

Reliable Instance Classifications in Law Enforcement

Stijn Vanderlooy Eric Postma Karl Tuyls
Ida Sprinkhuizen-Kuyper

MICC-IKAT, Universiteit Maastricht, PO Box 616,
6200 MD Maastricht, The Netherlands

Abstract

Machine-learning classifiers are gaining interest in the domain of law enforcement. However, when classifiers are applied in this domain two fundamental difficulties are encountered. First, the cost of an incorrect classification is extremely high. Second, the error-cost distribution is not static. In this paper we relate these difficulties to the reliability of instance classifications. Accordingly, we define two requirements to be met by classifiers. It is shown that an approach based on isometrics in Receiver Operator Characteristic (ROC) space is able to construct classifiers that satisfy both requirements. Due to the efficiency and generality of the approach, classifiers can be safely used to solve law enforcement tasks.

1 Introduction

In the last years, terrorist acts and other criminal activities had a strong influence on the politics of nations worldwide. Preventing crime and providing security for civilians have become high-priority goals. To achieve these goals, governments allow more data storage and international data transfers for law enforcement. This has resulted in an overwhelming amount of information that needs to be processed and interpreted in an efficient way. The new form of law enforcement guided by data analysis is known as intelligence led policing [5].

In this paper we focus on law enforcement tasks that can be suitably solved by a machine-learning classifier. Two domain-specific problems arise when a classifier is applied. First, incorrect classifications have serious consequences, e.g., waste of limited resources and privacy violations when personal data are involved. Clearly, the cost of an incorrect classification (error cost) is high. A second problem is the dynamics of the error-cost distribution. This distribution describes the balance between the costs of false positives and false negatives. It varies with the societal and legal context. For example, when terrorist threats are high, ensuring public safety will become more important than securing civilians' privacy. If crime rates are low for a sustained time period, then securing privacy will be more important. The former implies that the error cost of failing to identify an upcoming terrorist attack is substantially higher than the error cost of violating the privacy of innocent civilians. The latter implies a smaller difference between the error costs.

To apply a classifier in law enforcement it has to deal with these two problems.¹ Accordingly, we define two requirements that a classifier has to meet. First, the classifier has to guarantee a minimum level of performance on each class. Second, in order to cope with a dynamic error-cost distribution, it needs to have the ability to adjust its performance in a timely fashion. We will show that so-called reliable classifiers [8, 9] meet both requirements.

The rest of this paper is organized as follows. Section 2 briefly outlines reliable classifiers. Section 3 shows how to construct these classifiers and explains why they satisfy the requirements. In Section 4 we apply the approach on a real-life legal classification task. Section 5 discusses benefits of applying reliable classifiers in law enforcement. Section 6 concludes the paper.

2 Reliable classifiers

We consider classification tasks with two classes: positive (p) and negative (n). To solve a task we construct a scoring classifier. This is a classifier that outputs two positive values $l(x|p)$ and $l(x|n)$ indicating the likelihood that an instance x is positive and negative, respectively. The score of an instance combines these values as follows:

$$l(x) = \frac{l(x|p)}{l(x|n)} \quad (1)$$

and is used to rank instances from most likely positive to most likely negative.

The classification of a new instance is obtained by applying a numerical threshold on the score. Instances with a score higher than or equal to this threshold are classified as positive. The remaining instances are classified as negative. Unfortunately, a scoring classifier is not optimal, i.e., there exists negative instances with a higher score than some positive instances. Therefore, the use of a threshold always results in (too) many incorrect classifications. This implies that we do not know if the classification assigned to a particular instance is correct.

To overcome this problem, an approach to reliable classifiers is defined [8, 9]. A reliable classifier is able to guarantee a specific performance on each class.² To ensure this guarantee, it may be needed to abstain from classifying ambiguous instances. This abstention is implemented as a filtering mechanism with two numerical thresholds $a > b$. An instance x is classified as positive if $l(x) \geq a$. If $l(x) \leq b$, then x is classified as negative. Otherwise, the instance is left unclassified since its score does not indicate the correct classification with sufficient certainty.

