# COMBINING ONE-CLASS CLASSIFIERS FOR MOBILE-USER SUBSTITUTION DETECTION

Oleksiy Mazhelis
*Department of Computer Science and Information Systems, University of Jyväskylä*
*P.O. Box35, FIN-40351, Jyväskylä, Finland*

Seppo Puuronen
*Department of Computer Science and Information Systems, University of Jyväskylä*
*P.O. Box35, FIN-40351, Jyväskylä, Finland*

Keywords: Wireless and mobile security, user verification, combining classifiers.

Abstract: Modern personal mobile devices, as mobile phones, smartphones, and communicators can be easily lost or stolen. Due to the functional abilities of these devices, their use by an unintended person may result in a severe security incident concerning private or corporate data and services. The means of user substitution detection are needed to be able to detect situations when a device is used by a non-legitimate user. In this paper, the problem of user substitution detection is considered as a one-class classification problem where the current user behavior is classified as the one of the legitimate user or of another person. Different behavioral characteristics are to be analyzed independently by dedicated one-class classifiers. In order to combine the classifications produced by these classifiers, a new combining rule is proposed. This rule is applied in a way that makes the outputs of dedicated classifiers independent on the dimensionality of underlying behavioral characteristics. As a result, the overall classification accuracy may improve significantly as illustrated in the simulated experiments presented.

## 1 INTRODUCTION

Today, mobile devices have become a convenient and often essential component assisting us in our everyday life. Some of the new abilities of these mobile devices are essential from the security perspective. Among them are i) the ability to store (private) data, ii) the ability to perform mobile e-transactions, and iii) the ability to access a corporate intranet. These abilities pose security concerns, since only the legitimate user of the device should be permitted to access the private data and the corporate intranet, or to carry out mobile e-transactions allowed to the device.

In order to ensure the legitimacy of a user, an authentication procedure is performed, usually consisting in entering PIN/password by a user. The authentication process is usually launched when the device is being turned on, or after idle time. However, many users find such protection mechanism inconvenient and do not use it (Clarke et al., 2002). As a result, their mobile devices appear insecure in the case they are lost or stolen. In this paper we will call as the tools of *user substitution detection* such kind of tools that through the detection of a substitution offer a base to build further security means rendering a mobile device useless for a non-legitimate person.

In this paper, the anomaly intrusion detection approach (Kumar, 1995) is followed i.e. the problem of user substitution detection is seen as the problem of detecting abnormal changes in the user behavior. It is assumed that the behavior of a user and a non-legitimate person (hereafter called impostor) will differ in some details, and that such differences can be automatically detected.

Different characteristics of the user behavior can be employed for the profile construction, and various aspects of user behavior can be reflected by these characteristics. These include, for example, typing peculiarities of a user (Monrose and Rubin, 2000), patterns of user mobility (Samfat and Molva, 1997), and application usage of a user. Some of such characteristics reflect low-level aspects of the user behavior (e.g. voice patterns and typing rhythms), and others correspond to high-level, goal-oriented aspects of the behavior or user preferences (as mobility patterns or patterns of device facilities usage). Taken together, they are expected to provide a comprehensive description of normal user behavior.

In many anomaly intrusion detection techniques, the term “anomaly” is interpreted in a probabilistic



Mazhelis O. and Puuronen S. (2004).
COMBINING ONE-CLASS CLASSIFIERS FOR MOBILE-USER SUBSTITUTION DETECTION.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 130-137
DOI: 10.5220/0002639901300137

sense, i.e. it corresponds to the observation of behavior with a low probability to be invoked by the legitimate user according to his past behavior. Various methods based on statistical probability modeling (Anderson et al., 1995; Burge and Shawe-Taylor, 1997; Cahill et al., 2000; Yamanishi et al., 2000; Schonlau et al., 2001), outlier detection (Aggarwal and Yu, 2001), clustering (Sequeira and Zaki, 2002; Eskin et al., 2002), etc. have been proposed to estimate how probable the current behavior is for the legitimate user. In attempt to reveal anomalies, most of these techniques analyze the whole set of available behavioral characteristics simultaneously. However, substantial disadvantages are inherited into this approach including:

- difficulties with learning when the variables are lumped into a single high-dimensional vector (Aggarwal and Yu, 2001; Xu et al., 1992); and
- difficulties with the normalization of variables having different physical meaning (Xu et al., 1992).

