Synergistic Classifier Fusion for Security Applications

Paul F. Evangelista
United States Military Academy
West Point, NY
paul.evangelista@usma.edu

Mark J. Embrechts, Boleslaw K. Szymanski
Rensselaer Polytechnic Institute
Troy, NY
embrem@rpi.edu, szymansk@cs.rpi.edu

ABSTRACT

Unbalanced, high-dimensional, binary classifications create challenges in a variety of systems and environments, most notably within physical security and computer security domains. The imbalance within these problems consists of a significant majority of the negative (healthy, non-intruding) class and a minority (unhealthy, intruding) positive class. Any system that needs protection from malicious activity, intruders, theft, or other types of breaches in security must address this type of problem. Given numerical data that represent observations or instances which require classification, many practitioners apply state of the art machine learning algorithms to aid in solving unbalanced classification problems. The unbalanced and high-dimensional structure of the data can trouble these learning methods. High-dimensional data poses a ``curse of dimensionality" which can be overcome through subspace modeling and intelligent fusion. A fundamental method for evaluation of the binary classification model is the receiver operating characteristic (ROC) curve and the area under the curve (AUC), and the intelligent fusion employed ties directly with the properties of this evaluation method. This work exposes the underlying statistics involved with ROC curves and leverages these properties to create synergistic classifier fusion through rankings. Decision ROC charts are a novel illustration that augment the ROC curve to provide a more complete representation of the classifier performance. Pseudo-ROC curves, created from simulated rankings utilizing principles based on the Wilcoxon-Rank sum or Mann-Whitney *U* statistic, provide novel insight into the behavior of classifier rankings. The critical finding involves the unique behavior of rankings for unbalanced classification problems and methods to capitalize on this behavior to improve classifier accuracy for unbalanced problems. Arguments presented include theoretical discussion, proof of principle through simulated classifier rankings examined with a factorial design, and experimental results with actual data including host-based and network-based intrusion detection datasets.

ABOUT THE AUTHORS

Paul F. Evangelista, Ph.D., is a Major in the US Army, currently serving as an Assistant Professor in the Department of Systems Engineering at the United States Military Academy, West Point, New York. He was commissioned as an Engineer and has served in numerous troop assignments from platoon to battalion level. His research interests include machine learning and computational intelligence.

Mark J. Embrechts, Ph.D., is an Associate Professor in the Department of Decision Sciences and Engineering Systems at Rensselaer Polytechnic Institute. He has joint appointments in the Department of Mechanical, Aerospace, and Nuclear Engineering and in the Department of Information Technology. He has pioneered courses in Neural Networks, Computational Intelligence, and Data Mining. His current interests relate to scientific data mining and applications of data mining to biotechnology and drug design.

Boleslaw K. Szymanski, Ph.D., is a Professor in the Department of Computer Science at Rensselaer Polytechnic Institute. He is the director for the Center of Pervasive Computing and Networking at Rensselaer, an IEEE Fellow, and editor in chief of *Scientific Programming*. His research interests include parallel processing, network-centric computing, algorithm design, programming languages, and operating system.

Synergistic Classifier Fusion for Security Applications

Paul F. Evangelista
United States Military Academy
West Point, NY
paul.evangelista@usma.edu

Mark J. Embrechts, Boleslaw K. Szymanski
Rensselaer Polytechnic Institute
Troy, NY
embrem@rpi.edu, szymansk@cs.rpi.edu

INTRODUCTION

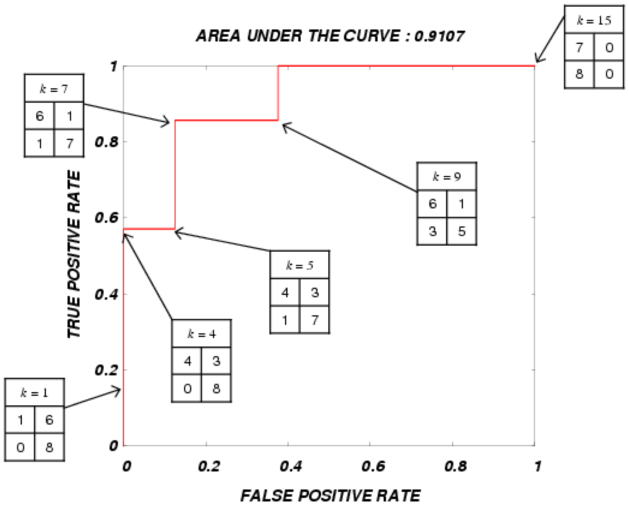
We live in a world of black box solvers and turn-key solutions. When the black boxes lose accuracy and the turn-key solutions struggle to fire the engine, knowing what happens inside the box, or what happens when you turn the key, enables anticipation and postures us for rapid problem solving. What better way is there to maintain the edge than to improve our ability to anticipate and speed problem solving? There is nothing wrong with harnessing the power of technological advances. There is also nothing wrong with improving our understanding of the first principles which support this technology. These first principles guide the current solutions, and our understanding of these first principles could help build new and improved solutions. Engineers and researchers have applied these first principles of signal detection to military applications for over half a century. The utility of these principles recently garnered attention in state of the art machine learning and data mining research communities, creating an interesting opportunity to merge knowledge and applications.