3 Reliable Classifiers via ROC Isometrics

Recently, an approach was introduced to construct a reliable classifier with a desired precision on both classes [9]. In subsequent work [8] analysis revealed that the approach can be used for other performance metrics as well, e.g., the

¹Note that the two identified problems can also occur in other domains, e.g., medical diagnosis.

²Thus, by definition, a reliable classifier satisfies the first requirement that we have defined.

F -measure. Thus, performance on each class can be measured with a variety of performance metrics. We will restrict to a classification task for which the minimum achievable performance is denoted in terms of precision on each class. We identify precision on each class with reliability. The reliability is said to be high when the precisions for both classes are high.

The key idea of the approach is to identify instances with uncertain classification. This is done on the basis of a scoring classifier and its representation in terms of a Receiver Operator Characteristic (ROC) curve. If the uncertain instances are left unclassified, then the corresponding classifier obtains the desired reliability. The next three Subsections 3.1, 3.2 and 3.3 discuss in detail this approach and how it allows us to handle error-cost distributions.

3.1 ROC Isometrics

Given a scoring classifier, its ROC curve shows the trade-off between false positive rate fpr and true positive rate tpr for each possible threshold on the score $l(x)$. Figure 1(a) shows an ROC curve.³ A concavity in this curve indicates local sub-optimality of the scoring classifier [3]. To remove concavities the convex hull of the ROC curve (ROCCH) is constructed.

It can be shown that for any point (fpr, tpr) on an ROCCH, a classifier can be constructed that has the performance represented by that point [7]. For simplicity of presentation, we assume that ROC curves are convex and all points can be obtained by a threshold.

Reliability is measured by the precision on the positive classifications ($prec_p^c$) and precision on the negative classifications ($prec_n^c$):

$$prec_p^c = \frac{tpr}{tpr + c fpr} \quad (2)$$

$$prec_n^c = \frac{tnr}{tnr + \frac{1}{c} fnr} \quad (3)$$

Here, $c = \frac{N}{P}$ with N and P the number of negative instances and positive instances, respectively. ROC isometrics are collections of points in the (fpr, tpr) plane with the same value for a performance metric [2]. Precision isometrics are defined by rewriting each precision equation to that of a line in ROC space:

$$tpr = \frac{prec_p^c}{1 - prec_p^c} c fpr \quad (4)$$

$$tpr = \frac{1 - prec_n^c}{prec_n^c} c fpr + 1 - \frac{1 - prec_n^c}{prec_n^c} c \quad (5)$$

The isometrics for positive precision are lines that go through $(0, 0)$. Isometrics for negative precision go through point $(1, 1)$. Varying the precision value results in a line with different slope. The higher the precision value, the closer the isometric

³Implicitly, an ROC curve also shows the trade-off between true negative rate tnr and false negative rate fnr since $tnr = 1 - fpr$ and $fnr = 1 - tpr$.

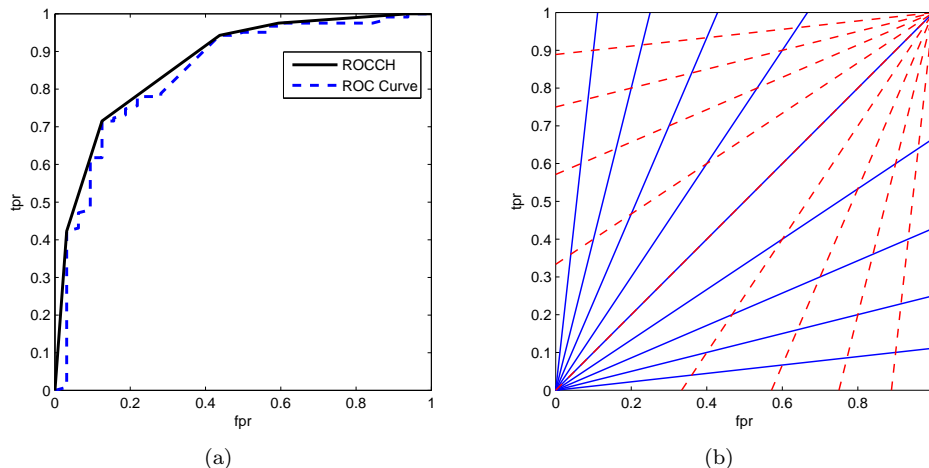


Figure 1: Constructions in ROC space: (a) an ROC curve with corresponding ROCCH, and (b) $prec_p^c$ -isometrics (solid lines) and $prec_n^c$ -isometrics (dashed lines).

