Hybrid Subspace Mixture Models for Prediction and Anomaly Detection in High Dimensions

Jenn-Bing Ong(✉) and Wee-Keong Ng

School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
{ongj0063,awkng}@ntu.edu.sg

Abstract. Robust learning of mixture models in high dimensions remains an open challenge and especially so in current big data era. This paper investigates twelve variants of hybrid mixture models that combine the G-means clustering, Gaussian, and Student t-distribution mixture models for high-dimensional predictive modeling and anomaly detection. High-dimensional data is first reduced to lower-dimensional subspace using whitened principal component analysis. For real-time data processing in batch mode, a technique based on Gram-Schmidt orthogonalization process is proposed and demonstrated to update the reduced dimensions to remain relevant in fulfilling the task objectives. In addition, a model-adaptation technique is proposed and demonstrated for big data incremental learning by statistically matching the mixture components' mean and variance vectors; the adapted parameters are computed based on weighted average that takes into account the sample size of new and older statistics with a parameter to scale down the influence of older statistics in each iterative computation. The hybrid models' performance are evaluated using simulation and empirical studies. Results show that simple hybrid models without the Expectation-Maximization training step can achieve equally high performance in high dimensions that is comparable to the more sophisticated models. For unsupervised anomaly detection, the hybrid models achieve detection rate $\gtrsim 90\%$ with injected anomalies from 1% to 60% using the KDD Cup 1999 network intrusion dataset.

Keywords: Mixture models · Coarse filtering · Model adaptation · Parameter rating · Dimensionality reduction · Incremental learning · Diffusion map

1 Introduction

Mixture models have been widely used in many applications such as speaker verification, background subtraction for real-time tracking, and biological applications. Efficient algorithm to learn mixture of Gaussians in high dimensions with small error bound has recently been demonstrated [6]. However, practical algorithms for robust and adaptive learning of high-dimensional mixture models

G. Cong et al. (Eds.): ADMA 2017, LNAI 10604, pp. 326–339, 2017.
https://doi.org/10.1007/978-3-319-69179-4_23

is still an open challenge [13]. This paper extends the work of a robust subspace mixture model initially developed by [2] for anomaly detection to predictive modeling in high dimensions. High-dimensional data is reduced to lower dimensional subspace using whitened principal component analysis. Diffusion Map (DM)-based coarse-filtering technique is robust to noise perturbation [5] and Student-t distribution Mixture Model (SMM) is robust to outliers [10], both provide a robust statistics of the model developed by [2]. The estimated SMM parameters are then used to form a Gaussian Mixture Model (GMM) statistics for predictive density estimation to ensure robustness and sensitivity to outliers. This paper aims to further investigate and improve the model performance and computing efficiency for high-dimensional predictive modeling and anomaly detection. The contributions of this paper are as follow[1].

- Twelve variants of hybrid mixture models that combine G-means clustering or K-Means clustering using Gaussian algorithm (KM) developed by [7], GMM, and SMM have been compared for predictive modeling and anomaly detection. Results show that simple hybrid models without the Expectation-Maximization (EM) step can achieve equally high prediction accuracy and anomaly detection rates comparable to the sophisticated models.
- For unsupervised anomaly detection, the noise can be removed by a DM-based coarse-filtering technique developed by [2]. However, without the coarse filtering, the hybrid models achieve detection rate $\gtrsim 90\%$ with injected anomalies from 1% to 60% in the KDD Cup 1999 network intrusion dataset. The top-down approach produces results that do not fluctuate with data sampling in contrast to the models using the DM-based coarse-filtering technique that process data in smaller chunks.
- For real-time batch data processing, a technique based on Gram-Schmidt orthogonalization process is proposed and demonstrated to update the reduced dimensions to remain relevant in fulfilling the task objectives. Existing work usually assumes same reduced dimensions for each batch of data or assumes spherical-Gaussian covariance so that the covariance remains conserved after re-projection from one set of dimension vectors to another.
- A model-adaptation technique is proposed and tested for incremental learning of GMM and SMM model parameters. This technique is different from previous [2,11,12] in that the adapted model parameters are computed by taking account the data size of new and older statistics, and a parameter is introduced in the technique to scale down the influence of older statistics in each iterative computation.
- Application of the parameter rating technique developed by [2] is demonstrated using KDD Cup 1999 network intrusion dataset (10% subset); the parameter ratings may be used to suggest mitigating actions for the different intrusion types or to label the data.