Besides, not all the variables may be present at the time the detection is performed. All these arguments justify the use of an alternative approach based on decision fusion (Dasarathy, 1994). Following this approach, the variables (called hereafter features) can be divided into subgroups processed by designated *classifiers*. Each of them is aimed at classifying the current values of assigned features as belonging to one of two classes: i) the *user class* that describes the normal user behavior using statistical models of feature value distributions, and ii) the *impostor class* reflecting the accumulated behavior of all possible impostors. By employing a *combining rule*, the final classification is produced based on the classifications provided by those designated classifiers.

In order to combine classifiers (and to adjust the combining rule), the knowledge about different classes is usually employed. However, while the user class can be modeled using the observed behavior of a legitimate user, almost no information may be available with respect to the impostor behavior. Therefore, the problem to be solved is that of *one-class classification* (Tax, 2001) whereby the target objects (the behavior of the user) is to be distinguished from all the other possible objects (the impostor behaviors).

Combining the classifications produced by several classifiers has been extensively explored as a mean to compensate the weaknesses of individual classifiers (Xu et al., 1992; Kittler and Alkoot, 2000; Tax and Duin, 2000). It has been shown that combining may result in significant reduction of classification errors (Kittler et al., 1998; Kuncheva, 2002). Different combining rules have been investigated, varying from simple fixed rules (as sum rule, product rule, and majority voting rule (Xu et al., 1992; Kittler et al., 1998)) to more complex trained rules (e.g. (Kittler and Alkoot, 2000)) as adopting stacked generalization approach (Wolpert, 1992). Most of the investigated combining rules deal with multi-class classification problem, where the task is to classify an instance presented by a vector of feature values into one class of the fixed set of alternative classes. These combining rules employ class related knowledge (e.g. distributions of the feature values for each class) to infer the final classification or to adjust the rule.

In combining one-class classifiers, where only the knowledge regarding one class is available, relatively few rules can be used. Among them are different modifications of voting rules as investigated by Xu et al. (Xu et al., 1992). More recently, Tax (Tax, 2001) reported the applicability of mean vote, mean weighted vote, product of weighted votes, mean of the estimated probabilities, and product combination of probabilities as combining rules for one-class classifiers. One of these rules, namely, the mean of the estimated probabilities rule, is reviewed below. In next section, this rule will be justified to be among the most suitable ones in the context of user substitution detection.

In one-class classification, an object $Z$ (presented by a vector $\mathbf{x}_i$ of the values of features from feature space $\mathcal{X}_i$, where $i$ designates $i$-th classifier) is classified into one of two classes $\{C_U, C_I\}$ where $C_U$ denotes the target class (later called user class) and $C_I$ denotes the class of outliers (later called impostor class) collecting all other objects not belonging to the target class. When $R$ classifiers are combined, each of them is assumed to represent its classification for $Z$ by a probability density function (pdf) for the user class $p(\mathbf{x}_i|C_U)$ (in fact, it is an estimation of $p(\mathbf{x}_i|C_U)$ produced by classifier $i$). In one-class classification, the pdf for the impostor class $p(\mathbf{x}_i|C_I)$ is assumed unknown.

Several rules based on the posterior probabilities $P(C_U|\mathbf{x}_i)$ have been investigated by Tax (Tax, 2001) for combining one-class classifiers. Different assumptions were made in order to infer the rules. Below, the *mean of the estimated probabilities (MP)* rule is represented that was produced under the following assumptions (Tax, 2001, pp. 118, 123):

- A1: $p(\mathbf{x}_i|C_I)$ is assumed to be independent of $\mathbf{x}_i$, i.e. it is distributed uniformly in the feature space $\mathcal{X}_i$. Using this assumption, $P(C_U|\mathbf{x}_i)$ is substituted with $p(\mathbf{x}_i|C_U)$.
- A2: classifiers operate using the same feature space, i.e. $\mathcal{X}_1 = \cdots = \mathcal{X}_R$. Then all $R$ classifiers provide the estimation of the same random variable $P(C_U|\mathbf{x}_1, \ldots, \mathbf{x}_R) = P(C_U|\mathbf{x}_i), \quad i = 1, 2, \ldots, R$.
- A3: the values of $p(\mathbf{x}_i|C_U)$ are estimated by the classifiers with the same zero-mean noise.