approaches point $(0, 1)$. Figure 1(b) shows $prec_p^c$ -isometrics and $prec_n^c$ -isometrics. The precision value is varied from 0.1 to 0.9 in steps of 0.1. In the next Subsection 3.2 we show how these lines can be used to construct a reliable classifier, i.e., a classifier with guaranteed reliability.

3.2 Design Framework

Given a ROCCH and a desired positive precision and negative precision, a positive precision isometric and a negative precision isometric are constructed. The intersection point of the $prec_p^c$ -isometric and the ROCCH is denoted as (fpr_a, tpr_a) . By definition, this intersection point represents a classifier with the $prec_p^c$ used to construct the isometric. In case of multiple intersection points, we choose the intersection point with the highest tpr value. Similarly, the intersection point (fpr_b, tpr_b) of the $prec_n^c$ -isometric and the ROCCH represents a classifier with negative precision defined by that isometric. In case of multiple intersection points, we choose the intersection point with the lowest fpr value. If either $(fpr_a, tpr_a) = (0, 0)$ or $(fpr_b, tpr_b) = (1, 1)$, then the desired classifier cannot be constructed. Otherwise, we can distinguish the following three cases as shown in Figure 2:

- **Case 1:** the isometrics intersect on the ROCCH
The classifier corresponding to this point has by definition the precisions defined by both isometrics. Therefore, this classifier has the desired reliability.
- **Case 2:** the isometrics intersect below the ROCCH
This classifier also has the desired reliability. However, the classifiers corresponding to any point on the ROCCH between (fpr_b, tpr_b) and (fpr_a, tpr_a) have higher reliability.

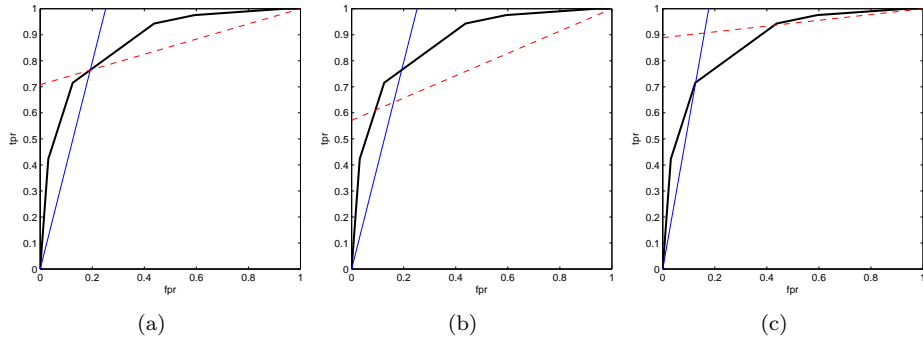


Figure 2: Location of the intersection point between a positive precision isometric and a negative precision isometric: (a) Case 1, (b) Case 2, and (c) Case 3.

- **Case 3:** the isometrics intersect above the ROCCH

There is no classifier with the desired reliability. Therefore, a number of ambiguous instances needs to be left unclassified. We define thresholds a and b of the abstention mechanism (see Section 2) to correspond with points (fpr_a, tpr_a) and (fpr_b, tpr_b) , respectively. It was shown that the reliable classifier with these thresholds has the desired reliability [8]. The construction of this classifier can be visualized by removing the instances in the ROCCH between points (fpr_a, tpr_a) and (fpr_b, tpr_b) . The removed instances are those left unclassified.

From the three cases we can conclude that a reliable classifier satisfies the first requirement, i.e., it is able to guarantee a desired level of reliability. In the next Subsection 3.3 we show that the second requirement is also satisfied.