The organization of this paper is as follows. Data pre-processing for dimensionality reduction and coarse filtering are provided in Sect. 2. The data processing

[1] For reproducibility, the Matlab scripts to run the simulation and experimental studies in this paper are obtainable from https://github.com/jennbing/hybrid-models.

to form the hybrid mixture models, model adaptation, and the parameter rating techniques are covered in Sect. 3. Section 4 evaluates the model performance by simulation/experimental studies and Sect. 5 concludes the work.

2 Data Pre-processing

2.1 Dimensionality Reduction Using Whitened PCA

Principal component analysis (PCA) is a linear mapping from a high-dimensional space to a subspace that captures the most variability in the data specified by a set of orthogonal/principal components (PCs). To extract the relevant components from different datasets, different number of PCs with the highest eigenvalues are tested to find the minimum required for better model performance. For batch processing, it is important to ensure that these minimum number of PCs are sufficient for each batch of data to fulfill particular objective; e.g., predictive modeling or anomaly detection. Suppose there exists additional PCs in new batch of data which are not spanned by the older set of PCs, the new dimensions can be appended by using the Gram-Schmidt orthogonalization process to remove the projections on older set of PCs. On the other hand, older PC may be discarded if the absolute value of the Pearson correlation with the set of new PCs is low (<0.5). The threshold can be determined from empirical experiments to ensure good model performance. The reason this updating technique is proposed because the projection of non-spherical Gaussian covariance from a set of orthonormal vectors to another is not conserved, therefore the PCs can only be appended or discarded, but not re-projected, during the updating process.

2.2 Coarse Filtering Using Technique Based on Diffusion Map

Diffusion Map (DM) is a non-linear technique that helps to discover the underlying manifold of high-dimensional data [5]. Given a set of d-dimensional data X, the similarity measure between two data points is defined as

$$\chi(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{\epsilon}\right) \tag{1}$$

where $||\bullet||$ is the Euclidean distance of the vectors in the ambient space $\mathbb{R}^d$. The scaling parameter is computed by the average smallest neighbouring distance [9]. The transition probability between two data points can be computed by normalizing Eq. 1. The diffusion distance is small if there are many short paths connecting two data points, which implies large transition probability between the two points. DM provides a representation in which the data points are clustered according to their connectivity, which is robust to noise perturbations [5]. In the diffusion space, inliers are expected to be clustered together, whereas outliers might be spread between several small clusters or scattered randomly. The inliers can then be identified by discovering the biggest connected component in the diffusion space using technique proposed in [2].

3 Hybrid Mixture Models

This section explains the methods to estimate the hybrid models' parameters, some of the algorithms can be found in [2]. Table 1 tabulates the sequence of data processing of the hybrid mixture models. The acronym for each hybrid model follows the sequence of data processing. For example, the sequence of KEG model is (1) KM (2) EM (3) GMM. For anomaly detection, the data is first coarse-filtered with a technique based on diffusion map described in Sect. 2.2 to remove anomalies before model training. GMM and SMM parameters are estimated using the EM algorithms described in [3] and [10] respectively. The EM initialization is provided by the K-means clustering using Gaussian algorithm (KM) developed by [7] that repeatedly splits every clusters until each approximates the Gaussian distribution statistically. The statistical test, which is based on the one-dimensional Anderson-Darling statistics, is valid for multi-dimensional Gaussian distribution. In addition, the mixture model parameters can also be learnt using KM directly and is theoretically shown to require near-optimal sample requirement with well-separated mixture components [4]. The reason this simple model is explored because EM algorithm often converges to local minimum in high dimensions. For computing efficiency, variance vectors are used in the mixture modeling instead of full covariance matrices. For KESG, the estimated SMM parameters are used to form GMM statistics; while for KEGS, the estimated GMM parameters form the SMM statistics assuming the degree of freedom is 1, this is justifiable because real data usually spreads out. Similar applies to other models. In addition, a model-adaptation technique is introduced for incremental learning of big data and the computation of a parameter rating

Table 1. Sequence of data processing of the proposed hybrid mixture models. The acronym for each hybrid model follows the sequence of data processing. For example, the sequence for DKEGS model is (1) DM (2) KM (3) EM (4) GMM (5) SMM.

Models	DM	KM	EM	GMM	SMM	Remark
KG		1		2		Prediction
KS		1			2	
KEG		1	2	3		
KES		1	2		3	
KESG		1	2	4	3	
KEGS		1	2	3	4	
DKG	1	2		3		Anomaly detection
DKS	1	2			3	
DKEG	1	2	3	4		
DKES	1	2	3		4	
DKESG	1	2	3	5	4	
DKEGS	1	2	3	4	5	

technique developed by [2] is presented here in order to examine the source of anomalies occurrences in the original feature space.

