Applied Soft Computing Technologies: The Challenge of Complexity,
Abraham, A.; Baets, B.d.; Koppen, M.; Nickolay, B. (Eds.), Springer Verlag, Berlin, 2006

TAMING THE CURSE OF DIMENSIONALITY IN KERNELS AND NOVELTY DETECTION

Paul F. Evangelista
Department of Systems Engineering
United States Military Academy
West Point, NY 10996

Mark J. Embrechts
Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute
Troy, New York 12180

Boleslaw K. Szymanski
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180

Abstract. The curse of dimensionality is a well known but not entirely well-understood phenomena. Too much data, in terms of the number of input variables, is not always a good thing. This is especially true when the problem involves unsupervised learning or supervised learning with unbalanced data (many negative observations but minimal positive observations). This paper addresses two issues involving high dimensional data: The first issue explores the behavior of kernels in high dimensional data. It is shown that variance, especially when contributed by meaningless noisy variables, confounds learning methods. The second part of this paper illustrates methods to overcome dimensionality problems with unsupervised learning utilizing subspace models. The modeling approach involves novelty detection with the one-class SVM.

1 Introduction

High dimensional data often create problems. This problem is exacerbated if the training data is only one class, unknown classes, or significantly unbalanced classes. Consider a binary classification problem that involves computer intrusion detection. Our intention is to classify network traffic, and we are interested in classifying the traffic as either attacks (intruders) or non attacks. Capturing network traffic is simple - hookup to a LAN cable, run tcpdump, and you can fill a hard drive within minutes. These captured network connections can be described with attributes; it is not uncommon for a network connection to be described with over 100 attributes [14]. However, the class of each connection will be unknown, or perhaps with reasonable confidence we can assume that all of the connections do not involve any attacks.

The above scenario can be generalized to other security problems as well. Given a matrix of data, $\mathbf{X}$, containing N observations and m attributes, we are interested in classifying this data as either potential attackers (positive class) or non attackers (negative class). If m is large, and our labels, $\mathbf{y} \in \mathbb{R}^{N \times 1}$, are unbalanced (usually plenty of known non attackers and few instances of attacks), one class (all non attackers), or unknown, increased dimensionality rapidly becomes a problem and feature selection is not feasible due to the minimal examples (if any) of the attacker class.

2 Recent Work

The primary model explored will be the one-class SVM. This is a novelty detection algorithm originally proposed in [27]. The model is relatively simple but a powerful method to detect novel events that occur after learning from a training set of normal events. Formally stated, the one-class SVM considers $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \in \mathcal{X}$ instances of training observations and utilizes the popular "kernel trick" to introduce a non linear mapping of $\mathbf{x}_i \rightarrow \Phi(\mathbf{x}_i)$. Under Mercer's theorem, it is possible to evaluate the inner product of two feature mappings, such as $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$, without knowing the actually feature mapping. This is possible because $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \equiv \kappa(\mathbf{x}_i, \mathbf{x}_j)$ [2]. Φ will be considered a mapping into the feature space, F, from $\mathcal{X}$.

The following minimization function attempts to squeeze R, which can be thought of as the radius of a hypersphere, 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 to allow for this. A free parameter, $\nu \in (0, 1)$, enables the modeler to adjust the impact of the slack variables.

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

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

The lagrangian dual of the one class SVM is shown below in equation 2.

$$\max_{\alpha} \sum_i \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\upsilon N} \text{ and } \sum_i \alpha_i = 1$$

Cristianini and Shawe-Taylor provide a detailed explanation of one-class SVMs in [24]. Stolfo and Wang [25] successfully apply the one-class SVM to the SEA dataset and compare it with several of the techniques mentioned above. Chen uses the one-class SVM for image retrieval [8]. Schölkopf et. al. explore the above formulation of the one-class SVM and other formulations in [23]. Fortunately there is also freely available software that implements the one-class SVM, written in C++ by Chang and Lin [7].

