Proc. International Joint Conference on Neural Networks 2005, Montreal, Canada, August 2005

Fuzzy ROC Curves for Unsupervised Nonparametric Ensemble Techniques

Paul F. Evangelista and Mark J. Embrechts
Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute
Troy, NY 12180
E-mail: {evangp,embrem}@rpi.edu

Piero Bonissone
GE Global Research
One Research Circle
Niskayuna, NY 12309
E-mail: bonissone@research.ge.com

Boleslaw K. Szymanski
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
E-mail: szymansk@rpi.edu

***Abstract*— This paper explores a novel ensemble technique for unsupervised classification using nonparametric statistics. Multiple classification systems (MCS), or ensemble techniques, involve considering several classification methods or multiple outputs from the same method and devising techniques to reach a decision. The performance of a binary classification system can be measured on a receiver operating characteristic (ROC) curve, and the area under the curve (AUC) is exactly the Wilcoxon Rank Sum or Mann-Whitney U statistic, both of which are nonparametric statistics based upon ranked data. Successful performance of an unsupervised ensemble can be measured through the AUC, and the performance of different aggregation techniques for the combination of the multiple classification system decision values, or rankings in this paper, is illustrated. Aggregation techniques are based upon fuzzy logic theory, creating the fuzzy ROC curve. The one-class SVM is utilized for the unsupervised classification.**

I. Introduction

The purpose of this paper is to illustrate effective ensemble methods for unsupervised classification in machine learning. The measure of effectiveness of the combination is displayed with fuzzy ROC curves. The problem motivating this research is the unbalanced, unsupervised, binary classification problem. We will specifically address results with a computer intrusion dataset. Receiver Operating Characteristic (ROC) curves illustrate the performance of binary classifiers [10]. The underlying nonparametric statistics of a binary classification system, specifically the formulation of the soft decision values into ranks, relates directly to the area under the ROC curve (AUC). Since our goal involves improving the AUC, we focus on ranks. Unsupervised learning methods often suffer from a curse of dimensionality [1], meaning that as dimensionality grows large, volume increases exponentially and distance measures become meaningless. Intelligent subspace modeling can overcome this curse of dimensionality.

II. Recent Work

Combinations of multiple classifiers, or ensemble techniques, is a very active field of research today. However, the field remains relatively loosely structured as researchers continue to build the theory supporting the principles of classifier combinations [16]. Significant work in this field has been contributed by Kuncheva in [14], [15], [16], [17]. Bonissone et. al. investigated the effect of different fuzzy logic triangular norms based upon the correlation of decision values from multiple classifiers [2]. The majority of work in this field has been devoted to supervised learning, with less effort addressing unsupervised problems [31]. The research that does address unsupervised ensembles involves clustering almost entirely. There is a vast amount of literature that discusses subspace clustering algorithms [25]. The recent work that appears similar in motivation to our technique include Yang et. al. who develop a subspace clustering model based upon Pearson's R correlation [33], and Ma and Perkins who utilize the one-class SVM for time series prediction and combine results from intermediate phase spaces[20]. The unsupervised learning discussed in this paper involves what is commonly referred to as novelty or anomaly detection, where class labels for training data are all negative or healthy cases.

The technique we propose illustrates that unsupervised learning in subspaces of high dimensional data will typically outperform unsupervised learning in the high dimensional data space as a whole. Furthermore, the novel innovations of this paper include the following:

1. Intelligent subspace modeling, through the use of nonparametric statistics which seek orthogonal subspaces, will provide further improvement of detection beyond a random selection of subspaces.

2. Fuzzy logic aggregation techniques create the fuzzy ROC curve, illustrating improved AUC by selecting proper aggregation techniques.

