Event Detection Based on Nonnegative Matrix Factorization: Ceasefire Violation, Environmental, and Malware Events

Barry Drake[1], Tiffany Huang[1], Ashley Beavers[1], Rundong Du[1] and Haesun Park[1]

[1] Georgia Institute of Technology, 75 5TH St. NW STE 900, Atlanta, GA 30308-1018, United States of America
{Barry.Drake, Tiffany.Huang, Ashley.Beavers}@gtri.gatech.edu, {rdu, hpark}@cc.gatech.edu

Abstract. Event detection is a very important problem across many domains and is a broadly applicable encompassing many disciplines within engineering systems. In this paper, we focus on improving the user's ability to quickly identify threat events such as malware, military policy violations, and natural environmental disasters. The information to perform these detections is extracted from text data sets in the latter two cases. Malware threats are important as they compromise computer system integrity and potentially allow the collection of sensitive information. Military policy violations such as ceasefire policies are important to monitor as they disrupt the daily lives of many people within countries that are torn apart by social violence or civil war. The threat of environmental disasters takes many forms and is an ever-present danger worldwide, and indiscriminate regarding who is harmed or killed. In this paper, we address all three of these threat event types using the same underlying technology for mining the information that leads to detecting such events. We approach malware event detection as a binary classification problem, i.e., one class for the threat mode and another for non-threat mode. We extend our novel classifier utilizing constrained low rank approximation as the core algorithm innovation and apply our Nonnegative Generalized Moody-Darken Architecture (NGMDA) hybrid method using various combinations of input and output layer algorithms. The new algorithm uses a nonconvex optimization problem via the nonnegative matrix factorization (NMF) for the hidden layer of a single layer perceptron and a nonnegative constrained adaptive filter for the output layer estimator. We first show the utility of the core NMF technology for both ceasefire violation and environmental disaster event detection. Next NGMDA is applied to the problem of malware threat events, again based on the NMF as the core computational tool. Also, we demonstrate that an algorithm should be appropriately selected for the data generation process. All this has critical implications for design of solutions for important threat/event detection scenarios. Lastly, we present experimental results on foreign language text for ceasefire violation and environmental disaster events. Experimental results on a KDD competition data set for malware classification are presented using our new NGMDA classifier.

Keywords: Malware Detection · Event Detection · Perceptron · Clustering · Nonnegative Matrix Factorization · Adaptive Filtering · Hybrid Classifier · Topic Modeling · Classification

1 Introduction

In this paper, we present algorithms for a single layer perceptron (SLP) that combines two algorithms into a new hybrid classification framework and implements locally tuned receptive fields. Our new framework builds on the hybrid architecture first reported by Moody and Darken [1, 2] and generalized by Drake, et al [3]. We use this hybrid classification framework with new algorithms that can be updated on a per-sample or mini-batch (small number of samples) basis and are sensitive to the input domain of the data. Our classification framework has advantages over multilayer perceptron architectures and the support vector machine (SVM) [3]. Our new hybrid framework and algorithms will be presented within the context of three important event detection problems: malware, ceasefire, and environmental disaster event detection. We further extend our Generalized Moody-Darken Architecture (GMDA) framework to the nonnegative domain using the nonnegative matrix factorization (NMF). When combined with a nonnegativity constrained adaptive filter the extended GMDA framework is called the Nonnegative GMDA (NGMDA), which is able to discover discriminative features for classification and demonstrate better performance for nonnegative input data.

Our GMDA classifier uses a clustering method to find the centers of the activation units. The kmeans algorithm has typically been used to determine the activation unit centers. However, for nonnegative input data, we show in this paper experimental results that demonstrate a loss of information in the classifier, which increases the classification error. We extend the GMDA further with an objective function based on a constrained low rank approximation (CLRA) method called the nonnegative matrix factorization (NMF) [3, 4], which we have applied in numerous application domains for text analytics. CLRA methods [4] have played a crucial role as one of the most fundamental tools in machine learning, data mining, image processing, information retrieval, computer vision, signal processing, and other areas of computational science and engineering. Since the NMF objective function is formulated using nonnegative constraints, application of NMF to certain problems produces results that are more interpretable for many types of problems. In [3] we demonstrated the importance of incorporating nonnegative constraints for applications such as image processing, chemometrics, and text analytics. NMF has become a valuable tool for many applications such as clustering, subspace-based topic modeling, general dimension reduction (PCA-like), hyperspectral image processing, and many more. In this paper, NMF is utilized as the first stage of our new NGMDA classifier. In previous work [3] we demonstrated the efficacy of NMF-based GMDA with preliminary results and demonstrated comparable performance to a support vector machine (SVM) classifier. We show that implementing the output layer of the NGMDA with a nonnegative adaptive filter improves classification performance over unconstrained adaptive algorithms. An NMF event detection methodology is demonstrated for foreign language texts in Arabic (ceasefire violation events in Yemen) and Chinese (environmental disaster events). The classification results are demonstrated on event detections for malware (cybersecurity). Experimental results are presented on malware event classification, and event detection from text data. Thus, the experimental results demonstrate NMF for stand-alone applications and as part of a hybrid classification architecture, the NGMDA.