The dimensionality problem faced by the one-class SVM has been mentioned in several papers, however it is typically a "future works" type of discussion. Tax and Duin clearly mention that dimensionality is a problem in [27], however they offer no suggestions to overcome this. Modeling in subspaces, which is the proposed method to overcome this problem, is not an altogether novel concept. In data mining, subspace modeling to overcome dimensionality is a popular approach. Aggarwal discusses this in [1]. Parsons et. al. provide a survey of subspace clustering techniques in [21]. The curse of dimensionality is largely a function of class imbalance and our apriori knowledge of the distribution of $(\mathbf{x}|\mathbf{y})$. This implies that the curse of dimensionality is a problem that impacts unsupervised problems the most severely, and it is not surprising that data mining clustering algorithms, an unsupervised method, has come to realize the value of modeling in subspaces.

3 Analytical Investigation

3.1 The Curse of Dimensionality, Kernels, and Class Imbalance

Machine learning and data mining problems typically seek to show a degree of similarity between observations, often as a distance metric. Beyer et. al. discuss the problem of high dimensional data and distance metrics in [3], presenting a probabilistic approach and illustrating that the maximally distant point and minimally distant point converge in distance as dimensionality increases. A problem with distance metrics in high dimensional space is that distance is typically measured across volume. Volume increases exponentially as dimensionality increases, and points tend to become equidistant. The curse of dimensionality is explained with several artificial data problems in [15].

Kernel based pattern recognition, especially in the unsupervised domain, is not entirely robust against high dimensional input spaces. A kernel is nothing more than a similarity measure between two observations. Given two observations, $\mathbf{x}_1$ and $\mathbf{x_2}$, the kernel between these two points is represented as $\kappa(\mathbf{x}_1, \mathbf{x_2})$. A large value for $\kappa(\mathbf{x}_1, \mathbf{x_2})$ indicates similar points, where smaller values indicate dissimilar points. Typical kernels include the linear kernel, $\kappa(\mathbf{x}_1, \mathbf{x_2}) = \langle \mathbf{x}_1, \mathbf{x_2} \rangle$, the polynomial kernel, $\kappa(\mathbf{x}_1, \mathbf{x_2}) = (\langle \mathbf{x}_1, \mathbf{x_2} \rangle + 1)^p$, and the popular gaussian kernel, $\kappa(\mathbf{x}_1, \mathbf{x_2}) = e^{(-\|\mathbf{x}_1 - \mathbf{x_2}\|/2\sigma^2)}$. As shown, these kernels are all functions of inner products. If the variables within $\mathbf{x}_1$ and $\mathbf{x_2}$ are considered random variables, these kernels can be modeled as functions of random variables. The fundamental premise of pattern recognition is the following:

$$(\kappa(\mathbf{x}_1, \mathbf{x_2})|y_1 = y_2) > (\kappa(\mathbf{x}_1, \mathbf{x_2})|y_1 \neq y_2) \tag{3}$$

If this premise is consistently true, good performance occurs. By modeling these kernels as functions of random variables, it can be shown that the addition of noisy, meaningless input variables degrades performance and the likelihood of the fundamental premise shown above.

In a classification problem, the curse of dimensionality is a function of the degree of imbalance. If there are a small number of positive examples to learn from, feature selection is possible but difficult. With unbalanced data, significant evidence is required to illustrate that a feature is not meaningful. If the problem is balanced, the burden is not as great. Features are much more easily filtered and selected.