3.3 Changing error costs

Having established a framework for designing classifiers with a desired reliability, we now extend it to deal with changing error costs (the second requirement). We define $c(p, n)$ and $c(n, p)$ as the costs of a false positive and false negative, respectively. For example, if the task is to classify a person as guilty (class p) or non-guilty (class n), then $c(p, n)$ represents the cost of classifying a non-guilty person as guilty. In a similar way the error cost $c(n, p)$ can be defined. Clearly, both costs are extremely high from both societal and legal perspective.

Error costs are incorporated in the precision metrics and corresponding isometrics by setting $c = \frac{c(p, n)}{c(n, p)} \frac{N}{P}$. Thus, changes in the error-cost distribution are incorporated by adjusting the class distribution. The effect is a change in the slope of the isometrics.⁴ Therefore, the design framework of the previous Subsection 3.2 remains valid.

⁴For a detailed discussion see [2, 8].

4 Application to Law Enforcement

We apply a reliable classifier to solve the task of detecting credit card abuse in a timely fashion. Credit card transaction can be classified as fraudulent (class p) or non-fraudulent (class n). The score $l(x)$ of transaction x can be seen as a suspicion score: the higher the score, the more unusual the transaction. The error cost of a false positive includes a waste of time and money. Moreover, since the card holder of a possibly fraudulent transaction is contacted by the agency, the false positive cost also includes a privacy violation. The error cost of a false negative includes non-detection of criminal activities.

Credit card fraud detection is gaining popularity in law enforcement because the number of fraudulent transactions is increasing significantly and a high level of (organized) crime activity is involved [1]. For these reasons, law enforcement agencies are prepared (and forced) to cope with the problem. However, the limited number of human resources implies that only a very small subset of the suspicious transactions can be investigated. A reliable classifier can be used to find a small number of instances with very high reliability. Thus, the used resources and obtained benefits are optimized.

To solve the outlined task, we reproduced the experiment in [6] that resulted in the best overall classifier (Bayesian belief network).⁵ The minimum reliability for the classifier to be applied is defined by $prec_p^c = 0.9$ and $prec_n^c = 0.95$. The ROCCH of the network is shown in Figure 3(a). From the location of the intersection point of the isometrics can be seen that abstention is needed to obtain the desired reliability. The point (fpr'_a, tpr'_a) that represents the resulting reliable classifier is marked by a solid disc. The classifier abstains from classifying 2531 instances. The output of the classifier for the remaining 2469 instances are considered as reliable.

If we remove the unclassified instances in the ROCCH, then we obtain a new ROCCH, i.e., ROCCH2 shown in Figure 3(b). This ROCCH has a larger area under the ROC curve (AUC), more specifically 0.96 in contrast to 0.88 for the original ROCCH. Although ROCCH2 is constructed with only a subset of the data of ROCCH1, the increase in AUC shows that the approach successfully identified instances that would be classified incorrectly.

5 Discussion

The law enforcement domain is changing constantly and it has a large impact on the politics of nations worldwide. Moreover, decisions and corresponding actions need to be correct since it concerns security and safety for civilians. For these reasons, it is difficult to apply machine-learning classifiers. Moreover, in the broader domain of legal practice, most of the intelligent applications concern legal text retrieval systems or expert systems for administrative tasks. It has been shown that even these systems make a substantial number of errors in their decisions [4].

⁵An instance (transaction) is described by ten features and its class. The high error costs are assumed to be equal and the skew ratio is $c = \frac{3405}{1595}$ (the class distribution of the training set).