The history of signal detection illustrates the age and persistence of unbalanced classification problems. During the air defense battles of World War II, the value of radar and signal detection grew at an explosive rate. Radar techniques were relatively primitive and required significant human interaction and interpretation. The essence of the problem was simple. Radar detected incoming aircraft. A radar operator needed to be able to distinguish between friendly and enemy aircraft. Identifying a friendly aircraft as enemy (a false positive), created an expensive sequence of drills and defensive responses. This created the potential for fratricide with inbound friendly aircraft. There was also the danger of not alerting when actual enemy aircraft were inbound, a false negative. Radar represents one of the earliest signal detection problems which required humans to interact and interpret a technical measure, the radar signal, with the overall goal of classifying the observation as one of two classes. This is a binary classification problem. In order to measure the effectiveness of radar operators, the military recorded the performance of these radar operators. This performance measure became known as the radar receiver operating characteristic illustrated on the receiver operating characteristic (ROC) curve (Fan et. al, 2005). Radar became one of the earliest applications of signal detection theory. After World War II, atomic weapons boosted the importance of air defense. In the 1950s tremendous research efforts, such as the MIT led Project Charles, communicated the vast problems and gaps that existed in national air defense. The final report of Project Charles, originally a classified document, emphasized that in order to improve air defense the program would need significant manpower and a deliberate layered detection strategy in order to overcome costly false positives (Loomis, 1951).

Much of the enemy aircraft threat existed in remote areas where vast expanses, such as oceans or harsh northern territories further constrained the nation's ability to man an adequate air defense system. This marked the beginning of a long quest to automate signal detection. Naka and Ward explain the history of air defense coverage across Alaska and Canada and the critical need for an automated system that could defend this enormous territory (Naka and Ward, 2000).

Although ROC curves and automated signal detection has matured immensely since its inception over half a century ago, there is a continued effort to improve automation and vast use of ROC curves to support various types of binary decisions. Contributions of this work include novel methods that involve the automation and accuracy of binary prediction models. Although the applications presented do not involve radar and air defense, many similar threads exist between early research of automated binary decision making and this work. Today's binary classification problems have increased complexity and dimensionality, however today's prediction methods and computing resources provide tools and leverage necessary to tackle the problems. There is still a continued quest to automate these systems, reduce false positives, and simply improve overall accuracy. In many ways, this quest is no different from the quest that Naka and Ward discuss when they explain one of the original quests for automated signal detection that

Decision Value ($\hat{y}_i$)	Rank (R_i)	True Class (y_i)
2.893	1	1
2.208	2	1
1.664	3	1
0.991	4	1
0.889	5	-1
0.609	6	1
0.015	7	1
0.013	8	-1
-0.240	9	-1
-0.278	10	1
-0.808	11	-1
-1.257	12	-1
-1.437	13	-1
-1.750	14	-1
-1.864	15	-1
$U = W_{YX}/pb$		0.9107



		Predicted Label	
		positive	negative
		k = threshold	
Actual	positive	TP	FN
Label	negative	FP	TN

Figure 1. Evolving decision values into confusion matrices and an ROC curve. The key for the confusion matrices is shown in the lower right table that compares predicted and actual labels.

occurred over half a century ago (Naka and Ward, 2000).

Notation

This work applies to binary classification problems. Assume a matrix of real numbers, $\mathbf{X} \in \mathfrak{R}^{N \times m}$, as a given data set. **X** contains N instances or observations,
$\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_N$, where $\mathbf{x}_i \in \mathfrak{R}^{1 \times m}$. There are m variables to represent every instance. For every instance there is a label or class, $y_i \in \{-1,+1\}$. Predicted real valued labels will be referred to as $\hat{y}_i \in \mathfrak{R}^1$.

ROC Curve Background and ROC Decision Charts

ROC curves are a two dimensional graph and popular method to display the performance of a binary classification system. This curve plots the true positive (*TP*) rate on vertical axis and the false positive (*FP*) rate on the horizontal axis. The confusion matrix is closely related and essentially a subset of an ROC curve. The confusion matrix displays four numbers - true positive(*TP*), true negative(*TN*), false positive(*FP*), false negative(*FN*) - which illustrate the prediction performance of a classifier at a specific threshold. Figure 1 shows several confusion matrices at different points in an ROC curve. A confusion matrix can also display the accuracy of a multi-class problem, often illustrating model tendencies and which classes the model tends to ``confuse" (Kohavi and Provost, 1998).

It is also well known that the area under the curve (AUC), a value between 0 and 1, has a special probabilistic meaning. The AUC is equivalent to the probability that a randomly selected positive instance ranks above (has a smaller R_i) than a randomly selected negative instance. Let $R(\mathbf{x}_i), i \in (1,2,...,p)$ represent the rank of positive instances, and $R(\mathbf{x}_j), j \in (1,2,...,b)$ represent the rank of negative instances. This probability is also known as the Mann-Whitney *U* statistic as shown in equation 1.

$$U = P\{(R(\mathbf{x}_i) \mid y_i = 1) < (R(\mathbf{x}_j) \mid y_j = -1)\} \quad (1)$$

Much of the original study of rank statistics that apply to ROC curves can be attributed to Wilcoxon, Mann, and Whitney who used ranks to measure whether or not two random variables were statistically different. Wilcoxon's work was a cursory exploration to “obtain a rapid approximate idea of the significance of the differences in experiments” (Wilcoxon, 1945). The dataset utilized in his experiments measured the lethality of two different fly sprays. Mann and Whitney also used rank statistics to determine the statistical difference between two treatments, with much more of an emphasis on describing the underlying distribution of the *U* statistic (Mann and Whitney, 1947). Lehmann provides a comprehensive study of rank statistics in his book *Nonparametrics: Statistical Methods Based on Ranks*, thoroughly discussing the Wilcoxon Rank Sum statistic and Mann-Whitney *U* statistic (Lehmann, 1975).

