ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Large-scale nonlinear dimensionality reduction for network intrusion detection

Yasir Hamid[1], Ludovic Journaux[2], John A. Lee[3],
Lucile Sautot[4], Bushra Nabi[5], M. Sugumaran[1]

1- Department of C.S.E., Pondicherry Engineering College, Pondicherry-14, India
2- LE2I UMR6306, CNRS, Univ. Bourgogne Franche-Comté, AgroSup Dijon, France
3- Univ. catholique de Louvain, UCL/SSS/IREC/MIRO, Bruxelles, Belgium
4- AgroParistech, UMR TETIS , Maison de la télédétection, Montpellier, France
5- Department of Botany, University of Kashmir

Abstract. Network intrusion detection (NID) is a complex classification problem. In this paper, we combine classification with recent and scalable nonlinear dimensionality reduction (NLDR) methods. Classification and DR are not necessarily adversarial, provided adequate cluster magnification occurring in NLDR methods like *t*-SNE: DR mitigates the curse of dimensionality, while cluster magnification can maintain class separability. We demonstrate experimentally the effectiveness of the approach by analyzing and comparing results on the big KDD99 dataset, using both NLDR quality assessment and classification rate for SVMs and random forests. Since data involves features of mixed types (numerical and categorical), the use of Gower's similarity coefficient as metric further improves the results over the classical similarity metric.

1 Introduction

With internet now being ubiquitous in many aspects of daily life, keeping the networks secure has become critical and requires continuous effort [15]. In this context, an Intrusion Detection System (IDS) aims to discriminate the legitimate network connections against potentially harmful ones [13].

Effective and popular processing techniques for IDS are feature selection and extraction [11]. Regarding the former, works about IDS have explored a lot of feature selection techniques [18]. For the latter, though, the well known PCA remains largely dominant [3]. PCA reduces data dimensionality based on variance preservation but this simple criterion is not guaranteed to retain relevant information in data. In an IDS, PCA generally improves the detection rate of the dominant classes, i.e., Probe and Denial of Service (DOS), but it does not impact much on the detection rate of more marginal classes like User to Root (U2R) and Remote to Local (R2L)[12]. PCA is one of the oldest linear methods of dimensionality reduction (DR) and many nonlinear methods have been developed over the past decades [8]. A usual assumption of DR methods is that the intrinsic dimensionality of data is much lower than the ambient space dimensionality, e.g., because data lives on a submanifold. Recent nonlinear DR methods have proved to be effective in many applications, like visualization and exploratory analysis [16]. Scalability issues of these methods have been addressed lately [19].

In classification tasks like IDS, preprocessing with DR remains debatable, with possible gains (mitigation of the curse of dimensionality, noise attenuation) and losses (blindness with respect to the actual supervised task, degraded separability due to collapsing classes). This paper shows that cluster magnification in some DR methods [16] lead to a positive outcome. Another challenge is to exploit both qualitative and quantitative features in the dataset, which is tackled by replacing the Euclidean distance with Gower's dissimilarity coefficient [2].

The rest of the paper is organized as follows. Section 2 presents our material and methods (DR, quality assessment of DR, and Gower's coefficient). Section 3 presents the experimental results. Finally, Section 4 draws the conclusions.

2 Materials and methods

2.1 Real-world KDD99 dataset

The KDD99 dataset is the benchmark that is predominantly used for network intrusion detection[1]. It comes with mixed-type features, some being numeric `num_failed_logins`, `num_file_creations`, `num_access_files` and others being nominal like `Protocol`, `type_of_service`. A total of $M = 41$ features are present for each record, while the 42^{nd} is the class label. The dataset includes a total of 23 attacks spanning across four groups, User to Root, Remote to Local, DoS, and Probe. In addition to four attack groups, a class is associated with normal network connections. So, let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1\leq i\leq N}$ be the M-by-N dataset. For our experiments with KDD99, a subset of $N = 61906$ records was selected, a size that raises already significant scalability issues. Moreover this subset size also allow acceptable computing time.