III. Seeking Diverse Subspaces

It is widely known that diversity is desirable for classifier fusion [15], [16], and there are numerous methods for measuring classifier diversity for supervised classification. It is undesirable to have multiple classifiers that all make the same errors in the same direction; it is desirable to have classifiers making different mistakes on different instances, and when combined, synergistic performance occurs through intelligent combination. A simple regression model illustrates this idea. This model can be expressed as $\mathbf{Xw}=\mathbf{y}$, where $\mathbf{X} \in \mathbb{R}^{N \times m}$, $\mathbf{w} \in \mathbb{R}^{m \times 1}$, $\mathbf{y} \in \mathbb{R}^{N \times 1}$. If $\mathbf{y}$ is known for the training data, the measures of diversity shown in [15], [16] apply. However,

when **y** is either unknown or only contains the negative class, measures of diversity must involve **X**. Therefore, since our desire is to create subspaces of a high dimensional dataset, we seek diverse subspaces.

In order to measure this diversity, a distance measurement, d_{ij}, will be calculated for the i^{th} observation in every j^{th} subspace. The distance measurement used in this paper is the Euclidean distance of the observation point to the subspace centroid. However, other distance measurements should not be discounted, and nonlinear kernel distance measurements can also be considered. This distance measurement will provide the basis of our intelligent subspace modeling. Our interest is to find subspaces that are not correlated with respect to d_{ij}. If subspaces are correlated with respect to d_{ij}, these subspaces capture similar behavior. Uncorrelated subspaces indicate subspaces that are somewhat orthogonal, and we interpret this as diversity.

Figure 1 illustrates the idea of correlated versus uncorrelated subspaces based upon Kendall's W, a nonparametric measure of concordance which calculates agreement across multiple subspaces using only ranks. The vertical axis in figure 1 measures d_{ij}, and the ranking of the points is obvious from plots. Similar ordering occurs in the subspaces on the left with a higher W, however the subspaces on the right contain a much more random order, and W reflects a lower value.

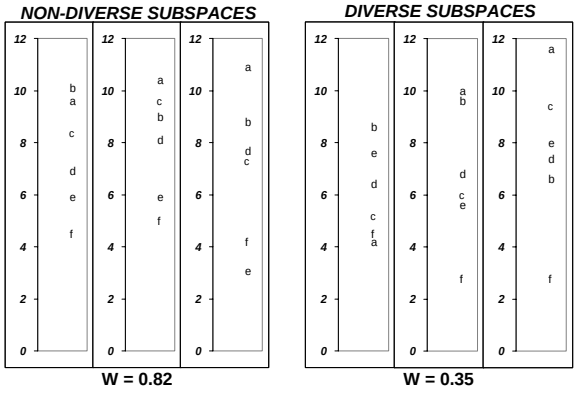

Fig. 1. A comparison of correlated and uncorrelated subspaces

Kendall's W provides a scalar measurement of agreement or disagreement of the rankings between subspaces. In order to compute W, first map $d_{ij} \rightarrow R_{ij}$ such that R_{ij} represents the rank of the i^{th} point distance-wise with respect to the j^{th} centroid. Assuming that there are N points in our training data and l subspaces, Kendall defines W as follows in [13]:

$$W = \frac{12S}{l^2(N^3 - N)} \tag{1}$$

$$\text{where } S = \sum_{i=1}^{N} \left(\left(\sum_{j=1}^{l} R_{ij} \right) - \frac{l(N+1)}{2} \right)^2$$

It is our hypothesis that diverse subspaces will create a small W, and this diversity will provide improved ensembles for unsupervised learning.

A. Finding Near-Optimal Subspaces with Evolutionary Computing

Since $(0 \leq W \leq 1)$, where perfect agreement equates to $W = 1$, our interest is to minimize W. To solve this minimization problem, our choice is a simple genetic algorithm.

Given standardized data matrix **X** (centered at 0 and divide by standard deviation), containing m variables that measure N observations, randomly create l mutually exclusive subspaces from the m variables. Assume there are k variables in every subspace if m is divisible by l. Our experience with the one-class SVM indicates that for $k > 7$, increased dimensionality begins to degrade performance, however this is simply a heuristic and a parameter that the modeler must choose based upon domain knowledge and desired complexity. These subspaces can be modeled as a chromosome where each element of the chromosome corresponds to a variable within a specific subspace. This is shown in figure 2 for an example where $m = 26$ and $l = 3$.