Under these assumptions, the MP rule is:

$$u_{mp}(\mathbf{x}_1, \ldots, \mathbf{x}_R) = R^{-1} \sum_{i=1}^{R} p(\mathbf{x}_i|C_U). \quad (1)$$

The MP rule is proposed as a mean to reduce the variance of (or, equally, suppress the noise in) the estimate. The final classification result using the MP combining rule is made by comparing the obtained $u_{mp}$ value with a threshold $t_{mp}$:

$$\begin{aligned} &\text{Decide} \quad Z \in C_U \quad \text{if} \quad u_{mp} \geq t_{mp}, \\ &\text{otherwise decide} \quad Z \in C_I. \end{aligned} \quad (2)$$

In this paper, we present a new modification of the MP rule and justify it as potentially appropriate for combining classifiers in the context of user substitution detection as improving the final classification accuracy. In the modified version, the outputs of the designated classifiers are made independent on the dimensionality of the underlying features. As a result, the reduction of the final classification error can be achieved. In this paper, our primary interest is in classifiers' outputs to be combined; thus the detailed design of individual classifiers is not considered (for an extensive discussion of individual classifiers the reader is suggested to consult e.g. (Samfat and Molva, 1997; Monrose and Rubin, 2000; Seleznyov, 2002)).

Many works in the intrusion detection domain addressed the problem of combining classifications of individual classifiers, e.g. (Anderson et al., 1995; Valdes and Skinner, 2000; Manganaris et al., 2000). The approaches most similar to ours are those employed in statistical component of NIDES (Anderson et al., 1995) and in (Ye and Chen, 2001), where probabilistic outputs of one-class classifiers are combined. In (Anderson et al., 1995), classifiers' outputs are mapped on half-normal distribution, and the sum of squares of transformed values is treated as following chi-square distribution. Similarly, in (Ye and Chen, 2001) classifiers' outputs are assumed normally distributed and chi-square test statistic is employed to combine them. However, in these works the outputs of classifiers should follow a predefined distribution, while no such constraints are imposed by the combining rule proposed in this paper.

The paper is organized as follows. In next section, the suitability of the MP rule in the context of user substitution detection is justified. The modification of this rule is introduced in section 3. Then, in section 4, the results of simulated experiments are presented wherein the original and modified mean probability rules are compared, and the benefits of the modification proposed are illustrated. Finally, section 5 discusses pros and cons of the modified rule and outlines the directions for future work.

# 2 MP COMBINING RULE FROM THE PERSPECTIVE OF USER SUBSTITUTION DETECTION

The MP rule is similar to (and may be considered as a special case of) a robust Sum rule for combining multi-class classifiers analyzed by Kittler et al. (Kittler et al., 1998). Following the Bayesian approach, the authors derived several combining rules for multi-class classifiers. The Product rule and the Sum rule are among them. The Product rule assumes that the values of the features $\mathbf{x}_1, \ldots, \mathbf{x}_R$ are conditionally independent. The Sum rule is inferred from the Product rule under the assumption that the posterior probabilities $P(C_j|\mathbf{x}_i)$ estimated by classifiers do not deviate significantly from the prior probabilities.

As compared against the Product rule, the Sum rule inference involves more assumptions and therefore may seem to be less realistic. However, as was shown by Kittler et al. (Kittler et al., 1998), the Sum rule is more robust to the errors in the estimates of the classifiers. As a result, its use is justified when the distributions of the values of the features are estimated by classifiers with a large error (Kittler et al., 1998; Tax et al., 2000).

The notational form of the Sum rule is similar to the MP rule above in the sense that both the MP and Sum rules are based on the sums of probabilities. Moreover, the MP rule can be derived from the Sum rule, as will be described below. This suggests that, by analogy with the results of Tax et al. (Tax et al., 2000) and Kittler et al. (Kittler et al., 1998) comparing the Product rule and the Sum rule, the MP rule is beneficial if the probability distributions are estimated by the classifiers with a great error.

In the context of the user substitution detection problem, the classifiers deal with peculiarities of human behavior, which is prone to changes over time. In addition, the data set available for learning the classifiers is usually quite limited. Therefore, it is likely that the error with which each classifier estimates the probability distribution will not be negligible. Then, the use of the MP rule may be justified as a robust combining rule for one-class classifiers in this context.