2.2 Gower's similarity coefficient

Gower's similarity coefficient is often used in an ecological study or in a modeling work. It is designed to measure (dis)similarity between two items or individuals that are characterized by both numerical and categorical features [2]. Gower's coefficient is defined as a weighted average of similarity scores, assigned in the pairwise comparison of items. The scores are computed differently depending on the type of the features: categorical (dichotomous, nominal, interval, or ratio-scale, ...) or numeric (real or integer quantities). Let us consider two items $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ with M features. Gower's coefficient is calculated as $\sum_{k=1}^{M} w_k S_k(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$, where w_k is a binary weight of the kth feature. The weight is one except if the kth feature value is not defined for $\boldsymbol{\xi}_i$ or $\boldsymbol{\xi}_j$; $S_k(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ then depends on type V_k of the kth feature:

$$S_k(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \left\{ \begin{array}{cl} \delta(\xi_{ik}, \xi_{jk}) & \text{if } V_k \text{ is qualitative} \\ \frac{|\xi_{ik} - \xi_{jk}|}{\max V_k - \min V_k} & \text{if } V_k \text{ is quantitative} \end{array} \right. , \tag{1}$$

where δ is Dirac's function, which is one iff $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ are from the same category.

[1] `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`

2.3 Dimensionality reduction

Dimensionality reduction is widely used for exploratory data analysis, as it can help users to visualize data in low-dimensional spaces. Many DR methods have been proposed in the literature [8], like *principal component analysis* (PCA)[5], *multidimensional scaling*[1], *Sammon's nonlinear mapping* [14], *stochastic neighbour embedding* [4], *t-distributed stochastic neighbour embedding* [16], *neighbour retrieval visualizer* [17], *Jensen-Shannon embedding*(JSE) [7], and recently *Multi-Scale Jensen-Shannon embedding* (Ms.JSE) [6]. Most of these approaches, though, scale rather poorly in a 'big data' context, namely, if N gets larger than a few thousand items, like in the KDD99 dataset. Except for PCA, these DR methods entail the computation of pairwise (dis)similarities in large N-by-N matrices, with quadratic time complexity. To overcome this limitations, pairwise relationships in SNE-like methods are aggregated and approximated with space partitioning trees [19]. In our experiments, we use in particular fast approximations of Elastic Embedding, SNE, symmetric SNE, NeRV, t-SNE, and weighted t-SNE [2]. Afterwards, we focus on the best-performing method and replace the Euclidean distance with Gower's coefficient in order to compare results.

2.4 Quality criterion to compare DR methods objectively

To compare the DR results, we use the quality assessment tools (QA) that quantify neighborhood preservation [9, 6]. If ν_i^K if the K-ary neighborhood of $\boldsymbol{\xi}_i$ and n_i^K is the K-ary neighborhood of its low-dimensional counterpart $\mathbf{x}_i$ found by DR, then $\mathrm{Q}_{\mathrm{NX}}(K) = \sum_{i=1}^{N} \frac{\nu_i^K \cap n_i^K}{KN}$ is the average neighborhood preservation. Rescaling into $\mathrm{R}_{\mathrm{NX}}(K) = \frac{(N-1)\mathrm{Q}_{\mathrm{NX}}(K)-K}{N-1-K}$ locates the considered embedding between a random one ($\mathrm{Q}_{\mathrm{NX}}(K) \approx K$, $\mathrm{R}_{\mathrm{NX}}(K) = 0$) and a perfect one ($\mathrm{Q}_{\mathrm{NX}}(K) = \mathrm{R}_{\mathrm{NX}}(K) = 1$). A synthetic scalar score over all sizes K is the AUC of $\mathrm{R}_{\mathrm{NX}}(K)$ in plot with log abscissae, namely, $\mathrm{AUC}(\mathrm{R}_{\mathrm{NX}}(K)) = \left(\sum_{K=1}^{N-2} \mathrm{R}_{\mathrm{NX}}(K)/K\right)\left(\sum_{K=1}^{N-2} 1/K\right)$. Like in DR methods, space-partitioning trees and incomplete sort (quick select) can improve scalability of DR QA.

3 Results and discussion

All DR methods listed in the previous section are applied to KDD99 to get 3D embeddings. For easier visualization, 2D embeddings corresponding to the best methods are also shown and discussed before reporting the classification results.