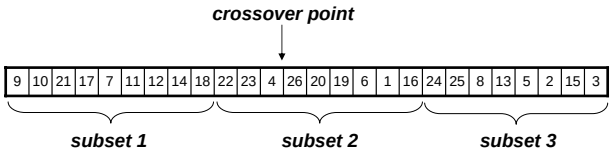

Fig. 2. Illustration of chromosome and subspaces

For the results included in this paper, the population size was 50 and the number of generations was 30. The selection of chromosomes followed the common roulette wheel process, where chromosomes with a better fitness receive a higher probability of being selected [24]. In order to retain the fittest chromosome found, an elitist rule retained the single best chromosome and passed it to the next generation. Immigrants, which were two new random chromosomes who took the place of the least fit chromosomes selected for breeding, migrated into every generation to help maintain diversity. After selecting a group of chromosomes for breeding, these chromosomes had a 40% chance of directly passing to the next generation and a 60% chance of crossover. Crossover occurred at the point shown in Figure 2.

The last step involved mutation. During both crossover and mutation, care was taken to ensure that only feasible chromosomes resulted. Each element had a 1% chance of encountering mutation. If an element was selected for mutation, the element was simply swapped with its mirror element. If the chromosome contains m elements, and the k^{th} element is selected for mutation, then the k^{th} element swaps with the $(m - k)^{th}$ element, which we refer to as its mirror. After intelligent selection of subspaces, the one-class SVM is utilized.

B. The one-class SVM

The one-class SVM is an outlier detection technique originally proposed in [26]. Stolfo and Wang [30] successfully apply the one-class SVM to the intrusion data set that we use in this paper. Chen uses the one-class SVM for image

retrieval[4]. Shawe-Taylor and Cristianini provide the theoretical background for this method in [29]. The simplest way to express the one-class SVM is to envision a hypersphere, and the object is to squeeze all of the training data into the tightest hypersphere feasible. Once the hypersphere is estimated, data that does not fit the hypersphere will be considered an outlier or not a member of the negative class. Consider the following formulation of the one-class SVM originally from [26]:

If we consider $X_1, X_2, ..., X_N \in \chi$ instances of training observations, and Φ is a mapping into the feature space, F, from χ.

$$\min_{R\in\mathbb{R},\zeta\in\mathbb{R}^N,c\in F} \quad R^2 + \frac{1}{vN}\sum_i \zeta_i \tag{2}$$

$$\text{subject to} \quad \| \Phi(X_i) - c \|^2 \leq R^2 + \zeta_i, \quad \zeta_i \geq 0 \text{ for } i \in [N]$$

This minimization function attempts to squeeze R, which can be thought of as the radius, as small as possible in order to fit all of the training samples. If a training sample will not fit, ζ_i is a slack variable so that a few points will not drive the hypersphere to an unnecessary size. A free parameter, v, enables the modeler to adjust the impact of the slack variables. The one-class SVM creates a corresponding decision function that takes on values generally ranging from -1 to +1, where values close to +1 indicate datapoints that fit into the ball and values of -1 indicate datapoints lying outside of the ball. The decision values from a one-class SVM indicate a degree of confidence either towards inclusion or exclusion. Converting these values into ranks invokes nonparametric statistical theory and also provides a method of comparison across subspaces. Empirical results indicate that fusion of rank statistics creates superior results. Since the goal of our research is to improve the area under the ROC curve through the fusion of nonparametric rank statistics, it is appropriate to show how these rank statistics relate to the ROC curve.