2 Background Material

As a review of the underlying components of the GMDA, we briefly describe the components of the GMDA for the hidden and output layers. First we provide some important mathematical properties of two neural network architectures. The Moody-Darken single layer perceptron (SLP) architecture has advantages over a multi-layer perceptron (MLP) architecture in both a fundamental mathematical sense and performance considerations. Both share the *universal approximation* property, which provides existence proofs of an interpolating set of basis polynomials for arbitrary inputs. One advantage of the SLP over a MLP architecture is the *best approximation* property, which guarantees that there is a set of approximating functions corresponding to all possible choices of the model parameters with one function from the set that minimizes the approximation errors [5, 6, 7].

The general equation for the GMDA is

$$\text{y} = b + \sum_{i=1}^{k_{hidden}} w_i \, f\left(\frac{\exp\|x - c_i\|^2}{2s_i^2}\right) \tag{1}$$

where x is the input data, c_i are the activation unit centers, w_i are the activation unit magnitudes computed by the output layer algorithm, f denotes the radial basis function (RBF) kernel, the summation is over the number of activation units in the hidden layer, b is a bias term, and y is the output. As stated above the c_i are computed using a clustering method (NMF for nonnegative data inputs), which performs a dimension reduction and, thus, reduces the computational complexity by decreasing the amount of data required for the output layer algorithm. The s_i are the estimated standard deviations of the computed clusters, which are based on the normalized sum of the distances between the samples in the cluster to its center and the output layer weights of the GMDA are computed by an adaptive filter [8, 5]. Various adaptive filters are possible for estimating the weights. Examples are LMS, Recursive Least Squares (RLS), and QR Decomposition Recursive Least Squares (QRDRLS) and many more. For this paper, we examine both unconstrained and nonnegative variants of adaptive filters, mainly the nonnegative LMS (NNLMS) [9] and our Adaptive Sequential Coordinate-wise Algorithm (ASCA).

First we review the SLP hidden layer algorithms from [3] followed by the background for the output layer algorithms. Throughout this paper $A \in \mathbb{R}^{mxn}$ where m is the number of features and n are the columns that represent the data. Assuming $rank(A) = r$, the low rank approximation of A, for $k \leq r$, is denoted $rank(\hat{A}) = k$, i.e., we approximate the matrix A with factors that, when multiplied together produce a matrix $\hat{A}$ that is close to A in some norm, usually the Frobenius norm [10], which we use throughout this paper. The Frobenius norm is analogous to the vector 2-norm (L_2) but for matrices. One of the most commonly used low rank approximations is the singular value decomposition (SVD) [10]. A general form of a low rank approximation can be expressed as

$$\min_{W,H} \|A - WH\|_F \tag{2}$$

where F denotes the Frobenius norm, $W \in \mathbb{R}^{mxk}$, and $H \in \mathbb{R}^{kxn}$ and k is the rank of $A \in \mathbb{R}^{mxn}$. Depending on the constraints on W and H, (1) may be a nonconvex optimization problem. However, the SVD gives the optimal global solution when the objective of equation (1) is *unconstrained*. Equation (2) will be the basic equation for discussing the clustering methods used in the SLP hidden layer. In the case of NMF, (2) becomes

$$\min_{W,H\geq 0} \|A - WH\|_F \tag{3}$$

where equation (3) is the same as (2) but with nonnegativity constraints on W and H, which makes (3) a nonconvex optimization problem, a very difficult problem to solve. Generally speaking, (2) achieves a global minimum while (3) is only guaranteed to achieve a local minimum at best. However, by choosing the best algorithms to solve (3) convergence to a local minimum is guaranteed. The choice of algorithm is extremely important in order to ensure convergence to a stationary point of the minimization surface [4]. In general, equation (3) performs clustering of the input data be it text (for documents) or pixels (images). Topic modeling discussed in a latter section is document clustering where each cluster is composed of semantically related terms (words) and k is user chosen as the desired number of clusters (topics) to compute. The W is the cluster *indicator* matrix and H is known as the cluster *membership* matrix. Thus, H can be used to retrieve documents within clusters (topics).

