72.

Swets, John A. (1988): Measuring the accuracy of diagnostic systems. In: Science, Vol. 240, pp. 1285-1293.

Turney, Peter D. (1995): Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. In: Journal of Artificial Intelligence Research (JAIR), Vol. 2, pp. 369-409.