IV. The Relevance of Ranks and ROC Curves

Receiver operating characteristic (ROC) curves provide an elegant, simple representation of the performance of a binary classification system. The curve presents the relationship between a false positive rate and a true positive rate across the full spectrum of operating points, which can also be considered the full spectrum of a continuous threshold value for a decision method. The nonparametric interpretation of the ROC curve and the AUC is discussed in [3], [11], [21]. Many machine learning algorithms provide a real number, or soft decision value, as the output. It is difficult, and often not necessary, to associate this output to a probability distribution. It is more meaningful to consider this output as a degree of confidence towards one class or the other. Interpreting this soft decision value as a ranking leads directly to the nonparametric statistics associated with the area under the ROC curve. The Wilcoxon Rank Sum statistic, which is directly proportional to the Mann-Whitney U statistic, provides this association.

The Wilcoxon Rank Sum statistic is based upon the rankings of observations which indicate their degree of membership in one of two different possible classes. Consider an experiment that involves computer intrusion detection. A host based intrusion detection system (IDS) monitors N workstations, and let us assume n workstations experience an actual attack or intrusion. $p = N - n$ of the workstations are not attacked. The IDS ranks the workstations from 1 to N, assigning a 1 to the workstation which seems most likely under attack and N to the workstation that appears least likely under attack. Let $S_i, i = 1..n$ represent the rankings of actual attacks, and let $S_j, j = 1..p$ represent the rankings of the workstations which were not actually attacked. The Wilcoxon Rank Sum Statistic, W_s, is equivalent to $\sum_{i=1}^{n} S_i$. Since the sum of all rankings is $\frac{1}{2}N(N+1)$, it follows that $W_r = \frac{1}{2}N(N+1) - W_s$. W_r can also be calculated as shown in equation 3 [19].

$$W_r = \sum_{j=1}^{p} S_j = \frac{1}{2}p(p+1) + \sum_{i=1}^{n}\sum_{j=1}^{p} \varphi(S_i, S_j) \tag{3}$$

$$\text{where } \varphi(S_i, S_j) = \begin{cases} 1 & \text{if } S_i < S_j \\ 0 & \text{otherwise} \end{cases}$$

The statistics $W_{XY} = W_s - \frac{1}{2}n(n+1)$ and $W_{YX} = W_r - \frac{1}{2}p(p+1)$ 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} and shown in equation 4.

$$U = \frac{W_{YX}}{pn} = \frac{1}{pn}\sum_{i=1}^{n}\sum_{j=1}^{p} \varphi(S_i, S_j) = \text{AUC} \tag{4}$$

Viewing this as a discrete probability problem, one can also claim that the AUC is equivalent to $P\{S_i < S_j\}$. It is helpful to visualize all of this with a simple numerical example. Given the same context of the example mentioned above, let us assume $N = 15, n = 7, p = 8$. The IDS ranks the workstations as shown in figure 3, and the resulting ROC curve is also shown.

The AUC is largely considered an excellent scalar performance measure for binary classification systems [3], [10]. We propose that ensemble techniques for binary classifiers should measure improved performance through the analysis of ROC curves. The purpose of ensemble techniques is to improve the accuracy of classification, and only through the use of ROC curves will it be apparent whether or not this improvement exists across the full spectrum of thresholds.

V. Fusion of Decision Values

1) Mapping into Comparable Decision Spaces: For each observation within each subspace selected, the classifier will produce a decision value, D_{ij}, where D_{ij} represents the decision value from the j^{th} classifier for the i^{th} observation. o_{ij} represents the ordinal position, or rank, of D_{ij} (for the

Rank	True Class
1	1
2	1
3	1
4	1
5	0
6	1
7	1
8	0
9	0
10	1
11	0
12	0
13	0
14	0
15	0

