http://www.icact.org/journal.asp

ICACT-TACT JOURNAL

Transactions on Advanced Communications Technology

Volume 6 Issue 4, Jul. 2017, ISSN: 2288-0003

Editor-in-Chief
Prof. Thomas Byeongnam YOON, PhD.

GIRI Global IT Research Institute

Journal Editorial Board

■ Editor-in-Chief

■ Editors

Dr. Youssef SAID, Tunisie Telecom, Tunisia
Dr. Noor Zaman, King Faisal University, Al Ahsa Hofuf, Saudi Arabia
Dr. Srinivas Mantha, SASTRA University, Thanjavur, India
Dr. Shahriar Mohammadi, KNTU University, Iran
Prof. Beonsku An, Hongik University, korea
Dr. Guanbo Zheng, University of Houston, USA
Prof. Sangho Choe, The Catholic University of Korea, korea
Dr. Gyanendra Prasad Joshi, Yeungnam University, korea
Dr. Tae-Gyu Lee, Korea Institue of Industrial Technology(KITECH), korea
Prof. Ilkyeun Ra, University of Colorado Denver, USA
Dr. Yong Sun, Beijing University of Posts and Telecommunications, China
Dr. Yulei Wu, Chinese Academy of Sciences, China
Mr. Anup Thapa, Chosun University, korea
Dr. Vo Nguyen Quoc Bao, Posts and Telecommunications Institute of Technology, Vietnam
Dr. Harish Kumar, Bhagwant institute of technology, India
Dr. Jin REN, North china university of technology, China
Dr. Joseph Kandath, Electronics & Commn Engg, India
Dr. Mohamed M. A. Moustafa, Arab Information Union (AIU), Egypt
Dr. Mostafa Zaman Chowdhury, Kookmin University, Korea
Prof. Francis C.M. Lau, Hong Kong Polytechnic University, Hong Kong
Prof. Ju Bin Song, Kyung Hee University, korea
Prof. KyungHi Chang, Inha University, Korea
Prof. Sherif Welsen Shaker, Kuang-Chi Institute of Advanced Technology, China
Prof. Seung-Hoon Hwang, Dongguk University, Korea
Prof. Dal-Hwan Yoon, Semyung University, korea
Prof. Chongyang ZHANG, Shanghai Jiao Tong University, China
Dr. H K Lau, The Open University of Hong Kong, Honh Kong
Prof. Ying-Ren Chien, Department of Electrical Engineering, National Ilan University, Taiwan
Prof. Mai Yi-Ting, Hsiuping University of Science and Technology, Taiwan
Dr. Sang-Hwan Ryu, Korea Railroad Research Institute, Korea
Dr. Yung-Chien Shih, MediaTek Inc., Taiwan
Dr. Kuan Hoong Poo, Multimedia University, Malaysia
Dr. Michael Leung, CEng MIET SMIEEE, Hong Kong
Dr. Abu sahman Bin mohd Supa'at, Universiti Teknologi Malaysia, Malaysia
Prof. Amit Kumar Garg, Deenbandhu Chhotu Ram University of Science & Technology, India
Dr. Jens Myrup Pedersen, Aalborg University, Denmark
Dr. Augustine Ikechi Ukaegbu, KAIST, Korea
Dr. Jamshid Sangirov, KAIST, Korea
Prof. Ahmed Dooguy KORA, Ecole Sup. Multinationale des Telecommunications, Senegal
Dr. Se-Jin Oh, Korea Astronomy & Space Science Institute, Korea
Dr. Rajendra Prasad Mahajan, RGPV Bhopal, India
Dr. Woo-Jin Byun, ETRI, Korea
Dr. Mohammed M. Kadhum, School of Computing, Goodwin Hall, Queen's University , Canada
Prof. Seong Gon Choi, Chungbuk National University, Korea
Prof. Yao-Chung Chang, National Taitung University, Taiwan
Dr. Abdallah Handoura, Engineering school of Gabes - Tunisia, Tunisia
Dr. Gopal Chandra Manna, BSNL, India