A simple explanation of this is to consider a two sample Kolmogorov test [22]. This is a classical statistical test to determine whether or not two samples come from the same distribution, and this test is general regardless of the distribution. In classification models, a meaningful variable should behave differently depending on the class, implying distributions that are not equal. Stated in terms of distributions, if x is any variable taken from the space of all variables in the dataset, $(F_x(x)|y = 1)$ should not be equivalent to $(G_x(x)|y = -1)$. $F_x(x)$ and $G_x(x)$ simply represent the cumulative distribution functions of $(x|y = 1)$ and $(x|y = -1)$, respectively. In order to apply the two sample Kolmogorov test, the empirical distribution functions of $F_x(x)$ and $G_x(x)$ must be calculated from a given sample, and these distribution functions will be denoted as $F^*_{N_1}(x)$ and $G^*_{N_2}(x)$. N_1 will equate to the number of samples in the minority class, and N_2 equates to the number of samples in the majority class. These empirical distribution functions are easily derived from the order statistics of the given sample, which is shown in [22]. The Kolmogorov two sample test states that if the supremum of the difference of these functions exceeds a tabled critical value depending on the modeler's choice of α(sum of probabilities in two tails), then these two distributions are significantly different. Stated formally, our hypothesis is that $F_x(x) = G_x(x)$. We reject this hypothesis with a confidence of $(1 - \alpha)$ if equation 4 is true.

$$D_{N_1,N_2} = \sup_{-\infty<x<\infty} |F^*_{N_1}(x) - G^*_{N_2}(x)| > D_{N_1,N_2,\alpha} \tag{4}$$

For larger values of N_1 and N_2 (both N_1 and N_2 greater than 20) and $\alpha = .05$, we can consider equation 5 to illustrate an example. This equation is found in the tables listed in [22]:

$$D_{N_1,N_2,\alpha=.05} = 1.36\sqrt{\frac{N_1 + N_2}{N_1 N_2}} \tag{5}$$

If N_2 is fixed at 100, and N_1 is considered the minority class, it is possible to plot the relationship between m and the critical value necessary to reject the hypothesis.

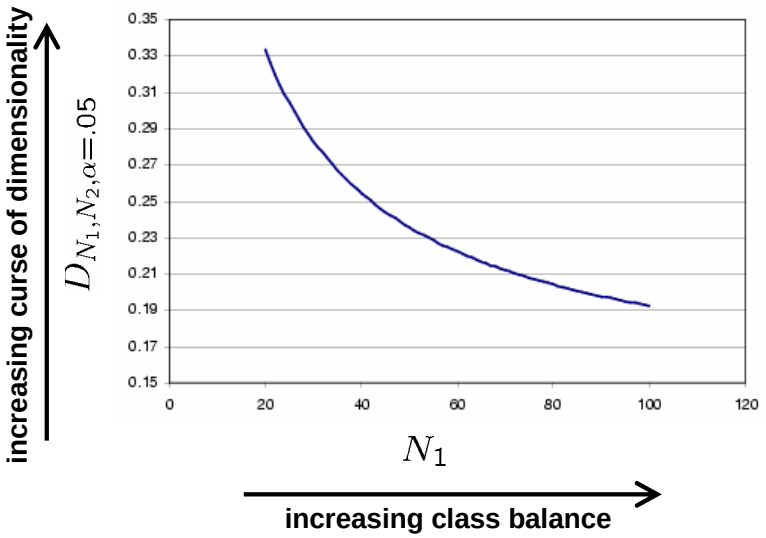

Figure 1. Plot of critical value for two sample Kolmogorov test with fixed N_2, $\alpha = .05$

Figure 1 illustrates the effect of class imbalance on feature selection. If the classes are not balanced, as is the case when $N_1 = 20$ and $N_2 = 100$, there is a large value required for D_{N_1,N_2}. It is also evident that if the classes were more severely imbalanced, D_{N_1,N_2} would continue to grow exponentially. As the classes balance, D_{N_1,N_2} and the critical value begins to approach a limit. The point of this exercise was to show that the curse of dimensionality is a function of the level of imbalance between the classes, and the two sample Kolmogorov test provides a compact and statistically grounded explanation for this.

3.2 Kernel Behavior in High Dimensional Input Space

An example is given in this section which illustrates the impact of dimensionality on linear kernels and gaussian kernels.

Consider two random vectors that will serve as artificial data for this example.

$$\mathbf{x_1} = (z_1, z_2, ..., z_m),\ z_i \sim N(0,1) \text{ i.i.d}$$
$$\mathbf{x_2} = (z_{1'}, z_{2'}, ..., z_{m'}),\ z_{i'} \sim N(0,1) \text{ i.i.d}$$
$$m' = m, \text{ and let } v_i = z_i z_{i'}$$

The expected value of v_i is zero. v_i is the product of two standard normal random variables, which follows an interesting distribution discussed in [12]. The plot of this distribution is shown in figure 2.