W_s	33
W_r	87
W_{XY}	5
W_{YX}	51
AUC	.9107

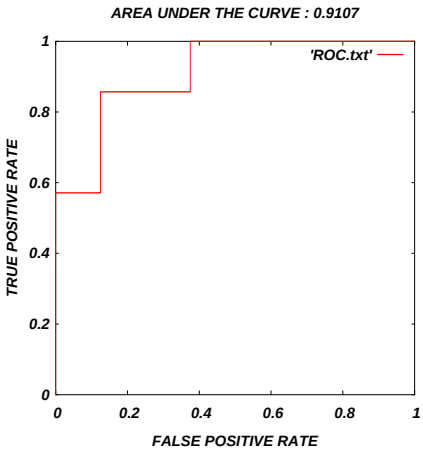

Fig. 3. Example of calculating AUC with ranks and resulting ROC curve

same classifier, meaning j remains constant). For example, if D_{71} is the smallest value for the 1^{st} classifier, $o_{71} = 1$. In order to incorporate fuzzy logic, o_{ij} must be mapped into a new space of real numbers, let us call Λ, where $\Lambda \in (0, 1)$. This mapping will be $o_{ij} \rightarrow \theta_{ij}$ such that $\theta_{ij} \in \Lambda$. For $o_{ij} \rightarrow \theta_{ij}$ this is a scaling procedure where all o_{ij} are divided by the number of observations, N, such that $\theta_{ij} = o_{ij}/N$.

2) Fuzzy Logic and Decisions with Contention: Utilizing fuzzy logic theory, T-conorms and T-norms can be considered for data fusion. Jang et. al. provide an explanation of T-conorms and T-norms in [12]. The choice between T-norms and T-conorms can often depend upon the type of decision. The medical community is cautious of false negative tests, meaning that they would rather have error on the side of falsely telling someone that they have cancer as opposed to letting it go undetected. The intrusion detection community is concerned about minimizing false positives, because too many false positives render an intrusion detection system useless as analysts slog through countless false alarms. In the realm of one-class SVMs, the original decision values will take on values ranging generally from -1 to +1, where values closer to +1 indicate observations that fit inside the ball or estimated distribution (indicating non-intruders), and values closer to -1 indicate outliers (potential intruders). Experimental results indicate that T-conorms create the best overall results amongst aggregation techniques, indicating that it is possible one subspace identifies a novelty that other subspaces do not identify. Figure 4 illustrates the domain of aggregation operators.

One problem with T-norms and T-conorms is that contention within aggregation is not captured. By contention we are referring to a vast difference of decision values between classifiers. However, contention can be captured and considered appropriately. Typically, if contention exists, a system needs to reflect caution. In other words, if we are minimizing false positives and contention exists in a decision, we may simply

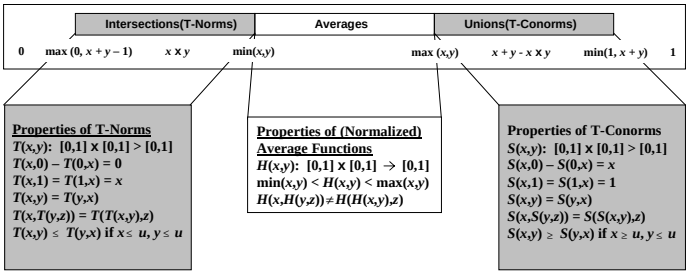

Fig. 4. Aggregation operators

choose negative or choose a different aggregator for contentious decisions. If contention exists in a medical decision, it is likely that the initial diagnosis will report positive (cancer detected) and then further tests will be pursued. There are numerous ways to measure contention, and one of the simplest is to consider the difference between the max and min decision values. If this difference exceeds a threshold, contention exists and it may be best to choose a different aggregator or make a cautious decision.

VI. Experimental Results

Two datasets have been explored for the experimental results. The first dataset, which we will refer to as the Schonlau et. al. or SEA dataset, is a computer intrusion dataset originally created by Schonlau and discussed in [6], [7], [8], [27], [28]. The original data captures UNIX commands from 50 different users. Masqueraders are simulated within these commands by probabilistically inserting other users (not from amongst original 50) commands into a stream of authentic commands. Only 2.56% of the command streams contain masquerades, creating an unbalanced dataset. The goal is to detect these masquerades. Much work has been done with this data, and the numerical representation of the SEA dataset is largely up to the modeler. The data used in this paper consists of the combination of text mining variables (described in [9]) and recursive data mining variables (described in [32]) derived from the SEA dataset. In total there are 26 variables.