Quality assessment of DR. Figure 1 shows the quality curves ($\mathrm{R}_{\mathrm{NX}}(K)$) and their AUC (in the legend box) for the 3D embeddings of all considered DR methods. Looking at all curves except the one for Gower t-SNE, we conclude that t-SNE outperforms the other methods, with the tallest and widest bump between $K = 10$ and 200, as well as highest AUC. Among those methods, t-SNE is the only method using discrepant Gauss/Student neighbourhoods in the high- and low-dimensional spaces, respectively. Thereby, it induces an exponential

[2] `http://research.cs.aalto.fi/pml/software/ne/`

stretch of distances and cluster magnification [10]. The curve of Gower t-SNE is not straightforward to interpret, since QA tools do not account for the change of metric occurring with Gower's coefficient. As discussed below, its use is expected to improve classification results, since it exploits qualitative features that can increase class and cluster separability, as illustrated below.

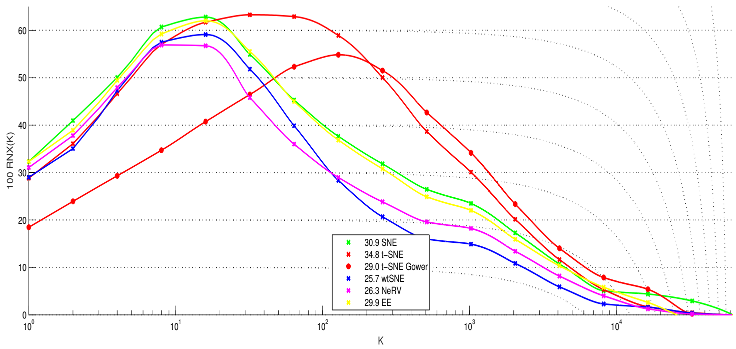
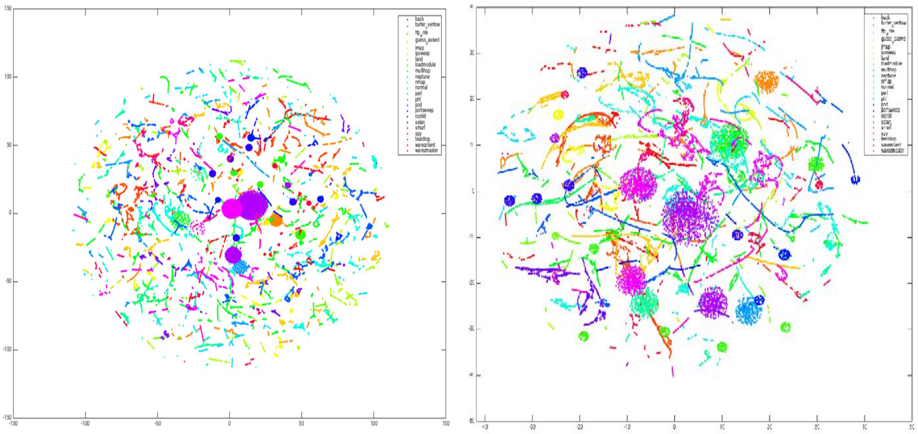


Figure 1: Quality curves ($\mathrm{R_{NX}}(K)$) and their AUC (legend) for all methods.

Best emdeddings of KDD99 with t-SNE. For visualization purposes, and since t-SNE outperforms other methods, Figure 2 shows 2D embeddings after running t-SNE with the Euclidean distance and Gower's coefficient. Some clusters seem to be monodimensional dashes, while disc-shaped ones would reflect a larger intrinsic dimensioanlity. Visual inspection shows that Gower t-SNE separates clusters even better than regular t-SNE. This is indirectly confirmed in the 3D embeddings by the improved classification results below.

Classification results. Table 1 reports the classification performance metrics of SVMs and RFs, applied to all 3D embeddings of KDD99. Due to the clear superiority of t-SNE in terms of cluster separation and longer computation times, we used Gower's similarity coefficient only with this DR method. SVMs with Gaussian kernel was tried with five different bandwidths Γ and 10-fold cross-validation. The best model was kept. A random forest was constructed as an ensemble of random trees without any restriction on the depth of the tree and a batch size set to 100. The total number of iterations was explicitly set to 100 and the results were checked for 2 decimal points. To assess classification quality, we relied on: %*Classification accuracy*, *Precision*, *Recall*, and ROC AUC.