It is necessary to note that the MP rule was presented as a combining rule to be used with classifiers dealing with same feature spaces (assumption A2). In the substitution detection context, the classifiers are expected to work with different aspects of user behavior and, thus, they mainly use different sets of features thereby invalidating this assumption. However, two arguments can be presented for the use of the MP rule when the feature sets are different:

1. While classifications produced by different classifiers are based on different features, all the clas-

sifiers attempt to estimate the same probability $P(Z \in C_U)$, i.e. the probability that an object $Z$ belongs to the user class. Then the same reasoning as was adopted for the inference of the MP rule may be applied. Namely, using the above assumption (A1) and assuming the zero-mean estimation error of the classifiers, the averaging (i.e. the MP rule) may be employed to suppress the error.

2. The MP rule can be derived from the Sum rule wherein the equality of the feature spaces is not assumed. In the general case of $M$ classes, the Sum rule can be presented in a form (Kittler et al., 1998, p. 228):

$$\text{Decide} \quad Z \in C_m \quad \text{if}$$
$$(1-R)P(C_m) + \sum_{i=1}^{R} P(C_m|\mathbf{x}_i) =$$
$$\max_{j=1}^{M}[(1-R)P(C_j) + \sum_{i=1}^{R} P(C_j|\mathbf{x}_i)], \quad (3)$$

where $P(C_j),\ j = 1, 2, \ldots, M$ denotes prior class probabilities.

As could be seen, for every class $j$ the term $(1-R)P(C_j)+\sum_{i=1}^{R} P(C_j|\mathbf{x}_i)$ is calculated, and the object $Z$ is to be assigned to the class corresponding to the maximum value of the term. In one-class classification situation, the classifiers provide their classifications concerning only one class. Therefore, instead of searching maximum value, the comparison with a threshold $t$ may be performed in order to make a final classification.

By substituting the search for maximum with the comparison against a threshold, and by applying the above assumption (A1) to the Sum rule, the rule can be rewritten in a form:

$$\text{Decide} \quad Z \in C_U \quad \text{if}$$
$$(1-R)P(C_U) + \sum_{i=1}^{R} p(\mathbf{x}_i|C_U) \geq t_{mp},$$
$$\text{otherwise decide} \quad Z \in C_I. \quad (4)$$

which in essence represents the MP rule. The term $(1-R)P(C_U)$ is a constant; therefore, it could be united with the threshold. Similarly, the term $R^{-1}$ in equation (1) is a normalization factor which does not influence the final decision provided the threshold $t_{mp}$ is properly adjusted.

Thus, the MP rule can be seen to be a special case of the Sum rule when the classifiers to be combined are one-class ones. Consequently, the MP rule is expected to hold the advantage of the Sum rule when the probability distributions are poorly estimated by the classifiers.

The MP rule represents the average of the estimated probabilities of the values of the features. These estimated probabilities can be thought of as the approximations of the classifiers' *confidences* in the hypothesis that the object belongs to the user class. In turn, the outcome of the rule represents the average of the classifier confidences. This rule was inferred for combining the classifiers operating on the same feature space. However, when classifiers based on different feature spaces are to be combined, the estimated probabilities of the values of the features $p(\mathbf{x}_i|C_U)$ may be inefficient approximations of the classifiers' confidences. This is because the values of density functions depend on the unit of measure applied to a feature. If a measure has a unit $x$, then the output of pdf has the unit $1/x$. The features of different nature are likely to have different units of measure; moreover, different classifiers may be based on unequal number of features. As a result, the classifiers may apply different scales (e.g., the maximum for one may be less than the minimum value for another). Averaging the terms having different scales will result in a loss of information. Consequently, the accuracy of the final classification may become worse.

Thus, in order to improve the classification accuracy, it is necessary to make the classifier confidence dimensionless. In next section, a modified version of calculating the confidence value is introduced to address the above problem.

## 3 MODIFIED MP COMBINING RULE

As discussed in previous section, the probability estimates $p(\mathbf{x}_i|C_U)$ may have different scale for different classifiers depending on the nature of features and the number thereof. As a result, their averaging, as it is done in the MP rule, may be inefficient.

Therefore, it is desirable to replace an estimate $p(\mathbf{x}_i|C_U)$ with a dimensionless measure $u_i$ representing the degree of the classifier's confidence in the hypothesis that an object $Z$ belongs to the user class. With respect to the user substitution detection problem, the confidence value reflects how sure a classifier is that the legitimate user is interacting with the device.