3.1 Model Adaptation

Box's M test is used to statistically compare the sample variance from new and older statistics. After finding a match of a pair of mixture components' variance, Hotelling's T^2 test is used to compare and match the corresponding sample mean. The adapted parameters are estimated using Maximum A-Posteriori (MAP) estimation. Similar model adaptation technique has been developed for GMM by [12]. Our proposed technique differs from previous [2,11,12] in that the adapted parameters are computed by taking account the sample size of new and older statistics, and a parameter $f^{\rho}(\mathbf{c})$ is introduced in the technique to scale down the influence of older statistics in the iterative computation. Equation 2 summarizes the model adaptation for both GMM and SMM. Although Box's M test and Hotelling's T^2 test assume the pair of mixture components are multidimensional Gaussian-distributed, short of other alternatives, these statistical tests provide a more stringent criteria for matching SMM mixture components. Additionally, the EM algorithm to estimate the mixture model parameters does not guarantee to find a global optimum since the problem is non-convex and the final solutions depend on the initial parameter values. Therefore, a technique to combine the statistics of a mixture of parametric models for predictive density estimation is proposed in [2]. The technique can be easily parallelized in the expense of computational resources due to model independence [2]. Our model-adaptation technique can also be used to merge the model parameters estimated from multiple trial estimation on a given dataset, this saves the memory space from storing duplicate parameters from different trials.

$$\begin{aligned} Mixing\ coefficient &: \tilde{\omega}_i = \left(\alpha_i^{\omega}\omega_i^{new} + (1-\alpha_i^{\omega})\omega_i\right)\gamma \\ Sample\ mean &: \tilde{\mu}_i = \alpha_i^{\mu}\mu_i^{new} + (1-\alpha_i^{\mu})\mu_i \\ Sample\ variance &: \tilde{\sigma}_i^2 = \alpha_i^{\sigma}\left((\sigma_i^{new})^2 + (\mu_i^{new})^2\right) + (1-\alpha_i^{\sigma})(\sigma_i^2+\mu_i^2) - \tilde{\mu}^2 \\ Degree\ of\ freedom &: \tilde{v}_i = \alpha_i^{v}v_i^{new} + (1-\alpha_i^{v})v_i \end{aligned} \tag{2}$$

where γ is a normalization factor which ensures the adapted weights sum to unity, α_i^{ρ} is the data-dependent adaptation coefficient that are computed by $\alpha_i^{\rho} = \frac{n_i^{new}}{n_i^{new}+f^{\rho}(\mathbf{c})n_i}$, where $\rho \in \{\omega, \mu, \sigma, v\}$, $n_i^{new} = N^{new}\omega_i^{new}$ and $n_i = N\omega_i$ is the sample size estimates of the i-th mixture component, $f^{\rho}(\mathbf{c})$ is a function of the context ranges from 0 to 1 that characterizes the decay of the influence of older statistics in the iterative computation. The sample mean and variance vectors have been matched statistically between a pair of mixture components before adaptation, therefore $f^{\rho}(\mathbf{c})$ has a larger impact on the mixing coefficients and degree of freedom than the sample mean and variance. Additionally, the weights of unmatched components may be scaled down appropriately, one way to do this is by applying the normalization factor on the unmatched components but keeping the weights of the matched components unchanged.

3.2 Parameter Rating

To understand the source of anomaly occurrences in the original feature space, a technique was developed by [2] for parameter rating from the learnt subspace. An anomaly is detected when its logarithmic probability is extremely low. Let z_a be the observed anomaly in the projected subspace span by the PCs, the associated mixture component of the anomaly is given by

$$i^* \triangleq \underset{i}{\operatorname{argmax}} \{q_i^{\tilde{K}}(z_a) | 1 \leqslant i \leqslant M\} \tag{3}$$

$$q_i^{\tilde{K}}(z_j) = \frac{\omega_i N(z_j; \mu_i, \Sigma_i)}{\sum_{\tilde{k}=1}^{\tilde{K}} \omega_i N(z_{\tilde{k}}; \mu_i, \Sigma_i)} \tag{4}$$

where $N(z_{\tilde{k}}; \mu_i, \Sigma_i)$ is the probability density of a Gaussian mixture component, M is the number of mixture components, and $\tilde{K}$ is the number of samples. The explanatory vector, which represents the parameters that account for the anomaly, is computed by

$$\bar{x}_a = \phi_{i^*}(x_a) \triangleq \sigma_{q_{i^*}^{\tilde{K}}}^{-\frac{1}{2}}[S] \left| x_a - E_{q_{i^*}^{\tilde{K}}}[S] \right| \tag{5}$$

$$\begin{aligned} E_{q_{i^*}^{\tilde{K}}}[S] &= \sum_{\tilde{k}=1}^{\tilde{K}} q_{i^*}^{\tilde{K}}(z_{\tilde{k}}) x_{\tilde{k}} \\ \sigma_{q_{i^*}^{\tilde{K}}}[S] &= \sum_{\tilde{k}=1}^{\tilde{K}} q_{i^*}^{\tilde{K}}(z_{\tilde{k}})(x_{\tilde{k}} - E_{q_{i^*}^{\tilde{K}}})^2 \end{aligned} \tag{6}$$