As mentioned above, the output layer of the GMDA is implemented using adaptive algorithms such as adaptive filters [5]. Adaptive filters are commonly found in the signal processing literature and have applications in adaptive beamforming, direction finding, speech processing, and many more. A commonly used adaptive filter is the Least Mean Squares (LMS) algorithm, which is an approximation of the gradient descent method. The LMS algorithm is simple to implement and has some nice statistical properties but may suffer from slow convergence and misadjustment error, i.e., the learning curve (MSE) slowly approaches the asymptotic minimum value. In this paper, we use a NNLMS, which for the first time, is incorporated into a hybrid learning architecture such as the NGMDA for malware event detection.

In general, the learning rule of the output layer can be cast as a model-building problem via regression. Thus, the general form of this learning rule is

$$\min_{X} \|BX - D\|_F \tag{5}$$

where equation (5) is the general linear regression problem with multiple right hand sides. We solve (5) using one of the adaptive filter methods mentioned above as each data sample becomes available (adaptive) and for a single right hand side, which computes the output of the hidden layer. Both the hidden layer and the output layer of the GMDA or NGMDA can be updated on a per sample basis with suitably chosen algorithms. For example, for arbitrary data 'kmeans + LMS' (GMDA); for nonnegative data 'NMF + NNLMS' (NGMDA) is an implementation that can be updated/downdated in both the hidden and output layers. Another approach for the output layer, presented in [3], is the ASCA, the adaptive sequential coordinate-wise algorithm, which is an adaptive version of SCA, sequential coordinate-wise algorithm [11], for solving the nonnegative least squares (NLS) problem. We will show that the results using ASCA, rather than NNLMS in the output layer are superior, and that, with NMF as the hidden layer algorithm, NMF-based NGMDA outperforms kmeans for computing the hidden layer activation functions. This serves to emphasize the better numerical properties of our new ASCA algorithm used for the NGMDA output layer as reported in [3] and the superior performance achieved when a properly constrained algorithm for the data is used to compute the hidden layer, i.e., NMF rather than kmeans.

Now we review the importance of choosing the correct algorithm for the data [3]. We briefly review those considerations here with simulated half-moon data in Figure 1 and simulated chemical detections in Figure 2 below. These figures illustrate that the choice of algorithms is critical, though some algorithms may have higher computational complexity, the benefits may outweigh the costs. Since the half-moon data was real (+, 0, - values), Figure 1a, b shows results with kmeans as the hidden layer algorithm with 30 activation units. Figure 1 demonstrates that classification results can depend critically on the selected algorithm.

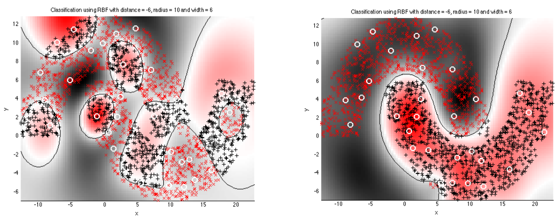


Fig. 1. a. GMDA with LMS adaptive filter as the output layer algorithm; b. GMDA with output layer algorithm RLS. We observe that LMS breaks down in this case while RLS obtains good classification results [1].

The chemical detection problem, Figure 2 a, b, uses Raman spectroscopy to discover chemical constituents of a chemical sample. For details see [3, 12]. There are two issues: The correct computational model for the data; Numerically robust algorithms that are not degraded by near collinearities (rank deficiency). The figures illustrate both cases for the generalized likelihood ratio test (GLRT) detection algorithm [13]. Note that the chemicals are labeled with species number in the library, which may be traced back to the wavenumber for a specific chemical. The tests were for three chemical species: two closely spaced and one more isolated at 0.3g/m^2.