Since the classification produced by a classifier is based on the estimated pdf of feature values, the confidence should be a function of this pdf. The combination rule for the classifications may be taken as the average of the classifier confidences:

$$u_{mc}(\mathbf{x}_1, \ldots, \mathbf{x}_R) = R^{-1} \sum_{i=1}^{R} u_i(p(\mathbf{x}_i|C_U)). \quad (5)$$

In order to be dimensionless, the confidence value can be calculated as a ratio of the estimated probability $p(\mathbf{x}_i|C_U)$ to its mean value $\overline{p}(\mathbf{x}_i|C_U)$. This mean value is equal to the probability of a random variable uniformly distributed in the feature space $\mathcal{X}_i$. This ratio is between zero and one when the estimated probability is less than its mean value, and is greater than one otherwise. In turn, when $u_i = 1$ the classifier $i$ can be said to have no arguments in favor or against the claim that an object $Z$ belongs to the user class (as it is in case the values of the features are uniformly distributed).

Further, in order to make the confidence more symmetric around the "no argument" value, the logarithm of the above ratio can be taken. This rescales the confidence value from the interval $[0, \infty)$ to the interval $(-\infty, \infty)$; the "no argument" case corresponds to the zero value of the confidence.

Finally, sigmoid transformation (Bishop, 1995) can be applied to map the confidence values into $(0, 1)$ interval. The produced confidence value can be calculated as

$$\begin{aligned} u_i(p(\mathbf{x}_i|C_U)) &= \frac{1}{1+\exp\left(-\ln\frac{p(\mathbf{x}_i|C_U)}{\overline{p}(\mathbf{x}_i|C_U)}\right)} = \\ &= \frac{p(\mathbf{x}_i|C_U)}{p(\mathbf{x}_i|C_U)+\overline{p}(\mathbf{x}_i|C_U)}. \end{aligned} \quad (6)$$

When the confidence value is close to one, the classifier is convinced of the presence of an object of the user class. Contrary, confidence values close to zero indicate the negligible classifier's confidence in the hypothesis that the object belongs to the user class. The value of 0.5 corresponds to the "no argument" case.

The use of this transformation function allows the confidence value to be interpreted as an approximation of the posterior probability. Indeed, using Bayes formula, it follows that the expression for confidence value (6) is equal to the posterior probability $P(C_U|\mathbf{x}_i)$ assuming that i) impostor cases are uniformly distributed in the feature space, and ii) the prior class probabilities are equal.

In this section, the modified version of the MP rule was proposed. This rule uses averaging over classifier confidences as dimensionless values. As a result, better classification accuracy is expected. In next section, the advantage of the proposed version over the basic MP rule will be evaluated.

# 4 PERFORMANCE EVALUATION

In this section, we compare the performance of the modified MP rule with the performance of the original MP rule. Two characteristics are often used to evaluate the performance (namely, accuracy) of a classifier distinguishing between a user and impostors. These are the false acceptance (FA) and false rejection (FR) error rates, denoted as $P_{FA}$ and $P_{FR}$ respectively. A false acceptance occurs when an impostor is classified as a legitimate user, and a false rejection occurs when a legitimate user is classified as an impostor. Another related measure is the probability of correct detection $P_D = 1 - P_{FA}$. The ideal performance is achieved when $P_D = 1$ and $P_{FR} = 0$. However, the ideal performance is usually impossible to achieve with real-world classifiers, and therefore, a tradeoff between the $P_D$ and $P_{FR}$ values is commonly set as a goal. The dependence between $P_D$ and $P_{FR}$ values can be represented by a so-called receiver-operating curve (ROC-curve) (Swets, 1988) that plots the $P_D$ values as a function of $P_{FR}$. The area above the curve characterizes the performance of a classifier; the smaller the area the better the performance. That is, the greater the probability of detection for a given false rejection rate is, the better is the performance of the classifier.

In order to plot a ROC-curve either for a single classifier or after combining several classifiers, the $P_D$ and $P_{FR}$ values are expressed as functions of a threshold value $t$:

$$\begin{aligned} P_D &= \int_{u(\mathbf{x})<t} p_{impostor}(\mathbf{x})d\mathbf{x}, \\ P_{FR} &= \int_{u(\mathbf{x})<t} p_{user}(\mathbf{x})d\mathbf{x}, \end{aligned} \quad (7)$$

where $p_{user}(\mathbf{x})$ and $p_{impostor}(\mathbf{x})$ denote the pdfs of feature values for the user and the impostor classes respectively, and $u(\mathbf{x})$ is the classification of a single classifier or the classification produced using a combining rule for several single classifiers.