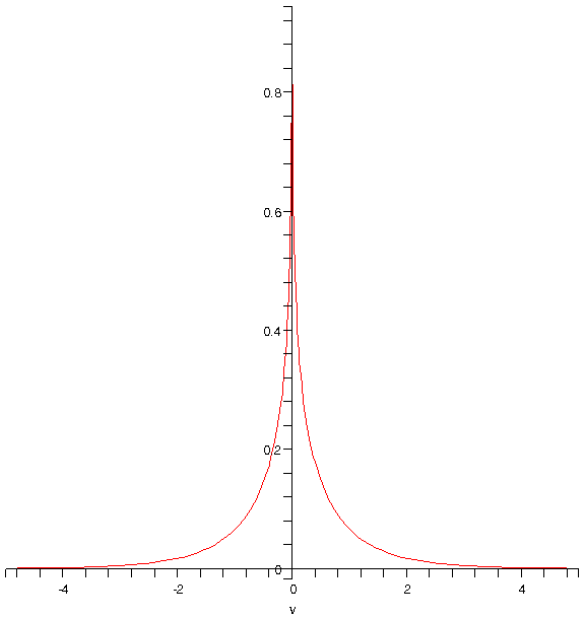

Figure 2. Plot of $v_i = z_i z_{i'}$

To find the expectation of a linear kernel, it is straightforward to see that $E(\langle \mathbf{x}, \mathbf{y} \rangle) = \sum_i v_i = E(z_1 z_{1'} + z_2 z_{2'} + ... + z_m z_{m'}) = 0$. The variance of the linear kernel can be found as follows:

$$f_{z_i, z_{i'}}(z_i, z_{i'}) \text{ is bivariate normal } \Rightarrow f_{z_i, z_{i'}}(z_i, z_{i'}) = \frac{1}{2\pi} e^{\frac{-(z_i^2 + z_{i'}^2)}{2}}$$

$$f_v(v) = \int_{-\infty}^{\infty} f_{z_i, z_{i'}}(z_i, \frac{v}{z_i}) \frac{1}{|z_i|} dz_i$$

$$E(v) = 0 \Rightarrow \text{ variance } = E(v^2) = \int_{-\infty}^{\infty} v^2 [f_v(v)] \partial v = 1$$

(verified by numerical integration)

Again considering the linear kernel as a function of random variables, $\kappa(\mathbf{x_1}, \mathbf{x_2}) = \langle \mathbf{x_1}, \mathbf{x_2} \rangle = \sum_{i=1}^{m} v_i$ is distributed with a mean of 0 and a variance of $\sum_{i=1}^{m} 1 = m$.

In classification problems, however, it is assumed that the distributions of the variables for one class are not the same as the distributions of the variables for the other class. Let us now consider v_- as a product of

dissimilar distributions, and v_+ as a product of similar distributions. Let $v_- = (z_i - 1)(z_{i'} + 1)$. v_- will be distributed with a mean of $\mu_- = E(z_i z_{i'} - z_{i'} + z_i - 1) = -1$, and a variance of 3 (verified through numerical integration). The linear kernel of the dissimilar distributions can be expressed as:

$$\kappa(\mathbf{x_1} - 1, \mathbf{x_2} + 1) = \sum_{i=1}^{m} v_-$$

This linear kernel is distributed with the following parameters:

$$\text{mean}_- = m\mu_- = -m, \text{variance } = m\sigma^2 = 3m$$

For the similar observations, let $v_+ = (z_i + 1)(z_{i'} + 1) = (z_i - 1)(z_i' - 1)$. The parameters of the kernel for the similar observations can be found in the same manner. v_+ is distributed with a mean of $\mu_+ = E(z_i z_{i'} + z_{i'} + z_i + 1) = 1$ and a variance of $\sigma^2 = 3$. The linear kernel of the similar distributions can be expressed as:

$$\kappa(\mathbf{x_1} + 1, \mathbf{x_2} + 1) = \sum_{i=1}^{m} v_+$$

This kernel is distributed with the following parameters:

$$\text{mean}_+ = m\mu_+ = m, \text{variance } = m\sigma^2 = 3m$$

The means and variances of the distributions of the linear kernels are easily tractable, and this is all the information that we need to analyze the effect of dimensionality on these kernels. In the above example, the mean of every variable for dissimilar observations differs by 2. This is consistent for every variable. Obviously, no dataset is this clean, however there are still interesting observations that can be made. Consider that rather than each variable differing by 2, they differ by some value ϵ_i. If ϵ_i is a small value, or even zero for some instances (which would be the case for pure noise), this variable will contribute minimally in distinguishing similar from dissimilar observations, and furthermore the variance of this variable will be entirely contributed. Also notice that at the rate of $3m$, variance grows large fast.

Based on this observation, an assertion is that for the binary classification problem, bimodal variables are desirable. Each mode will correspond to either the positive or negative class. Large deviations in these modes, with minimal variation within a class, are also desired. An effective model must be able to distinguish v_- from v_+. In order for this to occur, the model needs good separation between $mean_-$ and $mean_+$ and variance that is under control.

It is also interesting to explore the gaussian kernel under the same example. For the gaussian kernel, $\kappa(\mathbf{x_1}, \mathbf{y_2}) = e^{-\|\mathbf{x_1} - \mathbf{x_2}\|^2 / 2\sigma^2}$. This kernel is entirely dependent upon the behavior of $\| \mathbf{x_1} - \mathbf{x_2} \|^2$ and the modeler's choice of the parameter σ (which has no relation to variance).

Restricting our attention to $\| \mathbf{x_1} - \mathbf{x_2} \|^2$, an initial observation is that this expression is nothing more than the euclidean distance squarred. Also, if $\mathbf{x_1}$ and $\mathbf{x_2}$ contain variables that are distributed $\sim N(0, 1)$, then $(\mathbf{x_1} - \mathbf{x_2})$ contains variables distributed normally with a mean of 0 and a variance of 2.

Let $w = (z_i - z_{i'})^2$, implying that $w/2$ is a chi-squarred distribution with a mean of one (which will be annotated as $\chi^2(1)$). This also indicates that $w = 2\chi^2(1)$, indicating that w has a mean of 2 and a variance of 8 (verified by numerical integration).

Therefore, $\| \mathbf{x_1} - \mathbf{x_2} \|^2 = \sum_{i=1}^{m} w_i$ will have a distribution with a mean of $2m$ and a variance of $8m$. Notice that the variance grows much faster under this formulation, indicating even more sensitivity to noisy variables.

The purpose of the above example is to show how every variable added will contribute to the overall behavior of the kernel. If the variable is meaningful, the pattern contributed to the -1 class is not equivalent to the pattern contributed to +1 class. The meaningfulness of the variable can also be considered in terms of cost and benefit. The benefit of including a variable in a classification model is the contribution of the variable towards pushing $mean_-$ away from $mean_+$. The cost of including a variable involves the variance. This variance will be included regardless of the significance of the benefit.

3.3 The Impact of Dimensionality on the One-Class SVM

In order to illustrate the impact of dimensionality on kernels and the one-class SVM specifically, an experiment with artificial data was constructed. This data models a simple pattern involving standard normal distributions where the positive class and negative class have a difference of 2 between their means. This model can be presented as follows:

$$\mathbf{x}_{+1} = (z_1 + 1, z_2 + 1, z_3, ..., z_m),\ z_i \sim N(0,1) \text{ i.i.d}$$
$$\mathbf{x}_{-1} = (z_{1'} - 1, z_{2'} - 1, z_{3'} ..., z_{m'}),\ z_{i'} \sim N(0,1) \text{ i.i.d}$$

The true pattern only lied in the first two variables. All remaining variables were noise. Three types of kernels were examined: the linear kernel, polynomial kernel, and gaussian kernel. Only the results from the gaussian kernel are shown here, however degradation of performance occurred with all kernels. The performance metric used was the area under the ROC curve (AUC).