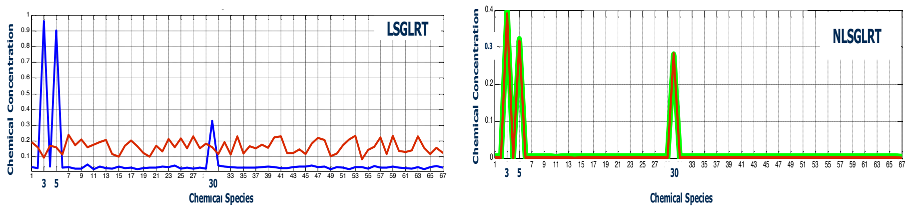


Fig. 2. a. Generalized likelihood ratio test *without* numerical problems (blue) and *with* numerical problems (red). Note that concentrations were not determined correctly and that there are no detections when numerical problems are present (red). b. detection results are clearly corrected with the right computational model: a. uses *standard* least squares; b uses *nonnegative* least squares.

With the above results shown in Figures 1 and 2, and knowing that NMF is the correct computational model for text data, in the next section we demonstrate its use for event detection with a human in the loop process on foreign text. The effectiveness of NMF for this application sets the stage for the development of the NGMDA classifier where we compare experiments with various combinations of computational models.

3 Observer-Based Event Detection from Foreign Language Text Using NMF-based Topic Modeling

In this section, we demonstrate the use of NMF topic modeling for processing nonnegative text data in order to discover latent information within a text corpus. Topic modeling is a methodology that clusters documents into semantically related words. Within each cluster (topic) the highest frequency keywords reveal the overall concept of the topic. Topic refinement is used to uncover topics that are more relevant to the domain of interest. Our approach uses a human-in-the-loop to interpret the NMF factors in order to accomplish search space reduction; since this approach does not depend on incorporation of complex language considerations, meta data, or a priori identification of target keywords, it can be used in contexts where there has been little natural language processing performed, where reliable meta data is not available, and where keywords are initially unknown. Since NMF is a subspace-based method we can easily access the original documents within certain topics and, if determined not to be relevant, eliminate them to retain only information relevant to our domain questions. After eliminating these documents from the original corpus, a reduced corpus is then analyzed using our topic modeling algorithm. In Figure 3 a, b is shown the process to refine the topic model where 3a shows the process and 3b shows the selection and elimination of irrelevant documents.

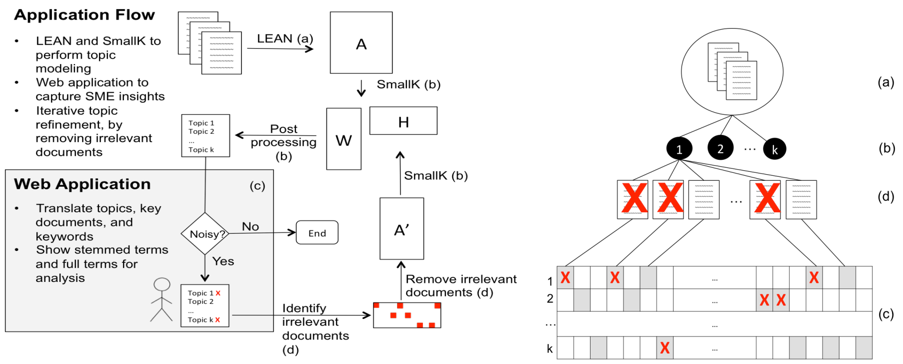


Fig. 3. a. Topic refinement process with human-in-the-loop; b. elimination of selected documents based on subject matter expert (SME) input.

A relevance score can be computed and a threshold set to determine whether or not a document should be eliminated from the corpus. The relevance score can be computed as

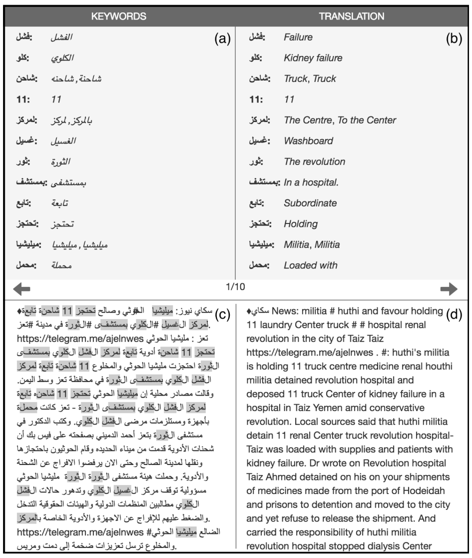
$$g_i > \max\left(g_i^T \times threshold\right) \quad \textbf{(6)}$$