$\bar{x}_a$ represents the scaled geometric difference vector between the anomaly and the sample mean associated with the mixture component i^*. The parameters are rated by their responsibility for the anomaly occurrence by sorting the entries in $\bar{x}_a$ in a descending order. In cases that a low confidence in the responsibility of a specific mixture component for an observed anomaly, Eq. 7 presents a soft parameter rating technique proposed by [2] that takes into account the deviation from all the mixture components.

$$\begin{aligned} q_i^M(z_j) &= \frac{\omega_i N(z_j; \mu_i, \Sigma_i)}{\sum_{m=1}^{M} \omega_m N(z_j; \mu_m, \Sigma_m)} \\ \bar{x}_a &= \mathbf{E}_q[\phi(x_a)] = \sum_{i=1}^{M} q_i^M(z_a)(\phi_i(x_a)) \end{aligned} \tag{7}$$

Both soft and hard parameter rating techniques can be applied when the SMM statistics is used, in this case, $N(z_{\tilde{k}}; \mu_i, \Sigma_i)$ should be replaced by SMM distribution. The soft parameter rating technique (Eq. 7) is more computing-efficient

than the hard one (Eq. 6) because there is no need to search for the associated mixture component for each anomaly. Notice that Eq. 4 is modified from Eq. 3 in [2]; the summation in the denominator is over the sample points instead of the mixture components as in [2], this makes more sense in computing the mean and variance in Eq. 6.

4 Experimental Evaluation

4.1 Simulation Studies

Figure 1(a) shows two simulated multidimensional Gaussian-distributed centers with white noise added. The noise constitutes one-third of the sample size. Two hybrid models for anomaly detection (see Table 1) are used to remove the white noise, the models differ in sophistication and therefore computing efficiency. Although KG is less sophisticated and hence more computing-efficient compared to DKESG, both models perform equally well in removing the white noise with appropriate logarithmic-probability threshold to identify the outliers. It will be

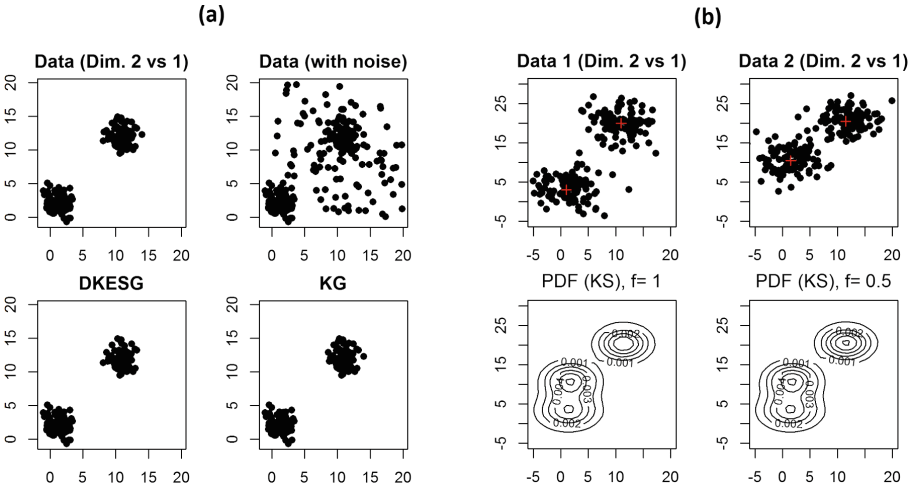

Fig. 1. **(a)** Top: Simulated multidimensional Gaussian-distributed data with two mixture components (left) and injected white noise (right). The first and second dimensions are plotted here and the distribution centers are $\mu_1 = (1, 2, ..., 10)$ and $\mu_2 = (11, 12, ..., 20)$ respectively with variance $\sigma^2 = (1, 1.5, ..., 5.5)$. The white noise comes from a uniform distribution within the range 0 to 20 in all dimensions. Bottom: Two hybrid models are used to remove the noise, DKESG is more sophisticated and hence less computing-efficient compared to KG but both models perform equally well in removing the noise. **(b)** Top: Two datasets with a common but slightly shifted multidimensional Gaussian-distributed center. The distribution centers are marked as red cross. Bottom: The predictive density of KS model after adaptation of the two datasets with the decay of influence of the older statistics, $f^{\rho}(\mathbf{c})$ set as 1 and 0.5 respectively.