In user substitution detection, the classifiers dealing with various behavioral characteristics as typing, application and service usage, etc. are to be employed. Unfortunately, neither the characteristics of such classifiers nor the rough data describing the above behavioral aspects in the context of mobile-device users are publicly available. Therefore, hypothetic classifiers are simulated in order to evaluate the modified MP rule. While the characteristics of these classifiers are likely to differ from the characteristics of classifiers to be employed in user substitution detection, it is expected that the difference is not critical since the modified and basic MP rules are abstract combining rules, and hence their characteristics are likely to hold for a variety of individual classifiers being combined.

In the following experiments, three classifiers are combined using the original MP rule and its proposed modification. Two different cases are investigated. First, hypothetical classifiers with noticeably different performance characteristics are studied. Second,

three classifiers whose characteristics are set according the classifiers employed for multi-modal user authentication are considered. In both cases, ROC-curves are used to compare the performance of combined classifiers.

In first case, three classifiers with noticeably different performance characteristics are combined. Both the user class and the impostor class are assigned statistical models of feature distributions described by normal continuous pdfs. Each classifier is based on only one feature and the features are assumed to be independent. To make the feature spaces bounded, the feature distributions are limited by the intervals $[a, b]$ with the density functions being normalized to integrate to unity. Note that while combining one-class classifiers does not involve any knowledge of the impostor class pdfs, they are needed to be able to evaluate the performance of the final combined classification. The characteristics of the hypothetical classifiers are given in Table 1.

Table 1: Characteristics of the classifiers with noticeably different characteristics

| Classifier | Model of the user class | | Model of the impostor class | | Bounds |
|---|---|---|---|---|---|
| | Mean | St. deviation | Mean | St. deviation | |
| 1 | 0 | 1 | 2 | 1.3 | [-2.6, 5.2] |
| 2 | 0 | 2 | 2 | 2.3 | [-2.6, 5.2] |
| 3 | 0 | 3 | 2 | 3.3 | [-2.6, 5.2] |

For all classifiers (more precisely, for all feature spaces), the user class distribution is spikier than the impostor class one. The distances between the means of the user and impostor classes are equal for all the classifiers. The difference of classifiers' performances is induced by the distinction of the standard deviations. The characteristics of the classifiers were intentionally selected so that their performances differ significantly. Correspondingly, the pdfs of the feature values estimated by classifiers have different scales.

For the second case, the classifiers' characteristics are assigned according published values (Verlinde et al., 2000). The corresponding classifiers analyze respectively a profile image, a frontal image, and voice characteristics of a user in order to verify his or her identity (Kittler et al., 1998). Each classifier uses an appropriate similarity measure to compare the measured values of the features against the corresponding values of a legitimate user as described in a previously established profile. In fact, every classifier transforms a multidimensional vector of input feature values into a one-dimensional output value indicating how likely the input vector is the one of the legitimate user. It was shown (Verlinde et al., 2000) that distributions of the classifiers' outputs might be approximated by normal distributions with the parameters shown in Table 2.

Table 2: Characteristics of the mono-modal identity verification classifiers

| Classifier | Model of the user class | | Model of the impostor class | | Bounds |
|---|---|---|---|---|---|
| | Mean | St. deviation | Mean | St. deviation | |
| Profile | 0.945 | 0.03 | 0.7 | 0.26 | [0, 1] |
| Frontal | 0.861 | 0.09 | 0.571 | 0.13 | [0, 1] |
| Vocal | 0.923 | 0.04 | 0.65 | 0.13 | [0, 1] |

In Figure 1, the ROC-curves of the hypothetical classifiers with noticeably different performance characteristics are shown along with the ROC-curves corresponding to the final classification produced by the original MP rule and by its modified version.

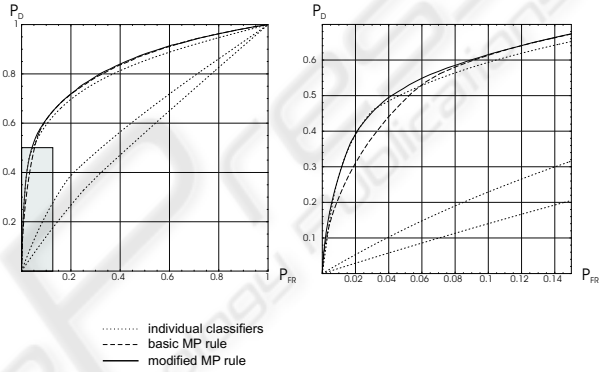


Figure 1: Results of combining three hypothetical classifiers with noticeably different characteristics. The right part of the figure is a magnification of the shadowed area

As illustrated by the figure, the modified MP rule outperforms the original MP rule for all reasonable $P_D$ or $P_{FR}$ values. The combined classification using the original MP rule may result in a performance that is poorer than the performance of a single best classifier as can be seen in the right part of the figure. At the same time, the performance of combined classification with the modified MP rule is at least comparable to the best classifier's performance.