The Wilcoxon Rank Sum Statistic, W_s, is equivalent to $\sum_{i=1}^{p} R_i$. Since the sum of all rankings is $\frac{1}{2}N(N+1)$, it follows that the sum of non-attack instance rankings is $W_r = \frac{1}{2}N(N+1) - W_s$. W_r can also be calculated as $\sum_{j=1}^{b} R_j$ (Lehmann, 1975). The statistics $W_{XY} = W_s - \frac{1}{2}p(p+1)$ and $W_{YX} = W_r - \frac{1}{2}b(b+1)$ are also popular forms of the Wilcoxon Rank Sum statistic, and it is this form of the statistic that relates to the area under the ROC curve. The Mann-Whitney *U* statistic, which is exactly equal to the area under the receiver operating characteristic curve, is directly proportional to the Wilcoxon rank sum statistic W_{YX} where $U = W_{YX} / pb$. Hanley and McNeil show in (Hanley and McNeil, 1982) that the area under the ROC curve equates to the Mann-Whitney *U* statistic.

Although ROC curves convey significant information, it is important that researchers understand the limitations and properties of these curves. Many authors, to include (Fawcett, 2003), (Bradley, 1997), and (Mason and Graham, 2002), advise due caution when using ROC curves to measure the performance of binary classifiers. ROC curves are non-parametric. There is value in the simplicity and pragmatism of displaying a classifier's output as non-parametric ranks, however it is without doubt that information is lost when we reduce the decision value from the classifier to a rank.

The eventual purpose of an ROC curve is to support a decision, however ROC curves provide an incomplete representation of the decision environment. Important information not included in ROC curves includes balance of classes and decision values. Practitioners value ROC curves for the insight that these curves provide to a classifier or detection device, however applying information from the ROC curve is difficult without some type of translation from ROC space to the decision space. The decision space for a binary classification problem involves considering the output of a model, $\hat{y}_i$, and comparing this output with a threshold, *t*.

Classifiers create decision values. These decision values, $\hat{y}_i$, exist on a spectrum in the real number realm ($\hat{y}_i \in \mathcal{R}^1$), where $\hat{y}_i$ represents nothing more than the classifier's judgment on the class membership of an observation. Classifiers base this judgment on some type of function created from other observations, so the decision value becomes the classifier's similarity or strength of belief metric indicating class membership. Many decision values follow some type of sigmoidal function, clustering in the center and existing in a lower density at the extreme ends of the spectrum. For some types of detection equipment, decision values could represent the output of a sensor.

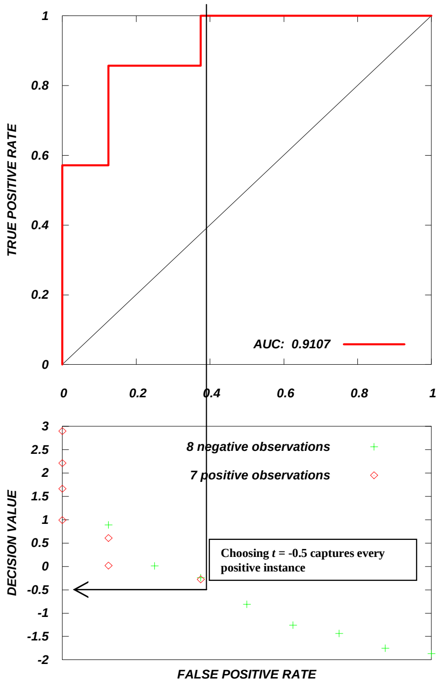


Figure 2. Decision ROC Chart displayed for the data in figure 1.

This could be a chemical measurement as in a medical test or an electronic signal measurement which exists in many types of detection scenarios. Regardless of the classification problem, the end result is that a decision must be made whether to classify an unlabeled instance as positive or negative. ROC curves illustrate the performance of a classifier, but they do not provide a complete picture or aid in classification decisions. Practitioners must bridge the gap between ROC curves and decision values.

The *Decision ROC Chart* is an extension to the ROC curve. It is a simple yet novel method to aid decision makers and assist in bridging the gap from ROC curves to decision values. The horizontal axis of an ROC curve measures the false positive rate of a classifier, and this false positive rate is simply a fraction of negative instances where the false positive rate equals $FP/(FP + TN)$. The extension to the ROC curve, referred to as the *Decision ROC Chart*, plots the false positive rate on the horizontal axis, enabling the bridge between the ROC curve and the decision value. The vertical axis represents the decision value, $\hat{y}_i$.

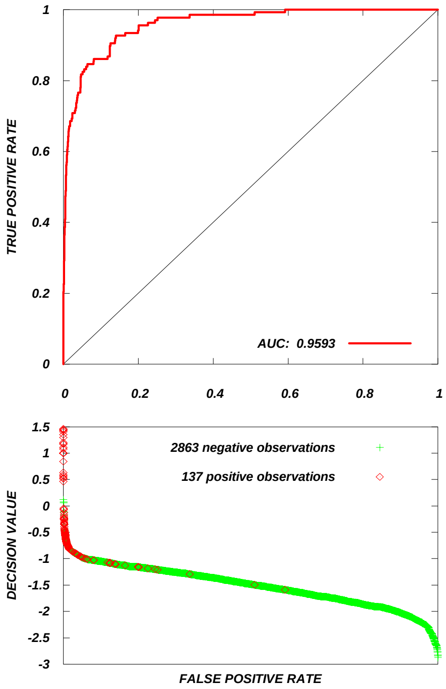


Figure 3. Decision ROC Chart for a large dataset.