shown later using empirical data that even without the EM step, simple hybrid model like KS shows high performance in anomaly detection. The model adaptation is demonstrated in Fig. 1(b) with two datasets sharing a common but slightly deviated multidimensional Gaussian-distributed center between the two datasets; the common distribution centers are $(11, 20)$ and $(11.5, 20.5)$ respectively. Each dataset also includes another non-colocated centers, which are $(1, 3)$ and $(1.5, 10.5)$ respectively. The standard deviation of all the distributions are set to 3. With high confidence level ($p \leqslant 0.0001$) during the matching of model mean and variance vectors using statistical methods, the results show that the proposed model-adaptation technique in Sect. 3.1 provides reasonable predictive density estimation with slight variation near the center of the common distribution between using parameter $f^{\rho}(\mathbf{c}) = 1$ and 0.5 to scale down the older statistics in model adaptation (see Sect. 3.1).

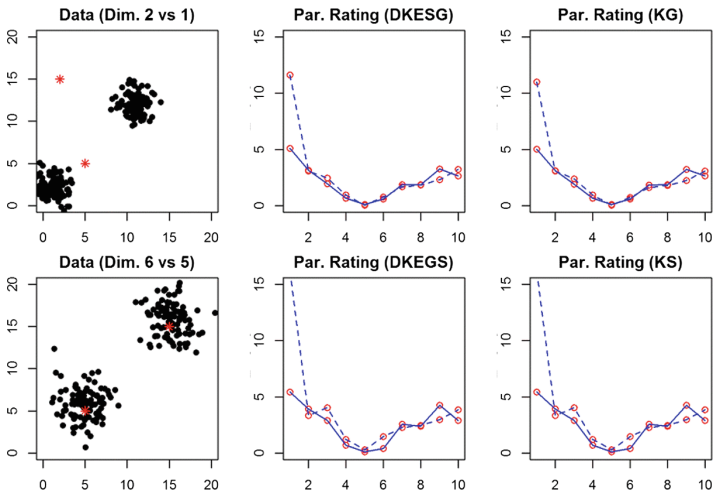

Fig. 2. First column from left: The first and second dimensions (top) and fifth and sixth dimensions (bottom) of simulated 10 dimensional data with two multidimensional Gaussian-distributed centers and two anomalies marked as red star. The two anomalies are $x_1 = (5, 5, ..., 5)$ and $x_2 = (2, 15, 15, ..., 15)$ respectively, which overlap with the distributions at about the 5th dimension but deviate at other dimensions. Remaining plots are the parameter rating using selected hybrid models. The blue lines correspond to the soft parameter rating using Eq. 7 and red circles are hard parameter rating using Eq. 6. (Color figure online)

Figure 2 simulates the data with the same multidimensional Gaussian distributions as in Fig. 1(a) top left plot, but with two anomalies inserted into the dataset to evaluate the parameter rating technique. All the selected hybrid models perform equally well by showing high parameter rating at the dimensions

where deviation from the Gaussian distributions occur. However, large sample size is required to form a reliable judgement of the parameter ratings because the EM algorithm is sensitive to initial parameter values and better statistics with higher confidence level can be obtained with larger sample size.

4.2 Empirical Studies

Table 2 tabulates the prediction accuracy of the hybrid mixture models on five popular datasets obtained from UCI Machine Learning Repository [1]. Training and testing on Adult dataset are conducted on two different given datasets (train: 32561, test: 16281), prediction accuracy on other datasets are computed using 10-fold cross validation on a single dataset. The highest prediction accuracy recorded in the repository are Adult (Forward Sequential Selection Naive-Bayes: 85.95%), Wine (Regularized Discriminant Analysis: 100%), and Breast Cancer (separating plane: 97.5%). For KDD Cup 1999 dataset, different prediction accuracy for different network intrusion types were reported using genetic algorithm [8]; i.e., normal (69.5%), probe (71.1%), denial of service (99.4%), user to root attacks (18.9%), and remote to user attacks (5.4%). On average, 88.2% detection rate is achieved in [8]. Overall, the proposed hybrid models perform reasonably well compared to other techniques especially in high-dimensional regime. It is also observed that without the EM step, KG and KS perform equally well compared to the sophisticated models for predictive modeling. To show that the algorithm is scalable, the full KDD Cup 1999 dataset with 41 attributes and close to 5 million instances is used to train and test the hybrid models for large-scale prediction in batch mode of 10^5 instances. The mixture components are adapted using the proposed model-adaptation technique described in Sect. 3.1. However, the variance vectors were not matched here because they are several orders of magnitude smaller than the mean and fluctuate wildly, i.e., only the mean vectors are matched in the adaptation process. The results are shown in Fig. 3, the highest prediction accuracy is observed when all the PCs are used. The EM algorithm to estimate GMM converges even with reduced dimensions $\gtrsim 40$ and produce higher prediction accuracy compared to the ones without the EM step (compare KEG and KEGS to KG and KS).