Previous work with this data includes Schonlau's uniqueness algorithm, explained in [28], achieving a 40% true positive rating before crossing the 1% false positive boundary. Wang [30] used one-class training based on data representative of only one user and demonstrated that it worked as well as multi-class training. Coull [5] applied bioinformatics matching algorithm for a semi-global alignment to this problem. Lee [18] built a data mining framework for constructing features and model for intrusion detection. Szymanski and Zhang applied a recursive data mining algorithm for frequent patterns to detect intruders [32]. Evangelista et. al. [9] applied supervised learning through kernel partial least squares to the SEA dataset.

Roy Maxion contributed insightful work with this data that challenged both the design of the data set and previous techniques used on this data [23], [22]. Maxion uses a "1v49" approach in [23], where he trains a Naive Bayes Classifier one user at a time using the training data from one user as true

negative examples versus data from the forty-nine other users (hence 1v49) as true positive (masquerader) examples. Maxion claimed the best performance to date in [23], achieving a true positive rating of 60% while maintaining a false positive rating of 1% or less. Maxion also examines masquerade detection with a similar data set that contains command arguments in [22].

The results shown in table I and figure 5 illustrate the improvements obtained through our nonparametric ensemble technique for unsupervised learning. The plot of the ROC curves shows the results from using 26 original variables that represented the SEA data as one group of variables with the one-class SVM and the result of creating $l = 3$ subspaces of features and fusing the results to create the fuzzy ROC curve. It is interesting to notice in the table of results that nearly every aggregation technique demonstrated improvement, with the most significant improvement in the T-norms.

TABLE I

RESULTS OF SEA DATA WITH DIVERSE AND NON-DIVERSE SUBSETS

	DIVERSE	NON-DIVERSE
T-norms		
minimum	.90	.84
algebraic product	.91	.85
minimum with contention	.86	.81
algebraic product with contention	.93	.90
T-conorms		
maximum	.84	.80
algebraic sum	.89	.85
maximum with contention	.82	.77
algebraic sum with contention	.86	.81

for contention, $t = .5$

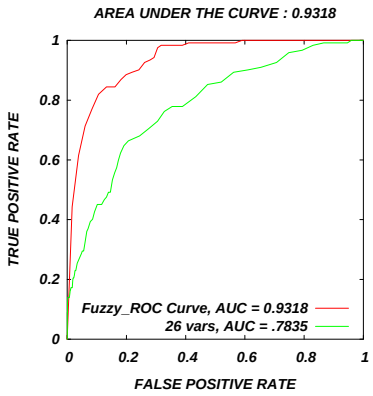

Fig. 5. ROC for SEA data using algebraic product with contention

The ionosphere data is included to illustrate the performance of our ensemble technique with a balanced benchmark data set. This dataset is available from the UCI repository, and it consists of 34 variables that represent different radar signals received while investigating the ionosphere for either good or bad structure. For this experiment we again chose $l = 3$.

It is very logical to ask why a simple dimension reduction technique, such as principal components, is not sufficient to

TABLE II

OVERALL RESULTS OF IONOSPHERE DATA WITH DIVERSE AND NON-DIVERSE SUBSETS

	DIVERSE	NON-DIVERSE
T-norms		
minimum	.96	.953
algebraic product	.61	.64
minimum with contention	.86	.92
algebraic product with contention	.85	.91
T-conorms		
maximum	.69	.70
algebraic sum	.69	.70
maximum with contention	.82	.88
algebraic sum with contention	.85	.91