The *Decision ROC Chart* captures much of the information absent from the ROC curve, to include the balance of the problem, distribution of the decision values, and perhaps most importantly the connection between the ROC curve and the decision value through the false positive rate. Figure 2 illustrates the *Decision ROC Chart* for the toy running problem from figure 1. Figure 3 illustrates the *Decision ROC Chart* for a larger dataset. In addition to aiding the threshold decision for a classification problem, the Decision ROC Chart provides clarity to ROC curves for those who are not familiar with this measure. The area under the curve is visibly seen as a probability with the *Decision ROC Chart*. It is much more believable, and understandable, that the probability of a positive value out ranking a negative value is 0.959 when illustrated as shown in figure 3.

Decision ROC Charts educate decision makers and students involved with classification. The purpose of these curves is simply to educate, assist in decision making, and provide a more complete picture of a decision environment.

Fusion of Multiple Models

The major contribution of this work involves introducing how model fusion for unbalanced datasets performs differently than model fusion for balanced data. Exposing this difference provides researchers with an additional parameter, the balance of the data, which can be considered when building ensembles of classification models.

Consider several sensors responsible for detecting some type of anomalous behavior. The sensors serve as sentries to a larger system. Suppose that every sensor reacts to every observation, evaluating or ranking the observation based upon a history of known behavior. Suppose that for each observation, some of the sensors have an opportunity to closely observe and measure an observation (a ``good" measurement), and some of the sensors remotely observe (a ``poor" measurement). All of the sensor measurements will be considered for the decision, and it is unknown which sensors closely observe and which ones remotely observe. Questions emerge from such a scenario. How should the measurements of these sensors fuse to create the best signal? What other considerations regarding the observed population should be included when choosing the fusion method? What other information should be considered when choosing the fusion method for this situation? These and several other questions will be addressed both through theoretical explanation and empirical results.

This scenario is also one potential application of synergistic classifier fusion. Synergistic classifier fusion is an ensemble technique designed for the unbalanced classification problem. Synergistic classifier fusion uses one additional piece of information to improve performance - assumed imbalance of the classes. With this simple assumption, it is possible to take advantage of the behavior of rank distributions and use `min` or `max` aggregators for synergistic performance. The experiments discussed in this work consider the classic case of model ensembles. With model ensembles, fusion of ranks is all the more important because the underlying distributions of the model decision values are unknown. Several important novelties stem from this work:

1. *Pseudo ROC curves*. ROC curves are a performance metric. However, it is also possible to examine the underlying statistics which create ROC curves to better understand classifier behavior. Typically ROC curves are built from a classification scenario and simply observed for what they are. However, how does an ROC curve with an AUC of 0.7 differ from an ROC curve with an AUC of 0.9? How does an ROC curve with an AUC of 0.9 that measures a classification problem with a 90% negative class differ from an ROC curve with an AUC of 0.9 that measures a classification problem with a 50% negative class? These questions can be explored with Pseudo ROC curves.

2. *Rank distributions from pseudo ROC curves*. Rank distributions illustrate the behavior of classifiers from a non-parametric position. ROC curves are non-parametric, and the underlying distributions of the ranks which create ROC curves are non-parametric. When comparing two classifiers, comparing them with non-parametric statistics makes sense. The underlying distribution of classifier decision values is unknown - using (non-parametric) ranks enables comparison of classifiers on a level field. These rank distributions also lead to consideration of the `max` and `min` aggregation or fusion metrics.

3. *The* `min` *and* `max` *aggregators provide robust classifier fusion for unbalanced classification problems*. This chapter will discuss the behavior of rank distributions for unbalanced classification problems - rank distributions of unbalanced classification problems behave with different likelihoods (discrete rank probabilities) than a balanced problem. It is possible to take advantage of this difference in likelihoods to improve classification.

Several included experiments illustrate synergistic classifier fusion, and underlying theory explains why this synergistic classifier fusion occurs. The fusion methods described provide consistently robust solutions to the security classification problem.

Pseudo-ROC Curves

A study of pseudo-ROC curves and rank distributions will provide support and insight to the underlying behavior of classifier fusion for the security classification problem. This discussion is critical in understanding why certain classifier fusion metrics work best when fusing multiple models in the security classification domain.

As previously discussed, ROC curves are based entirely upon ranks. Furthermore, the Mann-Whitney *U* statistic, which is equivalent to the area under the ROC curve, is also equivalent to the probability that any random positive instance is ranked higher than a negative instance. When referring to the rank of an observation, a higher rank is a smaller value, meaning that a rank of 1 is considered higher than a rank of 10, for example. Given this property, it is entirely possible to create pseudo-ROC curves.

ROC curves represent the performance of a binary classifier, based entirely upon how the binary classifier ranks the observations and the true class of these observations. However, if an assumption is made that a classifier has a certain discriminating ability reflected in the AUC or Mann-Whitney *U* statistic, artificial ranks can be created with pseudo-random numbers. The simplest way to accomplish this is assuming normal distributions for the sake of creating artificial ranks. Suppose *A* and *B* are two random variables such that $P(A > B) = U$, or equivalently $P(A - B > 0) = U$. If $W = A - B$, and it is assumed that W is a standard normal random variable, *A* and *B* can also be defined as normal random variables with a variance of .5. The following table provides examples of this relationship $w \sim N(\mu, \sigma^2)$ represents a normal distribution with a mean of μ and a variance of σ^2, with the same notation for the distributions of *a* and *b*.