Table 2. Prediction accuracy of the hybrid mixture models using different datasets obtained from UCI Machine Learning Repository [1].

Dataset	KG	KS	KEG	KES	KEGS	KESG
Iris (Instances: 150, Attributes: 4)	0.97	0.97	0.97	0.97	0.97	0.97
Wine (178, 13)	0.94	0.94	0.94	0.93	0.94	0.93
Adult (48842, 14)	0.81	0.81	0.81	0.80	0.81	0.80
Wisconsin Diagnostic Breast Cancer (569, 32)	0.96	0.96	0.96	0.94	0.96	0.94
KDD Cup 1999 (10% subset: 494020, 41)	0.90	0.90	0.86	N/A	0.86	N/A

The prediction accuracy fluctuates with the data sampling, this is likely a characteristic of the dataset which contains both predictable and less-predictable events. New PCs not spanned by the older set of PCs are appended using the Gram-Schmidt orthogonalization process described in Sect. 2.1, the unused PCs are not discarded in the updating process. It is observed that the SMM-based hybrid models (KS, KES, and KESG) do not perform well in reduced dimensions $\gtrsim 30$.

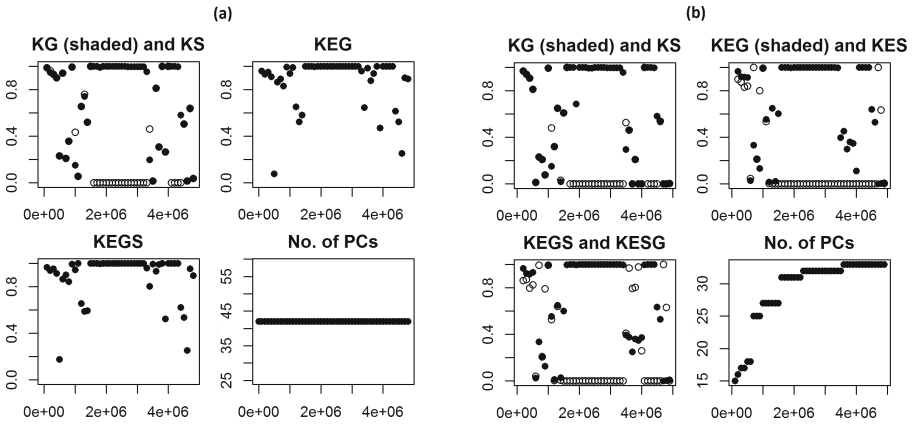

Fig. 3. Prediction accuracy of the hybrid models for batches of 10^5 instances using KDD Cup 1999 network intrusion dataset **(a)** with same number of PCs and **(b)** changing number of PCs for each batch of data.

Table 3 tabulates the detection rate and false positive rate of the hybrid mixture models using the KDD Cup 1999 dataset (10% subset). Normal-type network data of 95000 instances are extracted from the dataset and injected with different percentages of injected anomalies. The number of PCs required for anomaly detection is lesser compared to predictive modeling, in particular, only seven PCs were used here. The data is coarse-filtered with the technique based on diffusion map described in Sect. 2.2 to remove the anomalies before model training. The logarithmic-probability threshold for anomaly detection is set as the percentile of injected anomalies. It is observed that $\gtrsim$80% detection rate is possible with 1% to 60% injected anomalies with coarse-filtering technique based on DM. Higher percentage of injected anomalies biases the training model and lower percentage increases the false positive rate, hence present different challenges to unsupervised anomaly detection. However, the detection rates fluctuate with the data-sampling process because the DM-based coarse-filtering technique processes limited amount of data points at one time due to the need to compute the similarity distance between each pair of data points (see Sect. 2.2).