Figure 2 illustrates the performance of the profile, frontal and vocal classifiers as well as the performance provided by applying the original MP rule and its modified version.

As can be seen, in this case both combining rules improve the classification accuracy as compared with any single classifier. At the same time, the modified MP rule outperforms the original MP rule. The difference between them is especially remarkable for low values of FR rate (less than 0.1) as shown in the right-hand part of Figure 2.

Two conclusions may be made from the results above:

- For reasonable values of FR rate, the classification accuracy achieved with the modified version of the

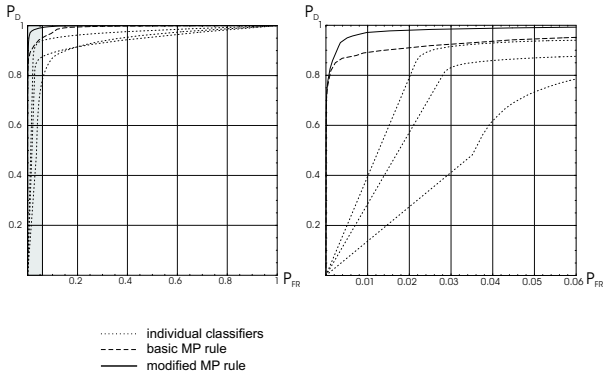


Figure 2: Results of combining profile, frontal, and vocal classifiers. The right part of the figure is a magnification of the shadowed area

MP rule is superior to the accuracy achieved with the original MP rule;

- In a situation when the combining classifiers using the MP rule results in worse classification accuracy than a single best classifier, the modified MP rule may still be beneficial.

Thus, the above results support the hypothesis that the modification of the MP rule wherein the estimations of classifiers' confidences are made dimensionless may be used to improve the overall classification accuracy compared to the original MP rule. At the same time, the modified MP rule still remains a sum-based one. Hence, the use of the modified MP rule may be justified when several heterogeneous classifiers are to be combined and their estimations of class-conditional probabilities of feature distributions are tampered with a noise.

# 5 DISCUSSION AND CONCLUSIONS

Above, the modified MP rule was introduced and its performance was compared with the performance of the original MP rule. In this section, the pros and cons of the modified MP rule in the context of mobile-user substitution detection are discussed, and topics for further research are outlined.

Two main advantages of the modified MP rule were already mentioned. First, the rule is robust to the classifiers' estimation errors that are expected to be significant in the case of the classifiers dealing with the behavioral aspects of a user. Second, the modified MP rule outperforms, at least in some situations, the original MP rule, mainly because the classifications of designated single classifiers are made independent on the dimensionality of the underlying features. Third, the modified MP rule appears to be superior with respect to the original MP rule for small FR error values. Keeping the FR error rate low is one of the essential requirements set by users to any substitution detection technique. Fourth, the modified MP rule can be made to take benefit of information about impostor behavior distribution when it exists. In the current version, as was explained above, the classifier's confidence value can be thought of as an approximation of the posterior probability $P(C_U|\mathbf{x}_i)$ assuming that impostor cases are uniformly distributed in the feature space using the constant probability density value $\overline{p}(\mathbf{x}_i|C_U)$. The assumed uniform distribution of impostor cases can be replaced with a better approximation provided relevant information about impostor behavior is available. For instance, if it is known that the impostor cases are normally distributed, then that constant value $\overline{p}(\mathbf{x}_i|C_U)$ is replaced with the appropriate normal pdf. This may further improve the classification accuracy.

There are at least three limitations inherited in the modified MP rule. First, the derivation of the original MP rule and its modified version assumes zero-mean estimation error of the classifiers. Should the mean be far from zero, combining classifiers using the MP rule will not suppress the error. Second, if the classifiers' estimation errors are (positively) correlated, then, even if they have zero mean, the averaging used in the MP rule may result in no benefits. Note however that the same two limitations hold for the original MP rule, too. Third, the incorporation of additional information about impostor distributions, when available, should be done with care. If the distribution is approximated with a significant error, then, in fact, the uniform distribution may appear to be a better approximation, and the use of erroneous approximation may result in worse final classification than the use of the constant value $\overline{p}(\mathbf{x}_i|C_U)$.