Dr. Il Kwon Cho, National Information Society Agency, Korea
Prof. Jiann-Liang Chen, National Taiwan University of Science and Technology, Taiwan
Prof. Ruay-Shiung Chang, National Dong Hwa University, Taiwan
Dr. Vasaka Visoottiviseth, Mahidol University, Thailand
Prof. Dae-Ki Kang, Dongseo University, Korea
Dr. Yong-Sik Choi, Research Institute, IDLE co., ltd, Korea
Dr. Xuena Peng, Northeastern University, China
Dr. Ming-Shen Jian, National Formosa University, Taiwan
Dr. Soobin Lee, KAIST Institute for IT Convergence, Korea
Prof. Yongpan Liu, Tsinghua University, China
Prof. Chih-Lin HU, National Central University, Taiwan
Prof. Chen-Shie Ho, Oriental Institute of Technology, Taiwan
Dr. Hyoung-Jun Kim, ETRI, Korea
Prof. Bernard Cousin, IRISA/Universite de Rennes 1, France
Prof. Eun-young Lee, Dongduk Woman s University, Korea
Dr. Porkumaran K, NGP institute of technology India, India
Dr. Feng CHENG, Hasso Plattner Institute at University of Potsdam, Germany
Prof. El-Sayed M. El-Alfy, King Fahd University of Petroleum and Minerals, Saudi Arabia
Prof. Lin You, Hangzhou Dianzi Univ, China
Mr. Nicolai Kuntze, Fraunhofer Institute for Secure Information Technology, Germany
Dr. Min-Hong Yun, ETRI, Korea
Dr. Seong Joon Lee, Korea Electrotechnology Research Institute, korea
Dr. Kwihoon Kim, ETRI, Korea
Dr. Jin Woo HONG, Electronics and Telecommunications Research Inst., Korea
Dr. Heeseok Choi, KISTI(Korea Institute of Science and Technology Information), korea
Dr. Somkiat Kitjongthawonkul, Australian Catholic University, St Patrick's Campus, Australia
Dr. Dae Won Kim, ETRI, Korea
Dr. Ho-Jin CHOI, KAIST(Univ), Korea
Dr. Su-Cheng HAW, Multimedia University, Faculty of Information Technology, Malaysia
Dr. Myoung-Jin Kim, Soongsil University, Korea
Dr. Gyu Myoung Lee, Institut Mines-Telecom, Telecom SudParis, France
Dr. Dongkyun Kim, KISTI(Korea Institute of Science and Technology Information), Korea
Prof. Yoonhee Kim, Sookmyung Women s University, Korea
Prof. Li-Der Chou, National Central University, Taiwan
Prof. Young Woong Ko, Hallym University, Korea
Prof. Dimiter G. Velev, UNWE(University of National and World Economy), Bulgaria
Dr. Tadasuke Minagawa, Meiji University, Japan
Prof. Jun-Kyun Choi, KAIST (Univ.), Korea
Dr. Brownson ObaridoaObele, Hyundai Mobis Multimedia R&D Lab , Korea
Prof. Anisha Lal, VIT university, India
Dr. kyeong kang, University of technology sydney, faculty of engineering and IT , Australia
Prof. Chwen-Yea Lin, Tatung Institute of Commerce and Technology, Taiwan
Dr. Ting Peng, Chang'an University, China
Prof. ChaeSoo Kim, Donga University in Korea, Korea
Prof. kirankumar M. joshi, m.s.uni.of baroda, India
Dr. Chin-Feng Lin, National Taiwan Ocean University, Taiwan
Dr. Chang-shin Chung, TTA(Telecommunications Technology Association), Korea

Editor Guide

■ Introduction for Editor or Reviewer

All the editor group members are to be assigned as a evaluator(editor or reviewer) to submitted journal papers at the discretion of the Editor-in-Chief. It will be informed by eMail with a Member Login ID and Password.

Once logined the Website via the Member Login menu in left as a evaluator, you can find out the paper assigned to you. You can evaluate it there. All the results of the evaluation are supposed to be shown in the Author Homepage in the real time manner. You can also enter the Author Homepage assigned to you by the Paper ID and the author's eMail address shown in your Evaluation Webpage. In the Author Homepage, you can communicate each other efficiently under the peer review policy. Please don't miss it!

All the editor group members are supposed to be candidates of a part of the editorial board, depending on their contribution which comes from history of ICACT TACT as an active evaluator. Because the main contribution comes from sincere paper reviewing role.

■ Role of the Editor

The editor's primary responsibilities are to conduct the peer review process, and check the final camera-ready manuscripts for any technical, grammatical or typographical errors.

As a member of the editorial board of the publication, the editor is responsible for ensuring that the publication maintains the highest quality while adhering to the publication policies and procedures of the ICACT TACT(Transactions on the Advanced Communications Technology).

For each paper that the editor-in-chief gets assigned, the Secretariat of ICACT Journal will send the editor an eMail requesting the review process of the paper.

The editor is responsible to make a decision on an "accept", "reject", or "revision" to the Editor-in-Chief via the Evaluation Webpage that can be shown in the Author Homepage also.

