Survey on Big Data Analysis Algorithms for Network Security Measurement

Hanlu Chen[1,2], Yulong Fu[2], and Zheng Yan[1,2,3](✉)

[1] State Key Lab of Integrated Services Networks, Xidian University, Xi'an, China
l286187488@qq.com, zyan@xidian.edu.cn
[2] School of Cyber Engineering, Xidian University, Xi'an, China
ylfu@xidian.edu.cn
[3] Department of Communications and Networking, Aalto University, Espoo, Finland

Abstract. With the development of network technologies such as IoTs, D2D and SDN/NFV, etc., convenient network connections with various networks have stepped into our social life, and make the Cyber Space become a fundamental infrastructure of the modern society. The crucial importance of network security has raised the requirement of security measurement on a heterogeneous networking system. However, the research on this topic is still in its infancy. According to the existing security evaluation schemes of intrusion and malware detection, we believe the network data related to security should be the key for effective network security measurement. A study of the algorithms in terms of data analysis for Data Dimension Reduction, Data Classification and Data Composition becomes essential and urgent for achieving the goal of network security measurement. In this paper, we focus on the problem of big data analysis methods for security measurement, and mainly investigate the existing algorithms in different processes of big data analysis. We also evaluate the existing methods in terms of accuracy, validity and their support on security related data analysis. Through survey, we indicate open issues and propose future research trends in the field of network security measurement.

Keywords: Data classification · Data dimension reduction · Network security measurement

1 Introduction

Nowadays, communication networks and social networks are becoming an indispensable part of our life, many areas like bioinformatics, medicine, education, agricultural, traffic management, and government departments are currently relying on these networks, and make the amount of network users increasing rapidly. People are getting more and more inseparable with networks. In this situation, two problems arouse our attention. Firstly, network attacks are emerging with the increasing amount of the network users. They may cause such network security threats as information disclosure, information fraud, network paralysis, and property damage. When using a networking service, users want to know its security level in order to avoid potential

Z. Yan et al. (Eds.): NSS 2017, LNCS 10394, pp. 128–142, 2017.
DOI: 10.1007/978-3-319-64701-2_10

loss. Secondly, when security incidences occur in a given network, a timely response mechanism for security threats requires quick measurement on network security. So a data trace back processing is required by network operators to ensure a secure networking service. Motivated by both network users and operators, it becomes essential to measure network security in an efficient and effective manner.

Generally speaking, security related data (in short security data) refers to the datasets that contain valuable information, which makes possible to figure out security issues, attacks, holes or threats by analyzing and mining them. The characteristics of the security related data are summarized as below: (1) Multi-class classification for data analysis. In a complex network, there are various types of security threats. Measurement results should be a composition of the analysis results based on all types of security related data instead of one or two types. (2) Big size. Due to the wide coverage of networks, the data collected for security measurement usually have high volume and high dimension. (3) Rich in security related information. Valuable information is carried by security related data for the sake of network security measurement. (4) Privacy issue. Due to the private information of users contained in security data, privacy issues should be taken into consideration.

The specific characteristics of security related data cause special challenges on network security measurement. Obviously, special and novel data analysis methods should be innovated and developed in order to overcome the potential challenges. However, the research in this field is still in its infancy. In this paper, we focus on the problem of big data analysis methods for security measurement, and mainly investigate the existing algorithms in different processes of big data analysis including Data Dimension Reduction, Data Classification and Data Composition. We also evaluate the existing methods in terms of criteria proposed by us in Sect. 2.

The rest of the paper is organized as follows. Section 2 introduces the preliminaries of network security related data analysis. Section 3 presents the algorithms of data dimension reduction and recent research on data composition. Section 4 reviews the existing schemes for data classification with comparison. Section 5 proposes open issues and future research trends. Finally, a conclusion is provided in the last section.

2 Preliminaries of Big Data Analysis

2.1 General Procedure of Big Data Analysis

In the past decades, a general data analysis procedure including data dimension reduction, data classification was often used in the areas of intrusion detection, malware detection, and medical diagnosis. In [1], Zhao applied a data mining method into intrusion detection. In [2] Jamdagni et al. utilized data dimension reduction and data classification methods to achieve an accurate and efficient real-time payload-based intrusion detection system. In [3] Bolzoni et al. presented Panacea method to classify attacks detected by an anomaly-based network intrusion detection system where Support Vector Machine and a rule induction algorithm called RIPPER were used. In [4], Li, Ge, and Dai studied a malware detection scheme using a Support Vector Machine (SVM) based approach. And in [5], Probabilistic Neural Network was used

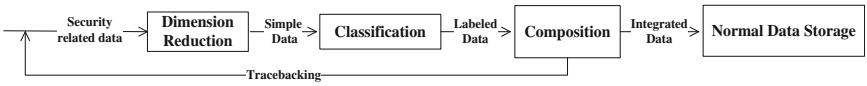

Fig. 1. A general data analysis procedure

for electrocardiogram beat classification task in order to help medical doctors in a decision-making process for heart problems. Herein, we introduce it from the perspective of network security measurement. And the whole procedure is shown in Fig. 1. We utilize dimension reduction and classification method to the network security related data, in order to get labeled data set. And then, after composition processing, the integrated data is stored for the network security measurement.