which is computed from the normalized columns of H in

$$\min_{W,H\geq 0} \|A - WH\|_F \quad \textbf{(7)}$$

Recall that H is called the topic (cluster) membership matrix, i.e., each column of the membership matrix tells us which document belongs to which topic and can even indicate how well a document represents the cluster (topic), which is the basis for equation (6). These relationships are the key elements for determining the documents to eliminate at each iteration of the refinement process in Figure 3a.

In order to demonstrate the effectiveness of this process on a real problem, ceasefire violation events in Yemen will be discovered roughly during the time period around May, 2016. NMF-based topic modeling and the topic refinement processes are used to discover these events from raw Arabic telegrams (social media not available), i.e., the telegrams are *not* translated to English before using our tools to discover topics.



Topic 1	
Keywords	Event
منطقة | غارة | جوية | تحالف العربي | المليشيا | لطيران | جبهة | الشقب Arab Coalition, attack, raids, district, Shaqab, front, aviation, militia	From May 20 to 22, the national army and Houthi militia fought for control of Taiz city center, resulting in Houthi control of many roads and national control of the city center.
Topic 2	
Keywords	Event
الغسيل | شاحنة | الفشل الكلوي | ادوية | محملة | ميليشيا | تحتجز | بمستشفى kidney failure, truck, Dialysis Center, hospital, detain, militia, loaded, medicine	On May 21st, the Houthi militia detained 11 trucks traveling to the Taiz Dialysis Center.
Topic 3	
Keywords	Event
باليستية | الحديدة | لجنة التهدئة | بصواريخ | الفرعية | محافظة بابل pacification committee, Hodeidah, ballistic, Babil province, subcommittee, missile	On May 18th, the Houthi militia sent ballistic missiles from Hodeidah to Taiz.

Fig. 4. a. Topic modeling on Arabic telegram text; b. findings based on topic refinement.

Figure 4a shows some of the intermediate processing steps used to refine the topic modeling. These are illustrated in Figure 4a, pane (a), (b), (c), and (d). (a) Stemmed keywords followed by the full keywords in order of frequency, e.g., target: targets, targeted, etc.; (b) translations of the full keywords for each stemmed keyword; (c) original text with highlighted keywords; (d) translated text. The relevant topics found are shown in Figure 4b, which shows 3 topics with events and their dates that indicate ceasefire violations. In Figure 4, results are shown for the geographical region around Taiz, Yemen. The analysis generally focused on major cities where ceasefire violations were more likely and the data sets were larger and richer in content. The key point is that specific events were discovered using a semi-supervised methodology based on the NMF.

Figures 5 a and b below show the results using our refined topic modeling process for environmental events in China during 2012. As above, the raw Chinese text is processed and NMF topic modeling is used to analyze the results. Revealed in Topic 1 is a straw fire that causes severe haze in Jiangsu and Hubei Provinces, which occurred around June 9. Topic 5 reveals a severe rain storm that killed dozens of people in Beijing on July 21.

A point to keep in mind is that these environmental events were discovered in a semi-supervised manner with a human in the loop. The topic modeling algorithm, however, is unsupervised and discovers the topics and related keywords automatically without any human intervention. Once the initial topic modeling is performed the top keywords are determined for each topic, which span the topic concept. A label can be determined for each topic and the topics can become categories for classification. This is consistent with the notion that clustering is often referred to as unsupervised classification.

<table>
<tr><th colspan="2">Topic 1</th></tr>
<tr><th>Keywords</th><th>Event</th></tr>
<tr><td>霾 | 秸秆 | 雾 | 空气 | 焚烧 | 武汉 | 能见度 | pm | 质量 | 污染

haze, straw, fog, air, incineration, Wuhan, visibility, pm, quality, pollution</td><td>Straw burning causes severe haze in Jiangsu Province and Hubei Province beginning June 9, 2012.</td></tr>
<tr><th colspan="2">Topic 2</th></tr>
<tr><th>Keywords</th><th>Event</th></tr>
<tr><td>云南省 | 地震 | 宁 | 丽江 | 彝族 | 盐 | 云南 | 宁乡 | 交界 | 灾区