for contention, $t = .5$

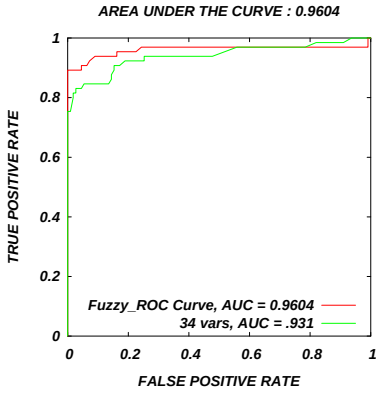

Fig. 6. ROC plot for ionosphere data with minimize aggregation technique

overcome the curse of dimensionality. Principal components capture variance, and by utilizing principal components the modeler is assuming that the variables from the training and testing data follow similar distributions, and furthermore that it is possible to identify novelties as an outlier along one of these axis of maximal variance, or principal component. If this novelty occurs as an outlier from only one variable, and this variable contributes minimal variance to the dataset in the training sample, it is likely that this novelty will go undetected. Furthermore, it is important that subspaces consist of meaningful dimensions for causality analysis[25].

For both the SEA data and the ionosphere data, principal components analysis was used as a dimension reduction technique to compare performance versus the ensemble method. For the SEA data, the PCA technique did improve classification reaching an AUC of .91, however the diverse subspaces with T-norm aggregation performed comparable to this and in some cases better. For the ionosphere data, however, the PCA technique actually degraded performance, achieving an AUC of .89 and falling well short of the best ensemble shown above.

T-norms provide the best aggregation for both datasets. This illustrates the idea of diversity, where different classifiers apparently correctly identify different positive cases. Every aggregator does not create improved performance with the

ionosphere data; actually, it is only the minimization operator that improves performance. Furthermore, the impact of diversity does not seem as significant with the ionosphere data as it is with the Schonlau data. It is possible that a different distance measure could be appropriate for the ionosphere data. Regardless of this, it is evident that improved performance is possible with intelligent ensemble methods which is a significant objective of this research.

VII. Conclusions

This paper illustrates the ability to improve unsupervised learning through intelligent subspace modeling and proper output fusion. Nonparametric representation of the output, or ranks, provides a comparable measure for data fusion. It is difficult to accurately represent the soft label output of a classification technique in terms of a parameterized density function, especially since this output should be bimodal as it represents both the positive and negative class. It is more consistent with the true meaning of the soft label output to arrange the output as ranks, especially if fusion of the label outputs is necessary.

It was empirically shown that Kendall's W statistic provides an overall measure of diversity across subspaces, and furthermore that this diversity will lead to improved classification results. Through careful experimentation and analysis of two different datasets, Kendall's W appears promising as a measure of diversity but does require further investigation with other distance measures. Further research in this area could include investigation of other distance measures, some of which could be kernel based.

References

[1] Charu C. Aggarwal and Philip S. Yu. Outlier Detection for High Dimensional Data. Santa Barbara, California, 2001. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data.

[2] Piero Bonissone, Kai Goebel, and Weizhong Yan. Classifier Fusion using Triangular Norms. Cagliari, Italy, June 2004. Proceedings of Multiple Classifier Systems (MCS) 2004.

[3] Andrew P. Bradley. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, Volume 30(7):1145–1159, 1997.

[4] Yunqiang Chen, Xiang Zhou, and Thomas S. Huang. One-Class SVM for Learning in Image Retrieval. Thessaloniki, Greece, 2001. Proceedings of IEEE International Conference on Image Processing.

[5] Scott Coull, Joel Branch, and Boleslaw K. Szymanski. Intrusion Detection: A Bioinformatics Approach. Las Vegas, Nevada, December 2001. Proceedings of the 19th Annual Computer Security Applications Conference.

[6] William DuMouchel, Wen Hua Ju, Alan F. Karr, Matthius Schonlau, Martin Theus, and Yehuda Vardi. Computer Intrusion: Detecting Masquerades. *Statistical Science*, 16(1):1–17, 2001.

[7] William DuMouchel and Matthius Schonlau. A Fast Computer Intrusion Detection Algorithm Based on Hypothesis Testing of Command Transition Probabilities. pages 189–193. The Fourth International Conference of Knowledge Discovery and Data Mining, August 1998.