- **Dimension reduction**
  As what we have described in the Introduction, network security related datasets usually have high dimension lead to complex computation and high cost. Data dimension reduction method can simplify high dimension datasets for more accurate and efficient data analysis in the later. As shown in Fig. 1, the input of the Dimension Reduction module is the security related data, and the output should be the simple datasets with lower dimension. Principal Component Analysis (PCA) [6, 7] and Linear Discriminant Analysis (LDA) [8] are the common linear dimensionality reduction algorithms used in the literature. Kernel PCA and Laplacian eigenmaps are also popularly used for nonlinear dimension reduction [9].
- **Data classification**
  The goal of classification algorithms is to construct a suitable classifier including a training phase and a testing phase according to a data set. The data classification can discover knowledge from security related data, which make network security measurement easily and intuitively. In Fig. 1, the input of classification is the simple data output from the dimension reduction module and the output of classification module is labeled data. We will introduce the processing in details in Sect. 4.
- **Data composition**
  Data composition refers to integrating data from various categories for knowledge discovery. It can effectively integrate labeled data in order to provide a more comprehensive and meaningful result than that offered by processing single data. As shown in Fig. 1, the input of the data composition module is the labeled data output from the classification module, and the output of the composition should be integrated data which will be stored to make a decision of the network security related data. According to the integrated data, network security level and a response mechanism for security incidents can be obtained. In addition, a traceback process of these data can be done by using the processed data in this phase to figure out the reason of security incidents. But few efforts have been made to research data composition in the literature based on our search.

2.2 Data Analysis Evaluation Criteria

Data analysis methods have been proposed in many research fields. To evaluate the effectiveness of existing approaches and compare their pros and cons, holistic criteria should be proposed to serve as an evaluation metrics for this survey work.

- **Data size and distribution**
  Size. Data analysis methods like dimension reduction and classification always have size limitation. Once data size limitation is exceeded, the error of data analysis result will increase.
  Distribution. Balance and imbalance are the two types of dataset distribution. Many novel classification algorithms are designed in order to classify imbalanced data with better performance than traditional algorithms.
- **Validity**
  Accuracy. Accuracy refers to correct judgment on security threats in network measurement, which demonstrates data analysis validity. Prediction results from classification can be divided into four classes including TP (true sample number predicted positive), FP (false sample number predicted positive), FN (false sample number predicted negative), and TN (true sample number predicted negative). *Accuracy rate* can be calculated as (TP + TN)/(TP + FN + FP + TN). *Error rate* is another criterion to judge the accuracy, which can be calculated as (FP + FN)/(TP + FP + TN + FN).
- **Efficiency**
  Network security related data have such characteristics as high volume and high dimension. This leads to big challenges on data processing. Analysis time is always used to indicate the efficiency of a data processing and analyzing method.
- **Data security and privacy**
  Network user's private information could be carried by the network security related data. They hope no personal information leakage when network security is measured.
- **Traceability**
  Traceback processing is used to find the reason why a security threat occurs. With traceability, it is possible to take a timely response to insecure network according to traceback results. The requirement of traceability guarantees solving network security problems correctly. Maybe, a hash function can be applied into security related data. The hash code of data can be added into the data label to support effective traceback.

3 Schemes for Data Dimension Reduction

Development of data collection and storage during the past few years have led to the curse of dimensionality in most sciences. Data dimension reduction can be used in the data mining task like data classification and clustering with high accuracy and efficiency. Datasets like network security related data are very high dimensional, which cause many challenges in data analysis. Thus, data dimension reduction is important for accurate and

Table 1. Comparison of dimension reduction algorithms

DR	Pros	Cons
Linear dimension reduction	Simple and efficient	But for non-linear problems, they cannot achieve a good result
Non-linear dimension reduction	Effective, support non-linear dataset structure	But with high complexity and high operation cost

DR: Dimension Reduction

efficient data classification. Herein, we review the data dimension reduction algorithms based on two categories: Linear Dimension Reduction and Non-linear Dimension Reduction. The pros and cons of the two classes are summarized in Table 1.

3.1 Linear Dimension Reduction

Linear dimension reduction is often used to reduce subset of original feature and then we can get linear combinations of original data.