Yunnan, earthquake, Ning, Lijiang, Yi, Salt, Yunnan, Ningxiang, junction, disaster area</td><td>On June 24, 2012, an earthquake occurred at the junction of Sichuan and Yunnan Provinces.</td></tr>
<tr><th colspan="2">Topic 3</th></tr>
<tr><th>Keywords</th><th>Event</th></tr>
<tr><td>黄河 | 流量 | 陕西省 | 洪峰 | 秒 | 立方米 | 洪水

The Yellow River, flow, Shaanxi Province, flood peak, seconds, cubic meters,flood</td><td>Flooding occurred in Shaanxi Province on July 27, 2012.</td></tr>
</table>

<table>
<tr><th colspan="2">Topic 4</th></tr>
<tr><th>Keywords</th><th>Event</th></tr>
<tr><td>热带 | 风暴 | 泰 | 海面 | 沿海 | 利 | 级 | 南海 | 移动 | 风力

tropical, storm, Ta-, sea surface, coastal, -lim, scale, South China Sea, move, wind force</td><td>Tropical storm "Talim" hit the South China Sea on June 18, 2012.</td></tr>
<tr><th colspan="2">Topic 5</th></tr>
<tr><th>Keywords</th><th>Event</th></tr>
<tr><td>遇难 | 确认 | 身份 | 66 | 搜寻 | 者 | 遗体 | 北京市

killed, confirmed, identity, 66, search, people, corpse, Beijing</td><td>A rainstorm in Beijing killed dozens of people on July 21, 2012.</td></tr>
<tr><th colspan="2">Topic 6</th></tr>
<tr><th>Keywords</th><th>Event</th></tr>
<tr><td>韦 | 森 | 特 | 台风 | 广东 | 登陆 | 沿海 | 海南 | 风力

Vi-, -cen-, -te, typhoon, Guangdong, landing, coastal, Hainan, wind force</td><td>Typhoon Vicente was in South China from July 23 to 24, 2012.</td></tr>
</table>

Fig. 5. Topic modeling on Chinese newspaper text; a. topics 1-3; b. topics 4-6.

The above examples of using NMF for topic model event detection and, in the next section, classification, doesn't nearly cover the utility for which this versatile mathematical method is capable: image segmentation, hyperspectral image processing, speckle removal from noisy images, non-stationary speech denoising, motion detection from video sequences, music analysis, bioinformatics applications, chemometrics, and many more.

4 NMF-based Generalized Moody-Darken Architecture (NGMDA) for Malware Event Detection: Experiments and Results

The GMDA can also be used as the clustering method for a hybrid classification algorithm by configuring the hidden and output layers with algorithms that constrain the solution to the nonnegative domain for nonnegative input data. In this section, the data is not composed of documents but rather network intrusion data, malware. By utilizing nonnegativity constrained low rank approximation in the hidden layer and a nonnegativity constrained output layer adaptation rule, the classifications can be mapped to the domain where interpretation of the results is possible, whereas without these constraints the results may not only be not interpretable, but also fail completely to provide meaningful solutions. This will be further illustrated by using NMF as the hidden layer algorithm for GMDA.

In Figures 1 and 2 above we illustrated that algorithm selection is an important consideration, which justifies the reengineering of the original Moody-Darken SLP architecture with algorithms that conform to the specifications indicated by the input data, e.g., text, image data, chemical concentrations, etc.

Several experiments with various algorithm combinations implemented in our GMDA framework were applied to the malware dataset[1] that is comprised of 21 attack types on network systems. The data consisted of 125973 training samples and 22543 testing samples with 41 features. For the experiments, the data was further processed so that all types of attacks were gathered into one class; instances of normal network traffic formed another class. Thus, the data was composed of two classes: malware and non-malware.

First, the analysis was focused on the performance of kmeans as the hidden layer with the LMS filter output layer algorithm as in the original Moody-Darken hybrid architecture [1]. The second major testing was performed on the extension of the original architecture to handle nonnegative input data. Thus, extending the original architecture to our GMDA. The third set of experimental results use NMF as the hidden layer algorithm with a nonnegativity constrained output layer adaptive filter, which is the NGMDA. The experiments culminate with NMF as the hidden layer and our ASCA nonnegative least squares implementation as the output layer algorithm NGMDA.