Table 3. Model performance in anomaly detection using KDD Cup 1999 computer network intrusion dataset (10% subset) with different percentages of injected anomalies. The dataset contains "normal" and "attack" data. The "normal" data is first extracted from the dataset and "attack" data is then artificially injected. The percentages of injected anomalies are calculated based on the ratio of artificially injected "intrusion" data into the extracted "normal" data.

Anomalies		KS	DKG	DKS	DKEG	DKES	DKEGS	DKESG
60%	DR	0.93	**0.94**	0.77	**0.94**	0.47	0.77	0.46
	FP	0.10	0.086	0.35	0.086	0.80	0.35	0.81
50%	DR	0.93	**0.96**	0.95	**0.96**	0.50	0.95	0.51
	FP	0.073	0.045	0.048	0.045	0.50	0.048	0.49
40%	DR	0.93	0.93	**0.94**	0.93	0.84	**0.94**	0.80
	FP	0.047	0.047	0.041	0.047	0.11	0.041	0.13
30%	DR	**0.90**	0.84	0.82	0.85	0.78	0.84	0.72
	FP	0.043	0.067	0.075	0.062	0.094	0.067	0.12
20%	DR	**0.92**	0.85	0.85	0.85	0.32	0.85	0.36
	FP	0.020	0.037	0.037	0.037	0.17	0.037	0.16
10%	DR	**0.93**	**0.93**	**0.93**	**0.93**	0.11	**0.93**	0.90
	FP	0.0079	0.0073	0.0079	0.0073	0.017	0.0079	0.011
5%	DR	**0.91**	0.81	0.80	0.81	0.16	0.80	0.17
	FP	0.0050	0.0099	0.011	0.0099	0.044	0.011	0.014
1%	DR	0.89	**0.92**	**0.92**	**0.92**	0.00	**0.92**	0.00
	FP	0.0012	0.00085	0.00075	0.00085	0.010	0.00075	0.010

Without coarse filtering, KS detection rate achieves $\gtrsim$90% and because of the top-down approach, the measured detection rates are robust to data sampling process.

Figure 4 shows the soft parameter rating for different network intrusion types using KDD Cup 1999 dataset (10% subset). The model is trained with normal-type network data and the KS statistics is used to compute the parameter rating. This is because KG statistics is too sparse due to the rapidly-decaying GMM tail distribution. For large number of anomalies, the soft parameter rating technique (Eq. 7) is more computing-efficient than the hard one (Eq. 6) because there is no need to search for the associated mixture component for each anomaly. Results show that there is overlap between the sources of anomaly occurrences from different network intrusion types; the parameter rating may be used to suggest mitigating actions for each intrusion type.