- Principal Component Analysis (PCA): PCA is one of the most typical methods in Linear Dimension Reduction. Its main idea is to find the least error direction in a sample space by computing the eigenvalues of the input data covariance matrix, so that the high dimension data is linearly transformed into low dimensional data. In [10], Selamat et al. used PCA together with hybrid multiclass SVM to enhance the performance of image face recognition. But the approach also used Discrete Wavelet Transform (DWT) to overcome the disadvantages of PCA in handling noises.
- Linear Discriminant Analysis (LDA): LDA is designed to maximize the distance between the different categories of features and minimize the distance between the features of the same category. In other words, it maximizes between-class distance and minimizes within-class scatter. When a LDA method performs dimension reduction, the maximum possible discriminatory information will be preserved. In [11], Lee et al. applied a LDA dimension reduction method for multi-labeled problems. From their experimental results, we can find that the LDA method can improve multiple class labels.
- Classical Multi-Dimensional Scaling (CMDS): CMDS can keep distance information in the lower dimension Euclidean space between the samples in high-dimension after dimension reduction as much as possible. For solving the problem of slow speed of CMDS, Qu et al. [12] proposed a divided-and-conquer based MDS (dcMDS) algorithm. It can significantly improve efficiency, if the intrinsic dimension of the dataset is much smaller than its size.

3.2 Non-linear Dimension Reduction

When data has inherently non-linear structure, it is difficult for linear dimension reduction methods to achieve dimension reduction effectively. It may lead to the loss of

structure information. If the two data points are near to one another in high-dimension, it should be also near in a reduced dimension space. The common methods like Isometric mapping (Isomap), Locally Linear Embedding (LLE), and Kernel PCA (KPCA) can transform data into a lower dimensional space with the non-linear structure of data retained.

- Isomap: Instead of Euclidean distance, Geodesic distance is used in Isomap. It ensures non-linear structure, when reducing dataset into a lower space. Compared with other dimension reduction methods, studies on Isomap are fewer. In [13], Cheng et al. proposed a pairwise-constraint supervised Isomap algorithm (PC-SIsomap) to achieve dimension reduction. In the approach, pairwise constraint information is introduced for replacing geodesic distance for the sake of obtaining a new distance. After dimensionality reduction, they employed the Back Propagation Neural Network (BP Neural Network) to map high dimension features into lower space in order to solve the problem of lacking samples. Finally, Support Vector Machine classification was used to evaluate the validity of the new method. We find that PC-SIomap can improve the classification accuracy and reduce the residual value.
- LLE: In LLE, the local weights are used to preserve local geometry in order to keep global non-linear structure of the datasets. In [14], Sun et al. proposed an effective feature fusion method based on LLE to overcome the challenges of handling different kinds of features and classification efficiency. It can also fuse features to a lower dimensional feature space than traditional LLE algorithms.
- Kernel PCA: KPCA uses the kernel method for principal component, similar to the kernel method in Support Vector Machine (SVM). First, KPCA maps the input samples to a high dimensional space using the kernel method. In the high-dimensional space, the data vary linearly. Thus, the data in the high dimensional space are corresponding to the non-linear data in a low dimensional space. In [15], Ha et al. combined a new C-KPCA (Custom KPCA) method created by combining a set of kernel functions with Support Vector Machine to improve classification accuracy and reduce classification time. First, they used Singular Value Decomposition (SVD) to reduce data dimension. Then, the proposed custom kernel function was used to map an input space to a higher-dimensional feature space than the former space. Experimental results showed that C-KPCA performs well in cancer classification process.

4 Scheme for Data Classification

Classification is the key method for data analysis. It is the purpose of data dimension reduction and the premise of data composition. Algorithms such as Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) are used commonly to solve classification problems. Due to their simplicity and accuracy, we mainly investigate these four classification methods. In Table 2, we make a summary about their pros and cons. And in Table 3, we compare various algorithms proposed recently by applying the criteria specified in Sect. 2.2.

Table 2. Algorithm comparison

CA	Pros	Cons
DT	Multi-class classification, no domain knowledge is needed, handle high dimensional data, simple, fast, robust, and good accuracy	But costs are associated with the decision attributes, classification accuracy, and memory requirements
NB	Simple, effective, efficient, and widely used	But it requires to making strong conditional independence assumptions
SVM	Accurate, robust, good for nonlinear samples and high dimensional pattern recognition	But performance is poor to big size data and high computational complexity quadratic programming problem, and cannot be directly used for multi-class classification
KNN	Simple, useful, effective, well-established and competitive classification performance	But high classification cost for large datasets, vulnerable to parameter selection, and difficult to determine an appropriate dissimilarity measurement