■ Deadlines for Regular Review

Editor-in-Chief will assign a evalaution group(a Editor and 2 reviewers) in a week upon receiving a completed Journal paper submission. Evaluators are given 2 weeks to review the paper. Editors are given a week to submit a recommendation to the Editor-in-Chief via the evaluation Webpage, once all or enough of the reviews have come in. In revision case, authors have a maximum of a month to submit their revised manuscripts. The deadlines for the regular review process are as follows:

Evalution Procedure	Deadline
Selection of Evaluation Group	1 week
Review processing	2 weeks
Editor's recommendation	1 week
Final Decision Noticing	1 week

■ Making Decisions on Manuscript

Editor will make a decision on the disposition of the manuscript, based on remarks of the reviewers. The editor's recommendation must be well justified and explained in detail. In cases where the revision is requested, these should be clearly indicated and explained. The editor must then promptly convey this decision to the author. The author may contact the editor if instructions regarding amendments to the manuscript are unclear. All these actions could be done via the evaluation system in this Website. The guidelines of decisions for publication are as follows:

Decision	Description
Accept	An accept decision means that an editor is accepting the paper with no further modifications. The paper will not be seen again by the editor or by the reviewers.
Reject	The manuscript is not suitable for the ICACT TACT publication.
Revision	The paper is conditionally accepted with some requirements. A revision means that the paper should go back to the original reviewers for a second round of reviews. We strongly discourage editors from making a decision based on their own review of the manuscript if a revision had been previously required.

■ Role of the Reviewer

Reviewer Webpage:

Once logined the Member Login menu in left, you can find out papers assigned to you. You can also login the Author Homepage assigned to you with the paper ID and author's eMail address. In there you can communicate each other via a Communication Channel Box.

Quick Review Required:

You are given 2 weeks for the first round of review and 1 week for the second round of review. You must agree that time is so important for the rapidly changing IT technologies and applications trend. Please respect the deadline. Authors undoubtedly appreciate your quick review.

Anonymity:

Do not identify yourself or your organization within the review text.

Review:

Reviewer will perform the paper review based on the main criteria provided below. Please provide detailed public comments for each criterion, also available to the author.

- How this manuscript advances this field of research and/or contributes something new to the literature?
- Relevance of this manuscript to the readers of TACT?
- Is the manuscript technically sound?
- Is the paper clearly written and well organized?
- Are all figures and tables appropriately provided and are their resolution good quality?
- Does the introduction state the objectives of the manuscript encouraging the reader to read on?
- Are the references relevant and complete?

Supply missing references:

Please supply any information that you think will be useful to the author in revision for enhancing quality of the paperor for convincing him/her of the mistakes.

Review Comments:

If you find any already known results related to the manuscript, please give references to earlier papers which contain these or similar results. If the reasoning is incorrect or ambiguous, please indicate specifically where and why. If you would like to suggest that the paper be rewritten, give specific suggestions regarding which parts of the paper should be deleted, added or modified, and please indicate how.

Journal Procedure

Dear Author,

- **You can see all your paper information & progress.**

- **Step 1. Journal Full Paper Submission**

Using the Submit button, submit your journal paper through ICACT Website, then you will get new paper ID of your journal, and send your journal Paper ID to the Secretariat@icact.org for the review and editorial processing. Once you got your Journal paper ID, never submit again! Journal Paper/CRF Template

- **Step 2. Full Paper Review**

Using the evaluation system in the ICACT Website, the editor, reviewer and author can communicate each other for the good quality publication. It may take about 1 month.

- **Step 3. Acceptance Notification**

It officially informs acceptance, revision, or reject of submitted full paper after the full paper review process.

Status	Action
Acceptance	Go to next Step.
Revision	Re-submit Full Paper within 1 month after Revision Notification.
Reject	Drop everything.

- **Step 4. Payment Registration**

So far it's free of charge in case of the journal promotion paper from the registered ICACT conference paper! But you have to regist it, because you need your Journal Paper Registration ID for submission of the final CRF manuscripts in the next step's process. Once you get your Registration ID, send it to Secretariat@icact.org for further process.

- **Step 5. Camera Ready Form (CRF) Manuscripts Submission**

After you have received the confirmation notice from secretariat of ICACT, and then you are allowed to submit the final CRF manuscripts in PDF file form, the full paper and the Copyright Transfer Agreement. Journal Paper Template, Copyright Form Template, BioAbstract Template,

Journal Submission Guide

All the Out-Standing ICACT conference papers have been invited to this "ICACT Transactions on the Advanced Communications Technology" Journal, and also welcome all the authors whose conference paper has been accepted by the ICACT Technical Program Committee, if you could extend new contents at least 30% more than pure content of your conference paper. Journal paper must be followed to ensure full compliance with the IEEE Journal Template Form attached on this page.