In justifying the use of the proposed modified MP rule in the user substitution detection, the hypothesis was made that the individual classifiers estimate the pdfs of the feature values of human behavior with a great error. In further work we plan to use real data describing mobile user behavior to test this hypothesis and to evaluate the practical capabilities of the proposed modified MP rule. Further work should also address the problem of possible positive correlation between the errors of individual classifications. Another topic for further research is to consider which available classifiers (or, more precisely, which available classifications) should be taken into account during the combining process.

# REFERENCES


Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 37–46. ACM Press.

Anderson, D., Lunt, T., Javitz, H., Tamaru, A., and Valdes, A. (1995). Detecting unusual program behavior using the statistical components of NIDES. SRI Technical Report SRI-CRL-95-06, Computer Science Laboratory, SRI International.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Burge, P. and Shawe-Taylor, J. (1997). Detecting cellular fraud using adaptive prototypes. In *AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management*, pages 1–8. AAAI Press.

Cahill, M., Lambert, D., Pinheiro, J., and Sun, D. (2000). Detecting fraud in the real world. Technical report, Bell Labs, Lucent Technologies.

Clarke, N. L., Furnell, S. M., Rodwell, P. M., and Reynolds, P. L. (2002). Acceptance of subscriber authentication methods for mobile telephony devices. *Computers & Security*, 21(3):220–228.

Dasarathy, B. V. (1994). *Decision Fusion*. IEEE Computer Society Press.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. (2002). *Data Mining for Security Applications*, chapter A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data. Kluwer.

Kittler, J. and Alkoot, F. (2000). Multiple expert system design by combined feature selection and probability level fusion. In *Proceedings of the Fusion'2000, Third International Conference on Information Fusion*, volume 2, pages 9–16.

Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Kumar, S. (1995). *Classification and Detection of Computer Intrusions*. Ph.D. thesis, Purdue University.

Kuncheva, L. (2002). A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286.

Manganaris, S., Christensen, M., Zerkle, D., and Hermiz, K. (2000). A data mining analysis of RTID alarms. *Computer Networks*, 34(4):571–577.

Monrose, F. and Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computing Systems (FGCS) Journal: Security on the Web (special issue)*.

Samfat, D. and Molva, R. (1997). IDAMN: An intrusion detection architecture for mobile networks. *IEEE Journal on Selected Areas in Communications*, 7(15):1373–1380.

Schonlau, M., DuMouchel, W., Ju, W., Karr, A., Theus, M., and Vardi, Y. (2001). Computer intrusion: Detecting masquerades. *Statistical Science*, 16(1):58–74.

Seleznyov, A. (2002). *An Anomaly Intrusion Detection System Based on Intelligent User Recognition*. Ph.D. thesis, Department of computer Science and Information Systems, University of Jyvskylä, Finland.

Sequeira, K. and Zaki, M. (2002). ADMIT: anomaly-based data mining for intrusions. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 386–395, Edmonton, Alberta, Canada. ACM Press.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1289.

Tax, D. (2001). *One-class classification*. Ph.D. thesis, Delft University of Technology.

Tax, D. and Duin, R. (2000). Experiments with classifier combining rules. In *MCS 2000*, volume 2 of *Lecture Notes in Computer Science*, pages 16–29. Springer-Verlag.

Tax, D., van Breukelen, M., Duin, R., and Kittler, J. (2000). Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485.

Valdes, A. and Skinner, K. (2000). Adaptive, model-based monitoring for cyber attack detection. In Debar, H., Me, L., and Wu, F., editors, *Recent Advances in Intrusion Detection (RAID 2000)*, number 1907 in Lecture Notes in Computer Science, pages 80–92, Toulouse, France. Springer-Verlag.

Verlinde, P., Chollet, G., and Acheroy, M. (2000). Multi-modal identity verification using expert fusion. *Information Fusion*, 1(1):17–33.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Xu, L., Krzyzak, A., and Suen, C. Y. (1992). Methods for combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435.

Yamanishi, K., Takeuchi, J.-I., Williams, G., and Milne, P. (2000). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 320–324. ACM Press.

Ye, N. and Chen, Q. (2001). An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 17(2):105–112.