CA: Classification Algorithm

4.1 Decision Tree (DT)

Decision Tree (DT) is a well-known top-down greedy classification method. In a given decision tree, the root, the set of nodes, and the set of edges are contained. Each internal node denotes a test on an attribute, each branch represents an output of the test, and each terminal node holds a class label. Generally, building a decision tree according to the selected attributes contains two steps including tree-growing step and pruning back step. Gini diversity is the common criterion used to split. And the pruning criterion is used to prevent "overfitting" by simplifying the tree structure with the goal of improving the classification accuracy in DT. Fierens et al. investigated the pruning criterion for probability trees [16]. DT algorithms such as ID3, ID4, ID5, C4.5, C5.0, and CART have been proposed during the past decade. Other DT-based algorithms such as boosted DTs, Rotation Forest and Random Forest are also used in practice. In [17] Choi et al. analyzed medical big data about foot disorder patients for efficient classification and analysis. They composed an independent variable (foot disorder record) into 24 attributes (e.g., sex, age, etc.). A DT prediction model was developed by them to obtain useful information effectively between foot disable groups and biomechanical parameters related to symptom. According to their results, 12 rules were generated to achieve classification and analysis process. To solve the cost problem of classification for multiple condition attributes, Chen et al. proposed a new algorithm [18]. For splitting, they selected a decision attribute under the cost constraint with the best-cost ratio. With experiments, they demonstrated that the algorithm is very effective to achieve good accuracy with limited costs.

What's more, in order to solve the problems of no dependencies considered among attributes, Yen et al. proposed a Neural Decision Tree (NDT) model [19] by combining neural network with traditional decision-tree to handle real world data. The neural network with a back-propagation (BP) model is used to find the dependencies among

attributes. Then, a traditional DT learning algorithm like C5 was applied to receive the training data and the results obtained by the neural network model to derive a more correct decision tree than traditional one. To achieve multi-class classification task accurately, Farid et al. introduced an adaptive naïve Bayes tree (NBTree) algorithm [20]. In their approach, NBTree nodes contain and split as common decision trees, but the leaves are replaced by naïve Bayes classifier to handle attributes. The approach uses decision tree induction to select a subset of attributes from training dataset to calculate the naïve assumption of class conditional independence. And then they used naïve Bayes classifier at the terminal node to deal with attributes for making the prediction. But the data size that can be supported by this algorithm is not as big as we expect.

4.2 Naïve Bayes (NB)

As a statistical classification method, NB can predict the probability of class membership. We call the probability as posterior probability. An instance will be classified into the category with maximum posterior probability using NB classification according to its prior probability. In order to reduce computation complexity, it supposes that all features are independent with each other, which is called feature independence assumption.

Recently, Naïve Bayes has been popularly applied in many fields, such as sentiment mining [21, 22], text classification [23] and some computation approaches [24]. Besides, to overcome the independence assumption, Bayes network was proposed. Bielza et al. wrote an overview of Bayesian network classifiers recently [25]. They comprehensively surveyed all kinds of discrete Bayesian network classifiers. However, learning the optimal structure of a Bayesian network from high dimension training datasets is almost impossible due to time and space consumption.

In order to solve the problem that available labeled data are limited, Jiang provided a fast and highly effective semi-supervised learning algorithm called Instance Weighted Naïve Bayes (IWNB) [26]. They use the maximal class membership probability estimated by the trained naïve Bayes to weight each instance in an unlabeled dataset. Finally, the training processing continues using both originally labeled data and newly labeled and weighted data. The approach can improve classification accuracy and efficiency, when unlabeled data sets are much bigger than labeled data. And Xue et al. [27] proposed a method called SWNB (SEIR immune-strategy-based instance weighting algorithm for naïve Bayes classification) to estimate accurate NB classifier for effective classification in the situation that the number of training instances is small. The method calculated optimal instance weight value automatically for each dataset in IWNB based on SEIR (Susceptible, Exposed, Infectious and Recovered) Immune strategy to obtain priori probability and conditional probability.

To weaken the attribute independence assumption, Webb et al. proposed a new classification method called Aggregating One-Dependence Estimators (AODE) [28]. The method averages all of the constrained class of classifiers to learn an aggregate of one-dependence classifiers. But Jiang et al. observed that the weights in different one-dependence classifiers are the same [29]. If assigning different weights to these one-dependence classifiers, the original model can be improved. So they gave another

model called Weighted Average of One-Dependence Estimators (WAODE). And as an application in text data analysis, Jiang et al. [30] proposed a novel model called Structure Extended Multinomial Naïve Bayes (SEMNB) based on WAODE in [29]. The algorithm was proved efficiency, accuracy and computational simplicity for real-world high-dimensional text data classification on NB classifiers.