➢ How to submit your Journal paper and check the progress?

Step 1. Submit	Using the Submit button, submit your journal paper through ICACT Website, then you will get new paper ID of your journal, and send your journal Paper ID to the Secretariat@icact.org for the review and editorial processing. Once you got your Journal paper ID, never submit again! Using the Update button, you can change any information of journal paper related or upload new full journal paper.
Step 2. Confirm	Secretariat is supposed to confirm all the necessary conditions of your journal paper to make it ready to review. In case of promotion from the conference paper to Journal paper, send us all the .DOC(or Latex) files of your ICACT conference paper and journal paper to evaluate the difference of the pure contents in between at least 30% more to avoid the self replication violation under scrutiny. The pure content does not include any reference list, acknowledgement, Appendix and author biography information.
Step 3. Review	Upon completing the confirmation, it gets started the review process thru the Editor & Reviewer Guideline. Whenever you visit the Author Homepage, you can check the progress status of your paper there from start to end like this, " Confirm OK! -> Gets started the review process -> ...", in the Review Status column. Please don't miss it!

Volume. 6 Issue. 4

What are the optimum quasi-identifiers to re-identify medical records?

Yong Ju LEE*, Kyung Ho LEE*

*School of Information Security, Korea University, Korea

sky4uni@korea.ac.kr, kevinlee@korea.ac.kr

***Abstract*—Recently, medical records are shared to online for a purpose of medical research and expert opinion. There is a problem with sharing the medical records. If someone knows the subject of the record by using various methods, it can result in an invasion of the patient's privacy. To solve the problem, it is important to carefully address the tradeoff between data sharing and privacy. For this reason, de-identification techniques are applicable to address the problem. However, de-identified data has a risk of re-identification. There are two problems with using de-identification techniques. First, de-identification techniques may damage data utility although it may decrease a risk of re-identification. Second, de-identified data can be re-identified from inference using background knowledge. The objective of this paper is to analyze the probability of re-identification according to inferable quasi-identifiers. We analyzed factors, inferable quasi-identifiers, which can be inferred from background knowledge. Then, we estimated the probability of re-identification from taking advantage of the factors. As a result, we determined the effect of the re-identification according to the type and the range of inferable quasi-identifiers. This paper contributes to a decision on de-identification target and level for protecting patient's privacy through a comparative analysis of the probability of re-identification according to the type and the range of inference.**

***Keyword*—Privacy, Re-identification, De-identification, Medical records**

I. INTRODUCTION

Recently, medical information has been shared to online for many purposes. Especially, it is needed to share patient information for the purpose of medical research and expert opinion [1]. On the other hand, sharing these data may result in an invasion of patient's privacy such as disclosure of diagnostic information via re-identification of medical records.

De-identification techniques have been used to address the problem of privacy and data sharing. In addition, once data is gathered, the conflict arises from two aspects of data use and privacy. It is expected for de-identification techniques to solve the conflict.

However, de-identified data has risk of a re-identification risk, and several studies have proven that it is possible to re-identify data which was de-identified. In addition, although de-identification strengthens the protection of privacy, it could damage the data utility. It can be found the relationship between data utility and disclosure risk like below Figure 1. In Figure 1, X-axis represents the data utility and Y-axis represents the disclosure risk. The disclosure risk of original data is the highest. When the protection techniques like de-identification are applied, the risk will become increasingly lower, however, the data utility will become increasingly lower at the same time [2]. In other words, it is important to find the level with the maximum data utility without exceeding risk threshold.

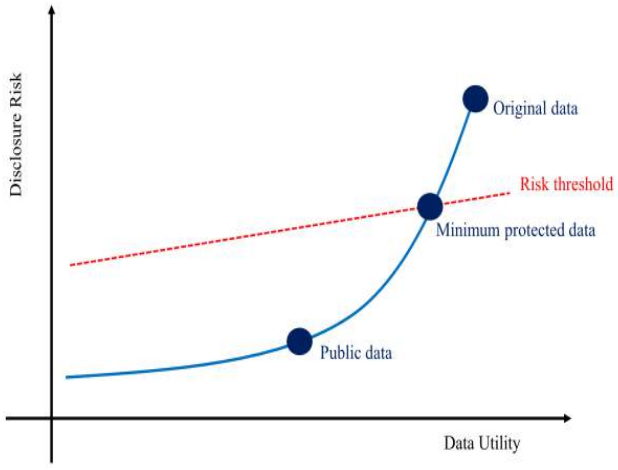


Fig. 1. Data Utility v.s Disclosure Risk

The purpose of this paper is to analyze factors affecting re-identification and to estimate probability of re-identification. From the result, we hope to decide proper de-identification level which can approve safety of personal information protection. To analyze the factors, we researched inference from background knowledge. Next, to estimate the probability of the re-identification, we used de-identified dataset provided in Statewide Planning and Research Cooperative System(SPARCS) of New York state Department of Health [3]. The result of this paper contributes to a decision on de-identification target and level for protecting patient's privacy through a comparative analysis of the probability of re-identification according to the type and the range of inference.