Table 1. One Class SVM (gaussian kernel) experiment for various dimensions on artificial data

Dimensions	AUC	R^2
2	0.9201	0.5149
5	0.8978	0.4665
10	0.8234	0.4356
50	0.7154	0.3306
100	0.6409	0.5234
250	0.6189	0.4159
500	0.5523	0.6466
1000	0.5209	0.4059

The gaussian kernels in this experiment were tuned using an auto-tuning method. Typically for gaussian kernels, a validation set of positive and negative labeled data is available for tuning σ. In unsupervised learning, these examples of positive labeled data do not exist. Therefore, the best tuning possible is to achieve some variation in the values of the kernel without values concentrated on either extreme. If σ is too large, all of the values will tend towards 1. If too small, they tend to 0. The auto tuning function ensures that the off-diagonal values for $\kappa(\mathbf{x}_{+1}, \mathbf{x}_{-1})$ average between .4 and .6, with a min value greater than .2.

4 A Framework to Overcome High Dimensionality

A novel framework for unsupervised learning, or anomaly detection, has been investigated to solve unsupervised learning problems of high dimension [10, 11]. This technique is designed for unsupervised models, however the fusion of model output applies to any type of classifier that produces a soft (real valued) output. This framework involves exploring subspaces of the data, training a separate model for each subspace, and then fusing the decision variables produced by the test data for each subspace. Intelligent subspace selection has also been introduced within this framework.

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 [18]. Significant work in this field has been contributed by Kuncheva in [16–19]. Bonissone et. al. investigated the effect of different fuzzy logic triangular norms based upon the correlation of decision values from multiple classifiers [4]. The majority of work in this field has been devoted to supervised learning, with less effort addressing unsupervised problems [26]. The research that does address unsupervised ensembles involves clustering almost entirely. There is a vast amount of literature that discusses subspace clustering algorithms [21]. 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 [28], and Ma and Perkins who utilize the one-class SVM for time series prediction and combine results from intermediate phase spaces [20]. The work in this paper has also been inspired by Ho's Random Subspace Methods in [13]. Ho's method randomly selects subspaces and constructs a decision tree for each subspace; trees are then aggregated in the end by taking the mean. Breiman's work with bagging [5] and random forests [6] was also a significant contribution in motivating this work. Breiman's bagging technique involves

bootstrap sampling from a training set and creating a decision tree for each sample. Breiman also uses the mean as the aggregator. The random forest technique explores decision tree ensembles from random subsets of features, similar to Ho's method.