4.3 Support Vector Machine (SVM)

SVM is a type of binary classifier that was proposed originally by Cortes and Vapnik for binary classification [31]. It has been introduced in the framework of Structural Risk Minimization (SRM) learning theory and used in statistical learning theory for machine learning. For non-linear SVM, kernel function is utilized to transform current feature into higher dimensional space. And in the higher dimensional data space, non-linear SVM can be trained as linear one. In [32], Keith et al. introduced an algorithm called Kernel Genetic Programming (KGP) to find near-optimal kernels. But the approach performed poor in large datasets. For wide application usage, Vapnik compared multi-class classification algorithms including one-against-all and one-against-one based on SVM by combining several binary classifiers together [33]. In the past few years, SVM is very popular in many fields like medicine [34], face detection [35], images classification [36]. And it became more popular and was widely used after SVM software LIBSVM was developed by Chang et al. [37].

For big data classification, Laachemi et al. developed a new approach for web services supervised categorization by combining Stochastic Local Search (SLS) with SVM [41]. SLS is a metaheuristic used for feature selection to select good attributes for SVM. For SVM, they implemented it based on LIBSVM [37]. Through comparison, we found that SVM+SLS approach is more accurate than other approaches like WEKA or NB in supervised classification of Web services.

Class-imbalance problem may exist in network security related data. Hao et al. [38] proposed a method called Maximal-margin Spherical-structured Multi-class Support Vector Machine (MSM-SVM) that use hyperspheres to solve class-imbalance problem. For each class, the hyperspheres are constructed by finding its center and radius. Each hyperspheres encloses all positive examples but excludes all negative examples. This approach is proved beneficial in dealing with imbalance problems compared with hyperplane-based multi-class SVM. As an application of SVM for malware detection, the approach was applied successfully in [39]. Comar et al. used macro-level binary classifier and micro-level classifiers to achieve classification with high precision. They employed random forest as the macro-level binary classifier, and the 1-class SVM method described in [38] as micro-classifiers. They used the thinking of hyperspheres in MSM-SVM into the 1-class SVM method to classify malwares and obtained a relatively accurate result. And to improve the performance of large datasets multi-class classification, Qing et al. [40] proposed a method called Least Squares Twin SVM Partial Decision Tree (LSTSVM-PDT) with partial binary tree constructed for multi-class classification and LSTSVM used in a non-terminal node.

4.4 K Nearest Neighbor (KNN)

As a popular and easy to implement nonparametric classification classifier, KNN has been widely used in multi-class classification problems [42]. Unlike eager learner methods such as DT, KNN belongs to a lazy learner method, which has no training stage. Traditional KNN is a classification method based on a distance function like the Euclidean distance. According to the labels of a sample's K closet neighbors, the sample is classified. The important problem in the area of KNN is how to choose the optimum value of K. For getting a better value of K than traditional KNN, Zhu et al. [43] presented a novel parameter-free concept called Nearest Neighbor (NaN) inspired by the friendship of human society. They used Natural Neighbor Eigenvalue (NaNE) in place of the parameter K in the traditional KNN method.

In order to select optimal nearest neighbors, Tang et al. [44] provided a classification method called Extended Nearest Neighbor (ENN). Unlike the traditional KNN, in addition to considering who the nearest neighbors of the test sample are, ENN also considers who considers the test sample as its nearest neighbors. The algorithm can learn from the global distribution by using all available training data to make a classification decision in order to improve pattern distribution. Similarly, İnkaya et al. [45] developed a parameter-free neighborhood classifier based on Surrounding Influence Region (SIR) decision rules to classify samples correctly in order to select optimal nearest neighbors. In their study, neighbors are determined according to distance, density and connectivity information.

Table 3. Scheme comparison

CA	Ref	V (%)		NA	S&P	T	MC	AD
C5.0	[17]	A	72.96	6610	×	×	×	Improving performance of clinical data analysis
DT	[18]	AE1	96.81	3772	×	×	×	Cost-constrained decision tree with multiple conditions
		AE2	84.01	3772				
NDT	[19]	Error rate	3.41	150	×	×	×	Improving accuracy and decision tree size and handling the problems of attribute dependencies
NBTree	[20]	A	99	150	×	×	√	Improving accuracy rates of multi-class classification problems
IWNB	[26]	A	97.00	699	×	×	√	Fast, simple and learning naïve Bayes for both labeled and unlabeled data
SWNB	[27]	A	98.67	3196	×	×	√	Self-adaptive instance weighted and accurate if the number of training instances is small
SEMNB	[30]	A	96.86	927	×	×	√	Simple, efficient, accurate and weakening attributes independence assumption

(*continued*)

Table 3. (*continued*)