For all of the experiments, the number of hidden nodes or centers for the hidden layer activation units was set to 15, and the number of trials or epochs to train the weight vector for the output layer was fixed at 1000.

The summary of the performance from the experiments is shown in Table 1 below.

[1] The UNB ISCX NSL-KDD malware dataset was obtained from http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html and a github site for the data used in this work: https://github.com/defcom17/NSL_KDD

Table 1. Misclassification errors (%) of the testing samples, Akaike Information Criterion (AIC) of the training samples, and Area Under Curve (AUC) of the testing samples on the UNB ISCX NSL-KDD binary class dataset for the proposed GMDA

Algorithms	Misclassification error in %	AIC	AUC
kmeans + LMS	**55.98**	**-222209**	**0.38**
kmeans + RLS	**69.15**	**-5621**	**0.38**
kmeans + NNLMS	**43.07**	**30.0**	**0.51**
NMF + NNLMS	**40.59**	**-10076**	**0.79**
NMF + ASCA	**33.07**	**-293620**	**0.86**

AIC was computed based on the sum of squared errors (SSE) from training the particular filter, weighted by the number of training samples (n) and penalized by the number of centers (k). It was computed as: $n\ln(SSE) - n\ln(n) + 2k$. Therefore, AIC is interpreted as a measure of model training accuracy, where a smaller value (or most negative value) represented better training accuracy.

Note that by replacing kmeans with NMF in finding the cluster centers, the performance of the GMDA improved significantly in terms of the metrics shown in Table 1 and the receiver operating characteristic (ROC) curves AUC (area under the curve). Another important item to note in Table 1 is that the AIC for kmeans + LMS seems to be acceptable. However, AIC measures the *training accuracy* only and indicates for this case that, given the input data and the algorithm choices, the algorithms were indeed trained well on the presented nonnegative data. But, the poor classification results underscore the mismatch between algorithm choice (unconstrained) and input data (nonnegative). This is also demonstrated by the kmeans + NNLMS line in Table 1 where the poorest results are shown. Another way to state this is that training the filter on results from an unconstrained dimension reduction algorithm for nonnegative data is not recommended. Both the training accuracy and ROC AUC are very poor. In fact, the AUC indicates that this combination is not much better than tossing a coin for classification.

The result suggests that NMF, with its increased interpretability, preserved the data relationships for constructing meaningful clusters that could be the basis for classification, but kmeans did not. The performance of NGMDA was improved further by running NMF with ASCA. With a dimension reduction algorithm that preserves cluster structure, such as NMF, and a nonnegative constraint on the adaptive filtering algorithm, such as NNLMS or ASCA, the performance of NGMDA in terms of classification and model training accuracy is greatly improved. NMF is again shown to be the right computational model for the data.

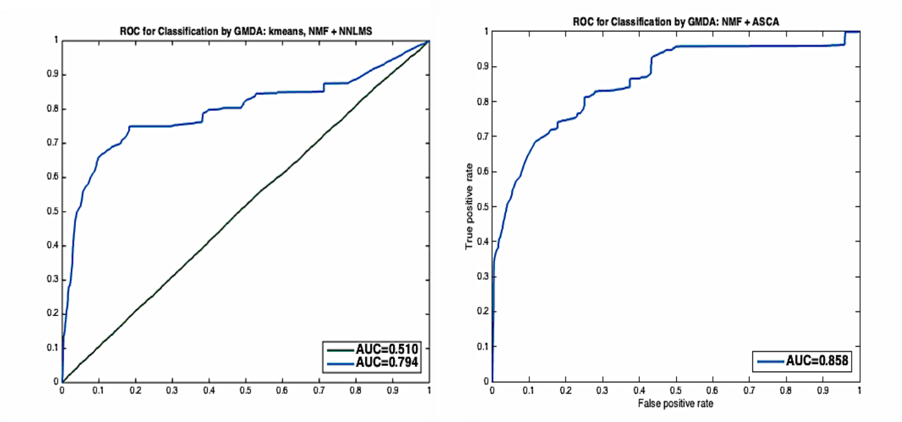


Fig. 6. a. kmeans + NNLMS (green) produced nearly a straight line, AUC = 0.51, while NMF + NNLMS (blue), AUC = 0.79; b. NMF + ASCA, AUC = 0.86 (best result)

To illustrate the performance of NGMDA for the analyses, the ROC AUC is shown in Figures 6 a and b. It is evident from these two figures that NGMDA easily outperforms the other methods tested.