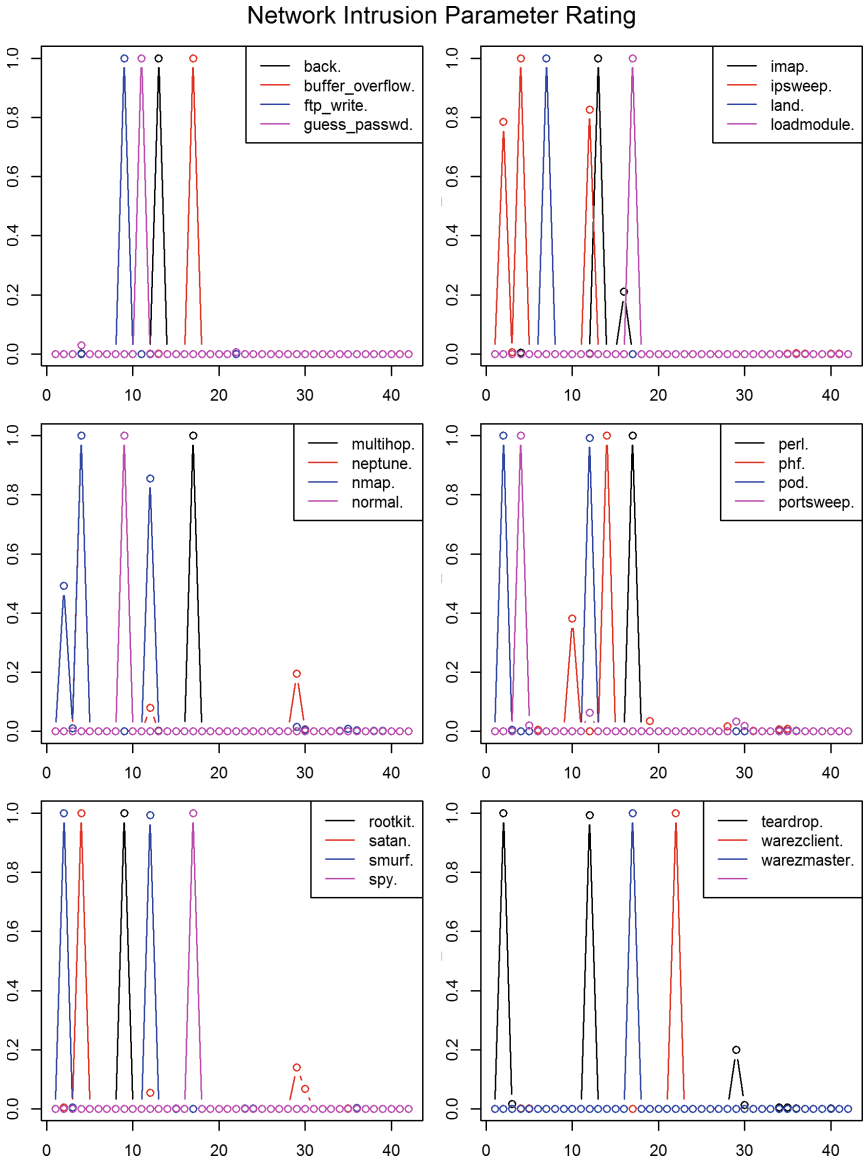

Fig. 4. Parameter rating of each network intrusion type of KDD Cup 1999 dataset (10% subset). The ratings are normalized to the range $[0, 1]$. Results show that majority of the sources of anomaly occurrences come from dimensions $\lesssim 20$ and may be used to suggest mitigating actions.

5 Conclusion

Twelve variants of hybrid mixture models have been assessed in terms of model performance and computing efficiency. In particular, KG and KS hybrid models are recommended for high-dimensional predictive modeling and anomaly detection respectively. This has implication for big-data applications because the EM algorithm may not be required to estimate mixture model parameters for high-dimensional data, which saves the computing cost. However, it is also found that whitened PCA reduces the dimensions and scales the subspace to a smaller one, which allows the EM algorithm to converge even with reduced dimensions $\gtrsim 10$. For real-time batch data processing, the proposed PC-updating technique based on Gram-Schmidt orthogonalization process is demonstrated; this technique can be used even if new dimensions are added in the original feature space. KS statistics is used to compute the parameter rating because GMM statistics is too sparse due to the rapidly-decaying tail distribution. Soft parameter rating is more computing-efficient than the hard one because there is no need to search for the associated mixture component for each anomaly. For anomaly detection, the detection rates measured from a bottom-up approach using DM-based coarse-filtering technique to remove the anomalies tend to fluctuate with the data sampling process. On the other hand, a top-down approach using KS statistics is demonstrated to achieve $\gtrsim 90\%$ detection rate from 1% to 60% of injected anomalies in the KDD Cup 1999 network intrusion dataset and this approach is more robust to the data sampling process.

References

1. Bache, K., Lichman, M.: UCI Machine Learning Repository. University of California Irvine School of Information (2013). http://www.ics.uci.edu/mlearn/MLRepository.html
2. Barkan, O., Averbuch, A.: Robust mixture models for anomaly detection. In: IEEE International Workshop on Machine Learning for Signal Processing (2016)
3. Bishop, C.M.: Pattern recognition and machine learning. Pattern Recogn. **4**(4), 738 (2006)
4. Chaudhuri, K., Dasgupta, S., Vattani, A.: Learning mixtures of Gaussians using the k-means algorithm, pp. 1–22 (2009). arXiv preprint arXiv:0912.0086
5. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmonic Anal. **21**(1), 5–30 (2006)
6. Ge, R., Huang, Q., Kakade, S.M.: Learning mixtures of Gaussians in high dimensions. In: STOC 2015 (2015)
7. Hamerly, G., Elkan, C.: Learning the k in k-means. In: Neural Information Processing Systems, pp. 281–288 (2003)
8. Hoque, M.S., Mukit, M.A., Bikas, M.A.N., Sazzadul Hoque, M.: An implementation of intrusion detection system using genetic algorithm. Int. J. Netw. Secur. Appl. **4**(2), 109–120 (2012)
9. Lafon, S.: Diffusion maps and geometric harmonics. Ph.D. thesis, Yale University, U.S.A, p. 97 (2004)

10. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Stat. Comput. **10**(4), 339–348 (2000)
11. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Proc. **10**(1–3), 19–41 (2000)
12. Song, M., Wang, H.: Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering. Intell. Comput. Theory Appl. **5803**, 174–183 (2005)
13. Vempala, S.S.: Technical perspective modeling high-dimensional data. Commun. ACM **55**(2), 112 (2012)