CA	Ref	V (%)		NA	S&P	T	MC	AD
MSM-SVM	[38]	A	99.45	N/A	×	×	√	Handling class-imbalance problems, accurate with optimal parameters
LSTSVM-PDT	[40]	A	96.34	2310	×	×	√	Supporting large datasets, accurate and efficient
		CT (s)	100.2					
SLS-SVM	[41]	A	86.81	364	×	×	√	Accurate in supervised classification of Web service
		CT (s)	117.68					
NaN	[43]	A	84.681	N/A	×	×	√	Parameter-free, handling different types of data and noise in data
ENN	[44]	ER	10.08	4601	×	×	√	Speeding up the searching of nearest neighbors and reducing complexity of computation
SIR	[45]	ER	4.45	683	×	×	√	Successful classification in datasets with density differences and arbitrary shapes

CA: Classification Algorithm, Ref: Reference, V: Validity, NA: Number of Instances with Maximal Accuracy, DS: Datasets, UCI: Benchmark datasets introduced in Sect. 4.3, S&P: Security and Privacy, T: Traceability, MC: Multi-class Classification, AD: Algorithm Description, AE1: Maximal Accuracy in Experiment1, AE2: Maximal Accuracy in Experiment2, A: Maximal Accuracy, CT: Classification Time, ER: Error Rate, N/A: Not Available, UCI/WEKA/ATF/QWS: The common dataset for data analysis

5 Open Issues and Future Research Trends

5.1 Open Issues

Based on the above review, we find the following open issues:

- Security and privacy for data analysis processing are important in some special cases, but few work focus on this problem. None classification methods reviewed above pay attention to security and privacy although there are some studies about privacy-preserving data mining as reviewed by Yan et al. [47]. Many methods were proposed for achieving accurate classification. Past work mainly focused on how to gain an effective method without concern on data privacy protection. If data owners would not like to disclose their data to data analyzer, the existing methods as reviewed above cannot be applied. Although homomorphic encryption and Secure Multiparty Computation have been applied to support ciphertext computation. Still, they are not mature enough to fully support the traditional data reduction and data classification algorithms. More crucially, high computation complexity and

overhead are a serious open issue of crypto-based schemes for privacy-preserving data reduction, data classification, and data composition.

- Traceability of data composition was seldom discussed in the literature as we have reviewed above. After data classification, the data will be used for knowledge discovery from original dataset without any tracking process on the data. But we need to know when and how the data is labeled with a special label. In network security measurement, traceability is essential in order to find the source of security treats and attacks in the network.
- There are very few literatures that utilize real network security related data to evaluate the performance of data analysis methods. As shown in Table 3, most of the classification algorithms were tested based on standard dataset. We should make additional efforts in order to show if they are appropriate to be applied into network security related data analysis for network security measurement.

5.2 Future Research Trends

With regard to network security measurement, we propose the following research directions based on the above literature review and discussion on open issues.

First, privacy-preserving network security measurement should be seriously studied in the whole process of security related data collection, transmission, processing and analysis. In [46] differential privacy model was embedded into NB classification for data privacy preservation. Perhaps, differential privacy model is a good trend to ensure security and privacy during data analysis. How to propose an effective scheme to combine the differential privacy model with classification together is an interesting topic.

Second, multi-class classification with sound performance is essential for analyzing security related data. From Table 3, we can see that most algorithms support multi-class classification, especially [19]. Binary classification methods should be extended to multi-class classification methods for wide usage. But due to the special characteristics of the network security related data, high performance, especially efficiency should be ensured. Highly efficient data analysis methods should be investigated in the future research.

Fourth, combining several methods together to improve the entire performance of data analysis for the purpose of produce a generic framework for security related data analysis is another direction. Classification is difficult for complex datasets. If we combine data reduction methods and data classification methods effectively, the data analysis could become more efficient, like the method used in [41]. What's more, combining different methods can achieve complementary advantages to avoid the limitation of a single algorithm as demonstrated in [18, 19, 40]. In addition, data fusion methods should be applied as a useful research method to kick out redundant, noisy and spam data in order to achieve high efficiency and trustworthiness of data analysis. This will be a very interesting research topic worth our investigation.

Finally, introducing data composition method into data analysis scheme in order to figure out network security level and meanwhile providing traceability and ability of provenance management for seeking the source of security threats will be a very

significant research topic. However, very few existing work take data composition methods into consideration based on our literature study and review because of its simplicity. No existing work integrates composition with provenance management, especially granular traceability.

6 Conclusions

In big data era, data analysis is a hot research topic for knowledge discovery. As a kind of big data, network security related data offer us great possibility to measure network security. In this paper, we review the methods of data analysis, mainly focusing on data reduction and data classification. We studied and analyzed some common data dimension reduction methods and mainly investigated the data classification algorithms proposed in recent year. We compared the performance of the reviewed classification algorithms and commented their support on security related data analysis. Based on the survey, we pointed out some open issues and proposed future research directions in the area of data analysis for network security measurement.