5 Conclusions

In this paper, the many examples of using the correct algorithm for the data were reviewed and extended with our experimental results on malware data. We have demonstrated new methods for event detection on a variety of event scenarios. The events were discovered from various text sources for ceasefire violations in Arabic text and environmental events from Chinese text. We also discovered malware events from a KDD data set containing network intrusion data. The text-based events discovered from our topic refinement process were very specific in terms of event type, location, severity of impact, and date. The technology could be used in a number of critical situational awareness applications, especially when embedded in a visual analytics system. Our new NGMDA is promising as a classifier for various nonnegative data sources where online, streaming updates are a requirement. The single layer perceptron architecture does not require the expensive backpropagation algorithm. Data can flow through the NGMDA as each sample or mini-batch of samples is available.

The motivation for these generalizations of the original Moody-Darken architecture has been to apply that architecture to a wider class of problem domains, which may impact social situational awareness, social conflict monitoring from text corpora and other data sources, image understanding and crop monitoring, and deep feature extraction for even better classification results. Like GMDA as in reported in [3], NGMDA inherits from the original Moody-Darken architecture fast training of locally tuned receptive fields, the best approximation property, and the ability to incorporate adaptability to changing data.

The results obtained on the KDD malware data appear to be promising for applying the NGMDA in other domains. The misclassification errors were very good when the computational model consistently utilized nonnegative constraints. Our future work will focus on deep feature extraction from a hierarchical NGMDA. We currently have preliminary results on image classification, again demonstrating that the correct computational model uses nonnegative constraints.

Acknowledgments. The authors thank Dr. Richard Boyd for his significant contributions to SmallK and LEAN, both of which were essential tools used in our topic refinement process. We would also like to thank Dr. Elizabeth Whitaker for the opportunity to present this work and the Information and Communication Laboratory (ICL) at Georgia Tech Research Institute for supporting this research. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) XDATA program grant FA8750-12-2-0309. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

1. Moody, J. and Darken, C.: Fast Learning in Networks of Locally-Tuned Processing Units.: Neural Computation 1(1):281-294, (1989)
2. Lippmann, R.P.: Pattern Classification Using Neural Networks.: IEEE Communications Magazine 27(11):47-54, (1989)
3. Drake, B, Huang, T., and Cistola, C (invited paper).: Malware Detection based on New Implementations of the Moody-Darken Single-Layer Perceptron Architecture: When the data speaks, are we listening?: Proceedings of the 7th International Conference on Applied Human Factors and Ergonomics, Human Factors in Cybersecurity, Orlando, Florida, USA, July 27-31, Springer, T. Ahram, W. Karwowski (Eds.), (2016)
4. Kim J., He Y., and Park, H.: Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework.: Journal of Global Optimization, 58:285–319, (2014)
5. Haykin, S.: Adaptive Filter Theory, 5th Edition,: Pearson Education, Inc. (2014)
6. Haykin, S.: Neural Networks and Learning Machines, 3rd Edition,: Prentice Hall (2009)
7. Bishop, C.: Neural Networks for Pattern Recognition,: Oxford University Press (2005)
8. Drake, B., Luk, F., Speiser, J. M. and Symanski, J.: SLAPP: a systolic linear algebra parallel processor.: IEEE Computer, 20(7):45–49, (1987)
9. Chen, J., Richard, C., Bermudez, J. and Honeine, P.: Nonnegative least-mean-square algorithm: IEEE Transactions on Signal Processing, 59(11):5225-5235, (2011)
10. Golub, G. H. and Van Loan, C. F. Matrix Computations. The Johns Hopkins University Press, Baltimore (2013)
11. Franc, V., Hlav´ac, V., and Navara, M.: Sequential coordinate-wise algorithm for the nonnegative least squares problem. Proceedings of the 11th International Conference on Computer Analysis of Images and Patterns (CAIP), pp. 407-414 (2005)
12. Drake, B., Kim, J., Mallick, M., and Park, H.: Supervised Raman spectra estimation based on nonnegative rank deficient least squares.: Proceedings 13th International Conference on Information Fusion, Edinburgh, UK, (2010)
13. Kay, Steven M.: Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory: Prentice Hall PTR (1998)