(a) t-SNE with the Euclidean distance (b) t-SNE with Gower's similarity coefficient

Figure 2: Scatter plots of the t-SNE embeddings, without and with Gower.

Table 1: Classification performance metrics

	% Classification		Precision		Recall		ROC Area		M.A.Error		RMSE	
RD Methods	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**
Normal Data	96.97	97.81	0.97	0.97	0.97	0.97	0.98	0.98	0.0012	0.0019	0.0319	0.0089
EE	98.92	98.55	0.98	0.98	0.98	0.98	0.97	0.99	0.0009	0.0055	0.0306	0.0379
SNE	99.17	98.81	0.99	0.98	0.99	0.98	0.99	0.99	0.0013	0.0058	0.3621	0.0366
NerV	97.81	97.95	0.97	0.97	0.97	0.97	0.98	0.998	0.0019	0.0089	0.0436	0.0496
***t*-SNE**	99.60	99.54	0.99	0.99	0.99	0.98	0.99	0.99	0.0004	0.0005	0.0198	0.0169
WTSNE	98.87	99.54	0.98	0.99	0.98	0.99	0.99	0.99	0.0015	0.0009	0.3116	0.0184
***t*-SNE with Gower**	**99.96**	**99.97**	**1**	**1**	**1**	**1**	**0.99**	**1**	**0**	**0.0002**	**0.0051**	**0.0056**

4 Conclusion

In the context of network intrusion detection, this paper proposes a methodology in two steps. First, data are embedded in a lower-dimensional space. In the case of KDD99, a scalable method of dimensionality reduction like Barnes-Hut t-SNE mitigates the curse of dimensionality and magnifies cluster separation (as well as class separation, indirectly), provided the target dimensionality is not too low, in order to prevent classes to collapse on each other. To extract maximal information from the KDD99 mixed-type features, we also combined t-SNE with Gower's similarity coefficient. Experimental results demonstrate the effectiveness of the proposed methodology.

Challenges for the future are: getting even better scalability, in order to process the entire KDD99 set; making the methodology fully parametric (the DR method and Gower's coefficient have no straightforward out-of-sample extension); and applying the proposed methodology to other datasets in order to generalize and develop a new approach of mixed data analysis.

References

[1] T.F. Cox and M.A. Cox. *Multidimensional scaling*. CRC press, 2000.

[2] J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, December 1971.

[3] F.E. Heba, A. Darwish, A.E. Hassanien, and A. Abraham. Principle components analysis and support vector machine based intrusion detection system. In *2010 10th International Conference on Intelligent Systems Design and Applications*, pages 363–367. IEEE, 2010.

[4] G.E. Hinton and S.T. Roweis. *Advances in neural information processing systems*, chapter Stochastic neighbor embedding, pages 833–840. 2002.

[5] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[6] J.A. Lee, D.H. Peluffo-Ordóñez, and M. Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[7] J.A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[8] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[9] J.A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing*, 72:1431–1443, 2009.

[10] J.A. Lee and M. Verleysen. On the role and impact of the metaparameters in t-distributed stochastic neighbor embedding. In Y. Lechevallier and G. Saporta, editors, *Proc. 19th COMPSTAT*, pages 337–348. Paris (France), August 2010.

[11] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36(1):16–24, 2013.

[12] W.-C. Lin, S.-W. Ke, and C.-F. Tsai. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-based systems*, 78:13–21, 2015.

[13] C.H. Rowland. Intrusion detection system, June 11 2002. US Patent 6,405,318.

[14] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.

[15] W. Stallings. *Network security essentials: applications and standards*. Pearson Education India, 2007.

[16] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[17] J. Venna, J. Peltonen, and K. et al. Nybo. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[18] C. Xiang, P.C. Yong, and L.S. Meng. Design of multiple-level hybrid classifier for intrusion detection system using bayesian clustering and decision trees. *Pattern Recognition Letters*, 29(7):918–924, 2008.

[19] Z. Yang, J. Peltonen, and S. Kaski. Optimization equivalence of divergences improves neighbor embedding. In *ICML*, pages 460–468, 2014.