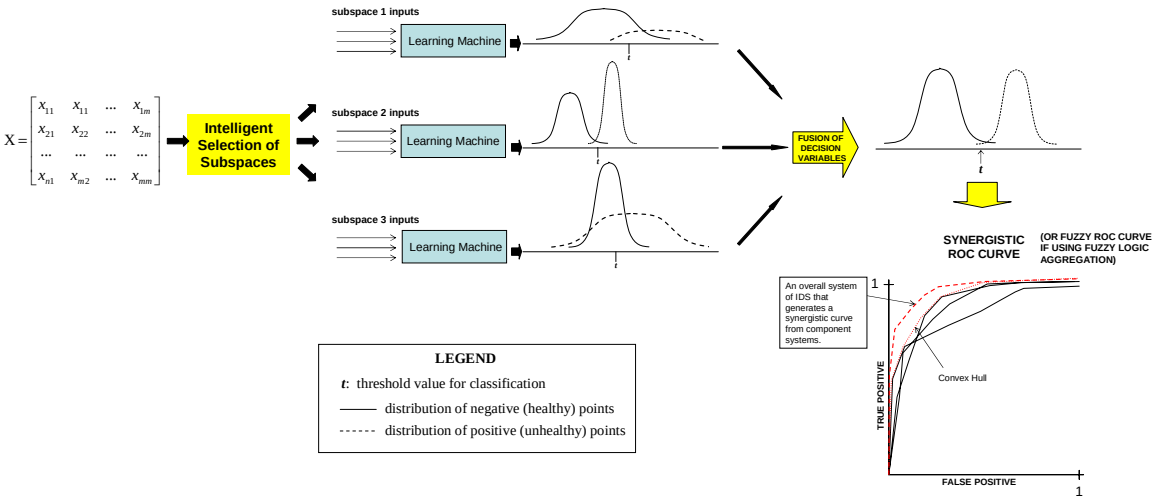

Figure 3. A sketch of subspace modeling to seek synergistic results.

Intersections(T-Norms)			Averages		Unions(T-Conorms)		
0	**max (0, $x + y - 1$)** *(bounded product)*	***x* x *y*** *(algebraic product)*	**min(x,y)**	**max (x,y)**	***x* + *y* - *x* x *y*** *(algebraic sum)*	**min(1, $x + y$)** *(bounded sum)*	**1**

Figure 4. Aggregation operators

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 following hypotheses show exceptional promise based on initial empirical results:

1. Intelligent subspace modeling 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.

Promising results from this approach have been published in [10, 11]. As previously discussed, aggregation of models with fuzzy logic aggregators is an important aspect. Given unbalanced data (minority positive class), it has been observed that fusion with T-norms behaves well and improves performance. Figure 4 illustrates the spectrum of fuzzy logic aggregators.

The results shown in table 2 and figure 5 illustrate the improvements obtained through our ensemble techniques for unsupervised learning. The plot of the ROC curves shows the results from using 26 original variables that represented the Schonlau et. al. (SEA) data [9] as one group of variables with the one-class SVM and the result of creating 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, especially in the SEA data, with the most significant improvement in the T-norms.

The ionosphere data 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 $l = 3$.

Table 2. Results of SEA data with diverse and non-diverse subsets

	SEA data	Ionosphere data
Base AUC (using all variables)	*.7835*	*.931*
T-norms		
minimum	.90	.96
algebraic product	.91	.61
T-conorms		
maximum	.84	.69
algebraic sum	.89	.69

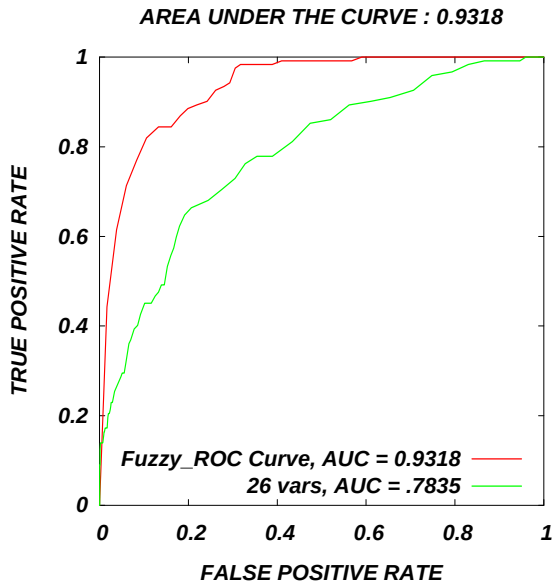

Figure 5. ROC for SEA data using algebraic product with contention

5 Discussion and Conclusion

There were two components to the research presented in this paper. The first component involved exposing the impact of the curse of dimensionality with kernel methods. This involved illustrating that more is not always better in terms of variables, but more importantly that the impact of the curse of dimensionality grows as class imbalance becomes more severe. Kernel methods are not immune to problems involving high dimensional data, and these problems need to be understood and managed.

The second component of this research involved the discussion and brief illustration of a proposed framework for unsupervised modeling in subspaces. Unsupervised learning, especially novelty detection, has important applications in the security domain. This applies especially to computer and network security. Future directions for this research include exposing the theoretical foundations of unsupervised ensemble methods and exploration of other ensembles for the unbalanced classification problem.