Table 1. Examples of Distributions Creating Artificial Ranks

$P(A>B)=.9$	$w \sim N(1.28,1)$	$a \sim N(1.28,.5)$	$b \sim N(0,.5)$
$P(A>B)=.8$	$w \sim N(1.28,1)$	$a \sim N(1.28,.5)$	$b \sim N(0,.5)$
$P(A>B)=.7$	$w \sim N(1.28,1)$	$a \sim N(1.28,.5)$	$b \sim N(0,.5)$

Given the distributions shown in table 1, it is possible to create random numbers which will behave with the desired probability of $P(A>B)=U$. This will also enforce that $P(R(a))>P(R(b)))=U$.

These rankings enable the creation of pseudo-ROC curves with an area under the curve equivalent to *U*. The essence of this method is that it allows for the study of ROC curves where control variables consist of the AUC, the number of positive examples, and the number of negative examples.

Pseudo-ROC curves serve a multitude of purposes. ROC curves are a very popular method to assess the performance of a binary classifier. Research involving ROC curves has largely been limited to the analysis of curves created by the output of models with real data. The study of ROC curves solely created from the output of classification models limits our ability to fully understand and explore the complete behavior of ROC curves and ranks. The study of pseudo-ROC curves places a number of parameters into the hands of the researcher - the discriminating power (reflected in the *U* statistic), proportion of the classes, and the total numbers of observations are all parameters controlled by the researcher with pseudo-ROC curves.

ROC theoretical research focuses extensively on the topic of the nonparametric statistics which impact ROC curves. This is primarily the Wilcoxon Rank Sum statistic and Mann-Whitney *U* statistic (Bradley, 1997; Fawcett, 2003; Hanley and McNeil, 1982). Exceptions to this include (Egan, 1975; Fawcett and Provost, 2001; Fawcett, 2001) where the authors have taken creative looks at ROC curves to include the application of game theory. However, the concept of the pseudo-ROC curve and use of this method to improve our understanding of ROC curves is a novel approach.

Rank Distributions

Given a *U* statistic and desired number of positive and negative instances, it is possible to create rank distributions. Let us consider *p* positive instances, *b* negative instances (choosing the letter *b* to signify a benign or negative observation), and *N* = *p* + *b* total observations. The rank distribution will be a discrete probability distribution, or probability mass function, with 1...*N* possible states, or ranks. Rank distributions reflect the likelihood that a particular rank is a positive or negative observation. For every case there is a given *U*, *p*, and *b*. This information is all that is necessary to create two rank distributions, one for positive observations and one for negative observations.

Simulation will be utilized to study these distributions. As stated in (Bertsekas and Tsitsiklis, 2002) estimating probabilities by simulation due to pragmatic necessity (because the analytical solution is very difficult) is an acceptable approach. Given a simulation that models behavior based upon true probabilities, the simulation estimates these probabilities with high accuracy. The combinatoric complexity and implications of order statistics involved with these rank distributions become problematic in creating an analytical solution for the mass functions of the rank distributions. This combinatoric complexity can be explored with a binomial distribution for a small number of instances, however that exploration is beyond the scope of this presentation.

Behavior of Fused Classifiers

Analyzing how these rank distributions behave provides insight for model fusion. Model fusion involves considering several models, all of which measure the same observations, and for each observation fuse the results of each model to arrive at a final decision value for each observation.

The fusion metric utilized in the most popular ensemble techniques such as random forest, bagging, and the random subspace method, is the average (avg) (Breiman, 1996; Breiman, 2001; Ho, 1998). The avg is a powerful aggregator, especially if all of the models possess roughly the same predictive power.

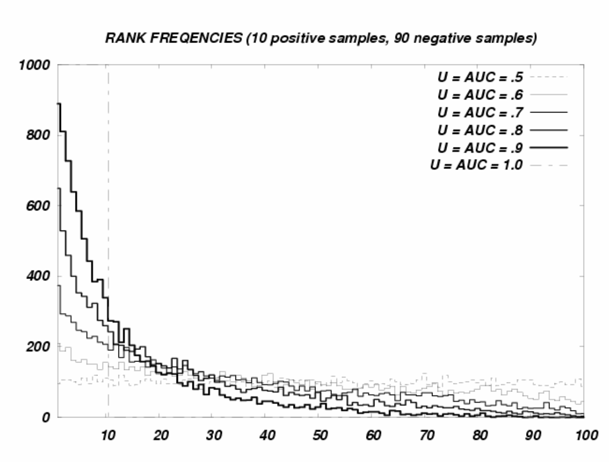


Figure 5. The histograms shown above illustrate how the rank frequencies for a minority class transition as *U* spans between 0.5 and 1.

Good prediction occurs for a model when the decision value distribution of the positive class achieves separation from the decision value distribution of the negative class. Fusion with the average function invokes the properties of the central limit theorem, and improved separation occurs as a result of variance reduction. This can be further explained in a brief example. Assume that three models each create a distribution for the decision values of the negative class with a mean of -1 and a variance of 1. Assume the distributions of the positive class have a mean of 1 and a variance of 1. The fused model, using the average aggregator, will create a distribution for the decision values of the negative class with a mean of -1 and a variance of 1/3. The positive class will have a mean of 1 and variance of 1/3. Tighter distributions for both the positive and negative classes create improved prediction.