Acknowledgment. This work is sponsored by the National Key Research and Development Program of China (grant 2016YFB0800704), the NSFC (grants 61672410 and U1536202), the Project Supported by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016ZDJC-06), the 111 project (grants B08038 and B16037), and Academy of Finland (Grant No. 308087).

References

1. Zhao, Y.: Network intrusion detection system model based on data mining. In: 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 155–160. IEEE, Shanghai, China (2016)
2. Jamdagni, A., Tan, Z., He, X., Nanda, P., Liu, R.P.: Repids: a multi tier real-time payload-based intrusion detection system. Comput. Netw. **57**(3), 811–824 (2013)
3. Bolzoni, D., Etalle, S., Hartel, P.H.: Panacea: automating attack classification for anomaly-based network intrusion detection systems. In: Kirda, E., Jha, S., Balzarotti, D. (eds.) RAID 2009. LNCS, vol. 5758, pp. 1–20. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04342-0_1
4. Li, W., Ge, J., Dai, G.: Detecting malware for android platform: an svm-based approach. In: 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 464–469. IEEE, New York, NY, USA (2015)
5. Banupriya, C.V., Karpagavalli, S.: Electrocardiogram beat classification using probabilistic neural network. IJCA Proc. Mach. Learn. Challenges Oppor. Ahead **1**, 31–37 (2014). MLCONF
6. Peason, K.: On lines and planes of closest fit to systems of point in space. Phil. Mag. **2**(11), 559–572 (1901)
7. Jolliffe, I.T.: Principal Component Analysis. 2nd edn. Springer Series in Statistics (2002)
8. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Acadamic Press, San Diego (1990)

9. Romdhani, S., Gong, S.: A multi-view nonlinear active shape model. Br. Mach. Vis. Conf. (BMVC) **10**, 483–492 (2002)
10. Selamat, M.H., Rais, H.M.: Image face recognition using Hybrid Multiclass SVM (HM-SVM). In: International Conference on Computer, Control, Informatics and ITS Applications (IC3INA), pp. 159–164. IEEE, Bandung (2015)
11. Lee, M., Park, C.H.: On applying dimension reduction for multi-labeled problems. In: Perner, P. (ed.) MLDM 2007. LNCS, vol. 4571, pp. 131–143. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73499-4_11
12. Qu, T., Cai, Z.: A fast multidimensional scaling algorithm. In: 2015 IEEE International Conference on Robotics and Riomimetics (ROBIO), pp. 2569–2574. IEEE, Zhuai, China (2015)
13. Cheng, J., Cheng, C., Guo, Y.: Supervised Isomap based on pairwise constraints. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012. LNCS, vol. 7663, pp. 447–454. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34475-6_54
14. Sun, B.Y., Zhang, X.M., Li, J., Mao, X.M.: Feature fusion using locally linear embedding for classification. IEEE Trans. Neural Netw. **21**(1), 163–168 (2010)
15. Ha, V.S., Nguyen, H.N.: C-KPCA: custom kernel PCA for cancer classification. In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. LNCS, vol. 9729, pp. 459–467. Springer, Cham (2016). doi:10.1007/978-3-319-41920-6_36
16. Fierens, D., Ramon, J., Blockeel, H., Bruynooghe, M.: A comparison of pruning criteria for probability trees. Mach. Learn. **78**(1), 251–285 (2010)
17. Choi, J.K., Jeon, K.H., Won, Y., Kim, J.J.: Application of big data analysis with decision tree for the foot disorder. Cluster Comput. **18**(4), 1399–1404 (2015)
18. Chen, Y.L., Wu, C.C., Tang, K.: Building a cost-constrained decision tree with multiple condition attributes. Inf. Sci. **179**(7), 967–979 (2009)
19. Yen, S.J., Lee, Y.S.: A neural network approach to discover attribute dependency for improving the performance of classification. Expert Syst. Appl. **38**(10), 12328–12338 (2011)
20. Farid, D.M., Rahman, M.M., Al-Mamuny, M.A.: Efficient and scalable multi-class classification using Naïve Bayes tree. In: 2014 International Conference on Informatics, Electronics & Vision (ICIEV), pp. 1–4. IEEE, Dhaka, Bangladesh (2014)
21. Sinha, H., Bagga, R., Raj, G.: An analysis of ICON aircraft log through sentiment analysis using SVM and Naive Bayes classification. In: International Conference on Information Technology (InCITe), The Next Generation IT Summit on the Theme-Internet of Things: Connect your Worlds, pp. 53–58. IEEE, Noida, India (2016)
22. Mertiya, M., Singh, A.: Combining Naive Bayes and adjective analysis for sentiment detection on Twitter. In: International Conference on Inventive Computation Technologies (ICICT), vol. 2, pp. 1–6. IEEE, Coimbatore, India (2016)
23. Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., Zhang, C.: Self-adaptive attribute weighting for Naive Bayes classification. Expert Syst. Appl. **42**(3), 1487–1502 (2015)
24. Naderpour, M., Lu, J., Zhang, G.: A fuzzy dynamic bayesian network-based situation assessment approach. In: 2013 IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–8. IEEE, Hyderabad, India (2013)
25. Bielza, C., Larrañaga, P.: Discrete Bayesian network classifiers: a survey. ACM Comput. Surv. (CSUR) **47**(1), 5 (2014)
26. Jiang, L.: Learning instance weighted Naive Bayes from labeled and unlabeled data. J. Intell. Inf. Syst. **38**(1), 257–268 (2012)
27. Xue, S., Lu, J., Zhang, G., Xiong, L.: SEIR immune strategy for instance weighted Naive Bayes classification. In: Arik, S., Huang, T., Lai, W.K., Liu, Q. (eds.) ICONIP 2015. LNCS, vol. 9489, pp. 283–292. Springer, Cham (2015). doi:10.1007/978-3-319-26532-2_31