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. Kristin P. Bennett and Colin Campbell. Support Vector Machines: Hype or Hallelujah. 2(2), 2001.
3. Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is "Nearest Neighbor" Meaningful? *Lecture Notes in Computer Science*, 1540:217–235, 1999.
4. Piero Bonissone, Kai Goebel, and Weizhong Yan. Classifier Fusion using Triangular Norms. Cagliari, Italy, June 2004. Proceedings of Multiple Classifier Systems (MCS) 2004.

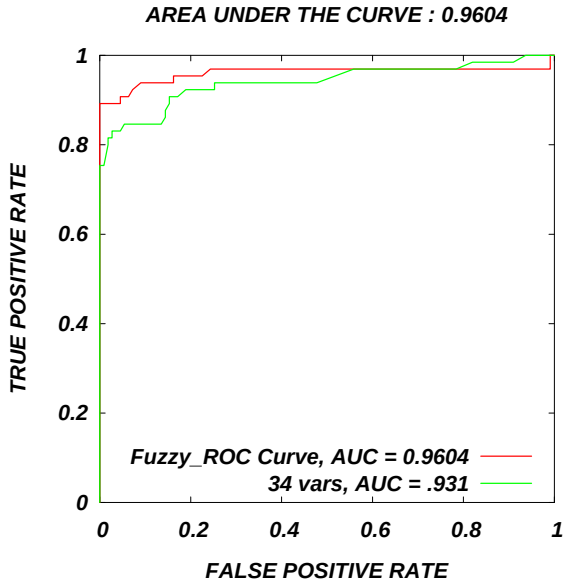

Figure 6. ROC plot for ionosphere data with minimize aggregation technique

5. Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
6. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
7. Chih Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. http://www.scie.ntu.edu.tw/ cjlin/libsvm, Accessed 5 September, 2004.
8. 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.
9. 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.
10. Paul F. Evangelista, Piero Bonissone, Mark J. Embrechts, and Boleslaw K. Szymanski. Fuzzy ROC Curves for the One Class SVM: Application to Intrusion Detection. Montreal, Canada, August 2005. International Joint Conference on Neural Networks.
11. Paul F. Evangelista, Piero Bonissone, Mark J. Embrechts, and Boleslaw K. Szymanski. Unsupervised Fuzzy Ensembles and Their Use in Intrusion Detection. Bruges, Belgium, April 2005. European Symposium on Artificial Neural Networks.
12. Andrew G. Glen, Lawrence M. Leemis, and John H. Drew. Computing the Distribution of the Product of Two Continuous Random Variables. *Computational Statistics and Data Analysis*, 44(3):451–464, 2004.
13. Tin Kam Ho. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.
14. Alexander Hofmann, Timo Horeis, and Bernhard Sick. Feature Selection for Intrusion Detection: An Evolutionary Wrapper Approach. Budapest, Hungary, July 2004. International Joint Conference on Neural Networks.
15. Mario Koppen. The Curse of Dimensionality. (held on the internet), September 4-18 2000. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5).
16. Ludmila I. Kuncheva. 'Fuzzy' vs. 'Non-fuzzy' in Combining Classifiers Designed by Boosting. *IEEE Transactions on Fuzzy Systems*, 11(3):729–741, 2003.
17. Ludmila I. Kuncheva. That Elusive Diversity in Classifier Ensembles. Mallorca, Spain, 2003. Proceedings of 1st Iberian Conference on Pattern Recognition and Image Analysis.
18. Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, Inc., 2004.
19. Ludmila I. Kuncheva and C.J. Whitaker. Measures of Diversity in Classifier Ensembles. *Machine Learning*, 51:181–207, 2003.
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. 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.

22. Vijay K. Rohatgi and A.K.Md. Ehsanes Saleh. *An Introduction to Probability and Statistics*. Wiley, second edition, 2001.
23. 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.
24. John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
25. 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.
26. 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.
27. David M.J. Tax and Robert P.W. Duin. Support Vector Domain Description. *Pattern Recognition Letters*, 20:1191–1199, 1999.
28. 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.