Why the Average and `min` Fusion Metrics Work

The disadvantage of the average aggregator involves the equal weighting and inclusion of all models, good and bad. When fusing security classification problem models, it is likely that some of the models are poor classifiers. Therefore, it is desirable to utilize fusion that is robust against poor classifiers without knowing which classifiers are poor. This is precisely what the `min` aggregator accomplishes. Given an unbalanced classification problem, the rank distributions which result clearly favor the highest rankings and quickly tail off (see figure 5). The behavior is remarkably similar to the exponential distribution. An interesting property of the exponential distribution involves the distribution of the `min` order statistic. Given an exponential random variable x distributed with a mean (and standard deviation) of θ, the distribution of the `min` of x, x_1, is exponential with a mean (and standard deviation) of θ/n. This is a well known property of the exponential distribution.

This property indicates that the distribution of x_1 contains less dispersion, concentrating in a tighter range. This concentration enables separation, however more importantly the `min` fusion metric creates robustness against poor classifiers. In unbalanced classification, we understand from our study of rank distributions that given a good model the probability of encountering a large rank value for a positive instance is small. It is more likely to observe a small rank value. The `min` fusion metric indiscriminately eliminates large rank values. This indiscriminant elimination works based on the fact that there are a small number of positive instances. Machine learning researchers will immediately question the contribution of this fusion metric since it is difficult to identify how this fusion metric improves classification. Haykin indicates in (Haykin, 1999) that one of the fundamentals in every classification or pattern recognition problem involves ensuring the inclusion of all available information. The `min` aggregator works based upon our assumption that there are a small number of positive instances. This information contributes and improves performance.

Table 2. A Toy Rank Fusion Problem

True Class	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_3$	$\mathbf{R}_4$	$\mathbf{R}_5$	min	avg	$R(\text{min}) = o_{i2}$	$R(\text{avg}) = o_{i2}$	$\frac{\text{min}+\text{avg}}{2} = \frac{o_{i2}+o_{i2}}{2}$	$R(\frac{\text{min}+\text{avg}}{2})$
0	8	15	5	6	10	5.005	8.807	16	8	12.005	12
0	16	16	19	7	17	7.006	15.003	18	20	19.000	19
0	17	18	13	15	9	9.008	14.400	19	17	18.004	18
0	15	20	8	19	1	1.002	12.601	2	15	8.504	8
0	11	13	7	1	2	1.008	6.804	4	3	3.509	3
0	14	9	4	9	18	4.007	10.808	12	12	12.008	13
0	6	6	9	18	3	3.004	8.402	10	5	7.503	6
0	13	7	20	10	5	5.003	11.007	15	13	14.008	16
0	18	17	11	3	13	3.002	12.410	8	14	11.006	11
0	7	8	16	4	7	4.003	8.402	11	4	7.508	7
0	19	19	6	8	19	6.001	14.202	17	16	16.510	17
0	12	11	14	20	15	11.003	14.404	20	18	19.008	20
0	4	10	15	5	11	4.009	9.000	13	9	11.005	10
0	5	12	17	11	6	5.003	10.209	14	11	12.509	14
0	20	3	18	17	16	3.001	14.801	7	19	13.007	15
1	3	2	10	16	12	2.001	8.605	5	6	5.508	5
1	9	4	3	14	20	3.003	10.007	9	10	9.509	9
1	1	5	1	12	14	1.003	6.609	3	2	2.508	1
1	10	1	12	13	8	1.001	8.804	1	7	4.005	4
1	2	14	2	2	4	2.005	4.801	6	1	3.505	2
AUC	0.87	0.85	0.83	0.44	0.43	-	-	0.88	0.853	-	0.920

It is useful to illustrate rank fusion through a toy problem. Table 2 shows three models, two of which perform adequately and one which appears to have no predictive power. The resulting ranks created by the `min`, `avg`, and (`min` + `avg`)/2 functions are also shown. Realize that the aggregation columns represent $R(f(o_{i1},\ldots,o_{i5}))$, not $f(R(o_{i1}),\ldots, R(o_{i5}))$.

Particularly for the min function, ties must be solved which is done simply at random. Following any aggregation, decision values are mapped to ranks, or o_{ij}. If desired, an interested reader could recreate the last six columns to reinforce the concept.

The Properties of Synergistic Fusion - a Factorial Design Illustration

There is a fundamental synergistic fusion property that emerges from the presented empirical results and theoretical discussion. Stated simply, this property claims that when fusing ranks, there is improved discriminating power from the min aggregator if the problem is unbalanced with a minority positive class. The opposite is true for the max aggregator if the problem is unbalanced with a minority negative class. The paper supports this property with a discussion that includes a statistical explanation of the behavior of ranks created in a classification problem. Another way to explore this property involves creating a factorial design. Factorial design stems from a body of knowledge known as design of experiments, or DOE. R.A. Fisher was a pioneer of DOE, and researchers give much credit to Fisher for the current studies involving DOE (Box, 1980). For a comprehensive collection of Fisher's work to include his DOE work, see (Bennett, 1974).The DOE for this problem involved four factors. These factors were derived from simply considering what parameters effect this fusion problem. These parameters, or factors, include:

1. **balance**: the number of observations (out of 1000) which are members of the negative class. This factor becomes the most important factor in the experiment. The primary hypothesis of this study claims that the balance of the problem closely relates to the utility of the min aggregator. This experiment will reinforce this hypothesis.

2. **AUC of "good'" models**: the assumed area under the curve (AUC) or predictive power of effective models. A major assumption includes that all of the ``good'' models predict with a specific accuracy.

3. **fraction "good"**: the percent of models predicting with the accuracy of the AUC, and all others have no predictive power (AUC = .5). If all of the models are "good", the average (avg) aggregator suffices. The min aggregator is more robust against these powerless classifiers. This is simply by favoring smaller ranks which statistically tend to be members of the positive class for good models classifying in an imbalanced environment (positive minority). The avg aggregator considers all of the models equally, and therefore becomes susceptible to meaningless ranks created by the "poor" models.