28. Webb, G.I., Boughton, J.R., Wang, Z.: Not so naive Bayes: aggregating one-dependence estimators. Mach. Learn. **58**(1), 5–24 (2005)
29. Jiang, L., Zhang, H., Cai, Z., Wang, D.: Weighted average of one-dependence estimators. J. Exp. Theor. Artif. Intell. **24**(2), 219–230 (2012)
30. Jiang, L., Wang, S., Li, C., Zhang, L.: Structure extended multinomial naive Bayes. Inf. Sci. **329**, 346–356 (2016)
31. Cortes, C., Vapnik, V.: Support-vector network. Mach. Learning **20**(3), 273–297 (1995)
32. Sullivan, K.M., Luke, S.: Evolving kernels for support vector machine classification. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, pp. 1702–1707. ACM, London, England (2007)
33. Vapnik, V.: The Nature of Statistical Learning. Springer, New York (1995)
34. Annam, J.R., Surampudi, B.R.: Inter-patient heart-beat classification using complete ECG beat time series by alignment of R-peaks using SVM and decision rule. In: International Conference on Signal and Information Processing (IConSIP), pp. 1–5. IEEE, Vishnupuri, India (2016)
35. Yao, M., Zhu, C.: SVM and adaboost-based classifiers with fast PCA for face reocognition. In: 2016 IEEE International Conference on Consumer Electronics-China (ICCE-China), pp. 1–5. IEEE, Guangzhou, China (2016)
36. Lee, S.B., Jeong, E.J., Son, Y., Kim, D.J.: Classification of computed tomography scanner manufacturer using support vector machine. In: 2017 5th International Winter Conference on Brain-Computer Interface (BCI), pp. 85–87. IEEE, Sabuk, South Korea (2017)
37. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
38. Hao, P.Y., Chiang, J.H., Lin, Y.H.: A new maximal-margin spherical-structured multi-class support vector machine. Appl. Intell. **30**(2), 98–111 (2009)
39. Comar, P.M., Liu, L., Saha, S., Tan, P.N., Nucci, A.: Combining supervised and unsupervised learning for zero-day malware detection. In: 2013 Proceedings IEEE INFOCOM, pp. 2022–2030. IEEE, Turin, Italy (2013)
40. Yu, Q., Wang, L.: Least squares twin SVM decision tree for multi-class classification. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1927–1931. IEEE, Datong, China (2016)
41. Laachemi, A., Boughaci, D.: A stochastic local search combined with support vector machine for Web services classification. In: 2016 International Conference on Advanced Aspects of Software Engineering (ICAASE), pp. 9–16 IEEE, Constantine, Algera (2016)
42. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)
43. Zhu, Q., Feng, J., Huang, J.: Natural neighbor: a self-adaptive neighborhood method without parameter K. Pattern Recogn. Lett. **80**, 30–36 (2016)
44. Tang, B., He, H.: ENN: extended nearest neighbor method for pattern recognition [research frontier]. IEEE Comput. Intell. Mag. **10**(3), 52–60 (2015)
45. İnkaya, T.: A density and connectivity based decision rule for pattern classification. Expert Syst. Appl. **42**(2), 906–912 (2015)
46. Vaidya, J., Shafiq, B., Basu, A., Hong, Y.: Differentially private Naive Bayes classification. In: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), pp. 571–576. IEEE, Atlanta, GA, USA (2013)
47. Yan, Z., Zhang, P., Vasilakos, A.V.: A survey on trust management for Internet of Things. J. Netw. Comput. Appl. **42**(2014), 120–134 (2014)