[8] William DuMouchel and Matthius Schonlau. A Comparison of Test Statistics for Computer Intrusion Detection Based on Principal Components Regression of Transition Probabilities. pages 404–413. Proceedings of the 30th Symposium on the Interface: Computing Science and Statistics, 1999.

[9] Paul F. Evangelista, Mark J. Embrechts, and Boleslaw K. Szymanski. Computer Intrusion Detection Through Predictive Models. St. Louis, Missouri, November 2004. Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining and Complex Systems.

[10] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Palo Alto, CA, 2003. Technical Report HPL-2003-4, Hewlett Packard.

[11] James A. Hanley and Barbara J. McNeil. The Meaning and Use of the Area Under the Receiver Operating Characteristic Curve. *Radiology*, 143(1):29–36, 1982.

[12] Jyh-Shing Roger Jang, Chuen-Tsai Sun, and Eiji Mizutani. *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Prentice Hall, 1997.

[13] M.G. Kendall. *Rank Correlation Methods*. Charles Griffin, London, 1948.

[14] Ludmila I. Kuncheva. 'Fuzzy' vs. 'Non-fuzzy' in Combining Classifiers Designed by Boosting. *IEEE Transactions on Fuzzy Systems*, 11(3):729–741, 2003.

[15] Ludmila I. Kuncheva. That Elusive Diversity in Classifier Ensembles. Mallorca, Spain, 2003. Proceedings of 1st Iberian Conference on Pattern Recognition and Image Analysis.

[16] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, Inc., 2004.

[17] Ludmila I. Kuncheva and C.J. Whitaker. Measures of Diversity in Classifier Ensembles. *Machine Learning*, 51:181–207, 2003.

[18] Wenke Lee and Salvatore J. Stolfo. A Framework for Constructing Features and Models for Intrusion Detection Systems. *ACM Transactions on Information and System Security (TISSEC)*, 3(4):227–261, 2000.

[19] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Inc., San Francisco, CA, 1975.

[20] Junshui Ma and Simon Perkins. Time-series Novelty Detection Using One-class Support Vector Machines. Portland, Oregon, July 2003. International Joint Conference on Neural Networks.

[21] S.J. Mason and N.E. Graham. 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, 2002.

[22] Roy A. Maxion. Masquerade Detection Using Enriched Command Lines. San Francisco, CA, June 2003. International Conference on Dependable Systems and Networks.

[23] Roy A. Maxion and Tahlia N. Townsend. Masquerade Detection Using Truncated Command Lines. Washington, D.C., June 2002. International Conference on Dependable Systems and Networks.

[24] Zbigniew Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs (2nd, extended ed.)*. Springer-Verlag New York, Inc., 1994.

[25] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace Clustering for High Dimensional Data: A Review. *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2004.

[26] Bernhard Scholkopf, John C. Platt, John Shawe Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the Support of a High Dimensional Distribution. *Neural Computation*, 13:1443–1471, 2001.

[27] Matthius Schonlau and Martin Theus. Intrusion Detection Based on Structural Zeroes. *Statistical Computing and Graphics Newsletter*, 9(1):12–17, 1998.

[28] Matthius Schonlau and Martin Theus. Detecting Masquerades in Intrusion Detection Based on Unpopular Commands. *Information Processing Letters*, 76(1-2):33–38, 2000.

[29] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[30] Salvatore Stolfo and Ke Wang. One Class Training for Masquerade Detection. Florida, 19 November 2003. 3rd IEEE Conference Data Mining Workshop on Data Mining for Computer Security.

[31] Alexander Strehl and Joydeep Ghosh. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617, December 2002.

[32] Boleslaw K. Szymanski and Yongqiang Zhang. Recursive Data Mining for Masquerade Detection and Author Identification. West Point, NY, 9-11 June 2004. 3rd Annual IEEE Information Assurance Workshop.

[33] Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. δ-clusters: Capturing Subspace Correlation in a Large Data Set. pages 517–528. 18th International Conference on Data Engineering, 2004.