4. **number of models**: the number of models fused. This has an important but subtle effect on the outcome that will be discussed.

Table 3. Design of Experiment

	balance	AUC of good models	fraction good	number of models
min value	500	0.8	0.4	2
max value	980	0.95	1	8
step	160	0.05	0.2	2

There were 4 levels explored for each factor creating 4^4 = 256 design points (dp). Each design point consisted of 30 repetitions, each repetition utilizing 1000 observations for each experiment. Three different fusion strategies were employed at each design point: min, (min + avg)/2, and avg. Table 3 illustrates the values of the parameters considered in the DOE.

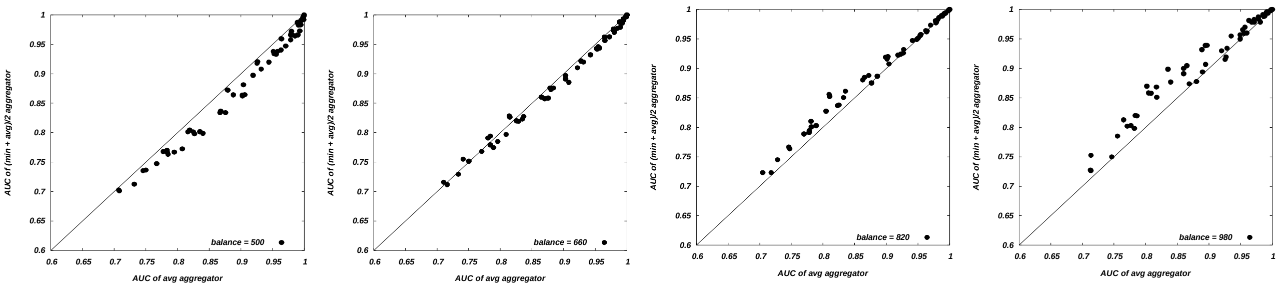


Figure 6. These plots compare the AUC achieved when fusing with the avg aggregator versus the (min + avg)/2 aggregator. Every plot contains the full spectrum of design points with the exception of the balance factor which is held constant for each plot. Notice that as imbalance increases, the (min + avg)/2 aggregator becomes more dominant.

The balance of the class ranges from 500 negative instances (completely balanced) to 980 negative instances (severely imbalanced) with a step of 160 between levels. The 'balance' parameter is not explored for a minority negative class. This is because the exact same property observed with the min aggregator for the minority positive class can be observed for the max aggregator if a minority negative class is considered. It is a symmetrical property that will not be shown for the sake of limiting redundancy. The AUC of the “good” models ranges from .8 to .95 with a step of .05 between levels. There was definitely a bias to explore the higher end of AUC values; further experimentation should consider exploring the lower range of AUC values. The fraction of “good” models replicates the unknown sensors or models in the ensemble which predict well, assuming that the others predict randomly (AUC = .5). This fraction ranges between .4 and 1 with a .2 step, exploring a slight minority of “good” models to observing a majority of “good” models. Brief experimentation indicated that having a fraction of “good” models less than .4 created poor and inconsistent performance across the board. The number of models ranged from 2 to 8 with a step of 2. Larger numbers of models severely favor the avg aggregator. Central limit theorem becomes stronger as the number of models increases therefore favoring the avg aggregator.

Table 3. Correlations between factors and the differences between aggregators.

	balance	AUC of good models	fraction good	number of models
(min+avg)/2 - avg	0.70	0.08	-0.17	-0.14
min - avg	0.58	0.16	0.06	-0.46

The min aggregator works based upon the simple premise that positive instances are more likely to rank as a low number when considering imbalanced (positive minority) problems. Given an imbalanced problem (minority positive class) with a small number of models to fuse, the **min** aggregator is less likely than the **avg** aggregator to be affected by a random model.

Table 3 illustrates the correlations observed between factors and the paired differences between aggregators. If there was no effect from the factor, the expected correlation is zero. The positive correlation that exists for the balance factor reinforces the premise that use of min aggregator improves classification for unbalanced classification. The strong negative correlation with the “number of models” factor and the difference between min and avg reinforces the concept that the avg aggregator becomes more dominant as the number of models increases. The avg aggregator served as the benchmark for this experiment because this is the aggregation technique accepted in state of the art ensemble methods. The min aggregator alone does not perform as well as a combination of the min aggregator with the avg aggregator. This is an interesting behavior that was discovered when exploring a spectrum of aggregators, T-norms and T-conorms, commonly known in the fuzzy logic literature. (Evangelista et. al., 2005) discusses the results of this initial experimentation.

Experimental Results

The focus of this paper is to explain some of the theoretical properties of rankings and discuss how understanding these properties can improve classifier fusion. During the exploratory work that led to the theoretical included theoretical discussion, experiments involving several classifier fusion strategies were performed. Experimental results using rank fusion with several large data sets has been published in (Evangelista, 2005).

Table 4. Experiments explored classifier fusion performance for these datasets. *m* represents dimensionality (number of variables), *N* represents the number of observations, *l* represents the bootstrapping leave out quantity. *p* represents the number of positive instances in the dataset.

Dataset	m	N	p	l	**comment**
P300	100	4200	2100	10	Courtesy of Wadsworth Lab, www.wadsworth.org
Ionosphere	34	351	126	5	UCI Repository
Schonlau	26	5000	231	5	www.schonlau.net
Sick	137	5393	250	10	see (Sick,2004)
Sick	137	5393	250	50	see (Sick,2004)

For each of these datasets, baseline classifier performance using the one-class support vector machine (with a linear kernel) was recorded. The experiment compared this baseline performance against several model fusion policies. Multiple models were created for each dataset using a bootstrapping of variables (Evangelista, 2005).