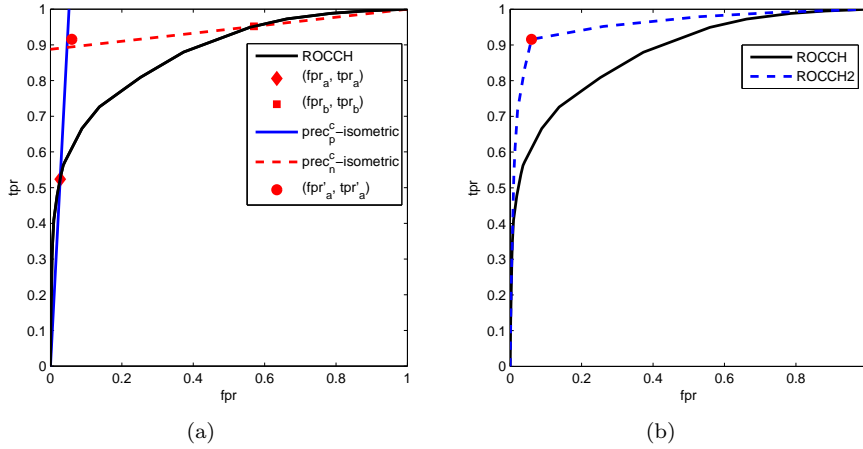


Figure 3: Fraud detection with a Bayesian belief network: (a) ROCCH with two isometrics defined for $c = \frac{3405}{1595}$, and $prec_p^c = 0.9$ and $prec_n^c = 0.95$, respectively. The disc symbol represents the reliable classifier, and (b) ROCCH2 is obtained by removing instances with uncertain classification.

Therefore, one cannot guarantee legally correct decisions and a correct treatment of civilians.

We argue that reliable classifiers are an appropriate solution to solve law enforcement tasks. Due to the massive amount of available data, reliable classifiers can improve the quality and reliability of information and support a more effective and efficient execution of law enforcement than was previously possible.

Our real-life application illustrated that reliable classifiers can be successfully used to cope with credit card fraud. In addition, the approach can be readily applied to other types of fraud detection with the same efficiency. We mention two examples: the detection of money laundering in an attempt to stop funds received by terrorist networks, and detecting computer intrusions to secure confidential data. In general, any application domain with high error costs benefits from reliable classifiers.

6 Conclusion

In this paper we focused on classification problems in law enforcement. We used an approach based on ROC isometrics to construct reliable classifiers. We have seen that this approach is generally applicable to construct a classifier with a predefined reliability. In the present context, reliability was measured by precision on each class but other performance metrics could be used as well, e.g., the F -measure. Furthermore, the approach has the ability to adjust its reliability in response to a changing error-cost distribution. Finally, we argued why reliable classifiers are desired in law enforcement. In conclusion, we state that reliable

classification significantly improves the viability of applying machine learning to the law enforcement domain.

Acknowledgments

The first author is sponsored by the Dutch Organization for Scientific Research (NWO), grant nr: 634.000.435. The third author is sponsored by the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

References

- [1] Richard Bolton and David Hand. Statistical fraud detection: a review. *Statistical science*, 17(3):233–255, 2002.
- [2] Peter Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, pages 194–201, Washington, USA, August 21-24 2003.
- [3] Peter Flach and Shaomin Wu. Repairing concavities in ROC curves. In *Proceedings of the 2003 UK Workshop on Computational Intelligence (UKCI 2003)*, pages 38–44, Bristol, UK, September 1-3 2003.
- [4] Marga Groothuis and Jörgen Svensson. Expert system support and juridical quality. In *Proceedings of the 13th Foundation for Legal Knowledge Based Systems Conference (JURIX 2000)*, pages 1–11, Enschede, the Netherlands, December 14-15 2000.
- [5] Paul De Hert, Wim Huisman, and Thijs Vis. Intelligence led policing ontleed. *Tijdschrift voor Criminologie*, 47(4):365–376, 2005. (in Dutch).
- [6] Sam Maes, Karl Tuyls, Bram Vanschoenwinkel, and Bernard Manderick. Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies (NF 2002)*, Havana, Cuba, January 16-19 2002.
- [7] Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.
- [8] Stijn Vanderlooy, Ida Sprinkhuizen-Kuyper, and Evgueni Smirnov. An analysis of reliable classifiers through ROC isometrics. In *Proceedings of the ICML 2006 Workshop on ROC Analysis (ROCML 2006)*, pages 55–62, Pittsburgh, USA, June 29 2006.
- [9] Stijn Vanderlooy, Ida Sprinkhuizen-Kuyper, and Evgueni Smirnov. Reliable classifiers in ROC space. In *Proceedings of the 15th Annual Machine Learning Conference of Belgium and the Netherlands (BENELEARN 2006)*, pages 113–120, Ghent, Belgium, May 11-22 2006.