The experiment iterated 30 times for each dataset. Each iteration involved a re-shuffling of the training / test data and a different bootstrap of variables for model fusion. Each iteration was compared against the baseline performance achieved for the established training / test split, and results were compared with various fusion policies.

Table 5. Paired *t*-test *p*-value results for various model fusion policies compared against baseline performance.

Dataset	algebraic product	min	(min+avg)/2	avg
Ionosphere	NS	0.00	0.00	NS
Schonlau	0.00	0.04	0.00	0.01
Sick, $l = 50$	0.04	NS	0.04	0.63
Sick, $l = 10$	0.00	0.28	0.02	0.00
P300	0.00	0.00	0.00	0.00

Table 5 illustrates the results from a one way hypothesis test where a small *p*-value indicates the average of the AUC achieved from the fusion policy was significantly greater that the average of the AUC achieved by the baseline model. Table 5 shows that the min aggregator combined with the avg aggregator is the only fusion policy that outperforms the baseline model for every dataset. A result of “NS” indicates a *p*-value that was greater than .5.

Conclusion

This paper presents several novel thrusts which present opportunities for improved ensembles as well as future directions for research with ensemble techniques. Simulating rank distributions and pseudo ROC curves provide insight into the non-parametric statistics behind ROC curves. The insight provided by controlling parameters such as prediction power, balance of classes, and number of models enables analysis which is not possible when analyzing fusion metrics and ROC curves created from actual data. Analysis with actual data limits control of the parameters.

Since the rank distributions and ROC curves are created from first principles and model generic, there is no concern of bias due to the characteristics of the data or behavior of a particular model. Results are general, and the general results of the simulation analysis provide a broader range of applicability for the research included in this chapter.

The final and perhaps most important finding in this chapter involves consideration of the `min` and `max` aggregators when fusing models. When comparing decision values or non-binary classification systems, the `avg` is essentially the only fusion metric considered. However, it is well known that there is a flaw of averages, including bias from outliers. The rank distributions studied in this chapter clearly illustrate that there are different likelihoods associated with balanced classification problems as opposed to unbalanced classification problems. The `min` or `max` aggregator capitalize on this difference in likelihoods and create improved results.

Reviewing first principles is often tedious, however a well grounded understanding of these principles can be the difference between improving a technique or merely knowing how to apply it. Maintaining the edge in a rapidly evolving technological world requires improving techniques and not merely applying them.

REFERENCES

Bertsekas, D.P., & Tsitsiklis, J.N. (2002) *Introduction to Probability*, Belmont, MA: Athena Scientific.

Box, J.F. (1980) R.A. Fisher and the Design of Experiments, 1922 - 1926, *The American Statistician, 34*(1), 1-7.

Bradley, A.P. (1997) The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms, *Pattern Recognition, 30*(7), 1145-1159.

Breiman, L. (1996) Bagging Predictors, *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001) Random Forests, *Machine Learning, 45*(1), 5-32.

Egan, J. (1975) *Signal Detection Theory and ROC Analysis*, New York: Academic Press, Inc.

Evangelista, P.F., Embrechts, M.J., & Szymanski, B.K. (2005), Data Fusion for Outlier Detection through Pseudo ROC Curves and Rank Distributions, *Proceedings of the International Joint Conference on Neural Networks*, Vancouver, Canada.

Evangelista, P.F. (2006) *The Unbalanced Classification Problem: Detecting Breaches in Security*, Doctoral dissertation, Rensselaer Polytechnic Institute.

Fan, J., Upadhye, S., & Worster, A. (2005). Understanding Receiver Operating Characteristic (ROC) Curves, *Canadian Journal of Emergency Medicine, 8*(1), 19-20.

Fawcett, T. (2001) Using Rule Sets to Maximize ROC Performance, *IEEE International Conference on Data Mining*, San Jose, CA.

Fawcett, T. (2003) ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, *Technical Report HPL-2003-4*, Hewlett Packard, Palo Alto, CA.

Fawcett, T., & Provost, F. (2001) Robust Classification for Imprecise Environments, *Machine Learning Journal, 42*(3), 203-231.

Bennett, J.H. (ed.) (1974) Collected Papers of R.A. Fisher (5 vols.), Adelaide: University of Adelaide.

Hanley, J.A., & McNeil, B. (1982) The Meaning and Use of the Area Under the Receiver Operating Characteristic Curve, *Radiology, 143*(1), 29-36.

Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, Second Edition, New Jersey: Prentice Hall.

Ho, T.K. (1998) The Random Subspace Method for Constructing Decision Forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844.

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning, 30*, 271-274.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco, CA: Holden-Day, Inc.

Loomis, F.W. (1951). Problems of Air Defense: Final Report of Project Charles. *MIT Report.*

Mann, H.B., & Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *The Annals of Mathematical Statistics 18*(1), 50-60.

Mason, S.J., & Graham, N.E. (2002) Areas Beneath the Relative Operating Characteristics (ROC) and relative operating levels (ROC) curves: Statistical Significance and Interpretation, *Quarterly Journal of the Royal Meteorological Society*, 128, 2145-2166.

Naka, F.R., & Ward, W.W. (2000). Distant Early Warning Line Radars: The Quest for Automatic Signal Detection, *Lincoln Laboratory Journal, 12*(2), 181-204.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods, *Biometrics Bulletin 6*(1), 80-83.