Manuscript received June 27, 2017. This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2015-0-00403) supervised by the IITP(Institute for Information & communications Technology Promotion) .

Kyung Ho LEE is with School of Information Security, Korea University, Seoul, Korea (corresponding author to provide phone: +82-2-3290-4885; e-mail: kevinlee@korea.ac.kr).

Yong Ju LEE was with School of Information Security, Korea University Seoul, Korea (e-mail: sky4uni@korea.ac.kr).

This paper is organized as follows. In chapter 2, we describe related terms, guidelines, de-identification techniques, re-identification research, risk management. In chapter 3, we describe factors affecting re-identification, data set, and probability of a re-identification. In chapter 4, we describe the result of the re-identification simulation. Finally, in chapter 5, we describe meaning of the result, limits of this paper and future work for enhanced research.

II. RELATED WORKS

A. Research on related terms

We describe the definitions about personal information, de-identification, anonymization, and re-identification. First, we describe how to define personal information in the laws from each nation. In USA's Privacy Act, the act defines term 'record' as any item, collection, or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a photograph [4]. Also in Children's online privacy protection Act, the act defines term 'personal information' as individually identifiable information about an individual collected online, including (A) a first and last name, (B) a home or other physical address including street name and name of a city or town, (C) an e-mail address, (D) telephone number, (E) a Social Security number, (F) any other identifier that the Commission determines permits the physical or online contacting of a specific individual or (G) information concerning the child or the parents of that child that the website collects online from the child and combines with an identifier described in this paragraph [5].

In EU Data Protection Directive, the directive defines term 'personal data' as any information relating to an identified or identifiable natural person ('data subject'). And an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity [6]. Since then, General Data Protection Regulation(GDPR) which replaced EU Data Protection Directive appeared. According to the draft of GDPR published in 2012, it defined term 'personal data' as any information relating to a data subject [7]. In the draft, personal data was defined in a broad sense. Since then, according to final version of GDPR published in 2016, it defines term 'personal data' as any information relating to an identified or identifiable natural person ('data subject'). And an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person [8]. As seen in final version of GDPR, the notion of identification was included in defining personal data.

In Canada's Privacy Act, it defines term 'personal information' as information about an identifiable individual that is recorded in any form including, without restricting the generality of the foregoing. The Act divides the term into detailed 13 items such as information relating to the race, any identifying number, the address, the views of another individual about the individual, etc [9]. The scope of the term 'personal information' is more specifically described in Canada.

In Japan's Act on the Protection of Personal Information, it defines term 'personal information' as information about a living individual which can identify the specific individual by name, date of birth or other description contained in such information (including such information as will allow easy reference to other information and will thereby enable the identification of the specific individual) [10].

In Republic of Korea's Personal Information Protection Act, it defines term 'personal information' as information that pertains to a living person, including the full name, resident registration number, images, etc., by which the individual in question can be identified, (including information by which the individual in question cannot be identified but can be identified through simple combination with other information) [11].

So far we have discussed the definition of personal information that is described in the laws and regulations of each country. In most countries, when defining personal information, we know that their definition are based on whether it can identify the individual. On the contrary, if the criteria to identify the individual is ambiguous, there may be some confusion in defining personal information. In other words, the criteria deciding whether the individual can be identified is very important in defining personal information.

Next, we describe de-identification. In ISO/TS 25237:2008(E), it defines term 'de-identification' as general term for any process of removing the association between a set of identifying data and the data subject [12]. And, de-identification makes it hard to learn if the data in a data set is related to a specific individual, while preserving data utility [13].

In ISO/TS 25237:2008(E), it defines term 'anonymization' as process that removes the association between the identifying data set and the data subject [12].

In NISTIR 8053, it defines term 're-identification' as the process of attempting to discern the identities that have been removed from de-identified data [13]. In other words, re-identification occurs when breaking de-identification by identifying an individual who is the subject of the data [14]. Because an important goal of de-identification is to prevent re-identification, re-identification is sometimes called re-identification attack. Meanwhile, re-identification is attempted by various reasons such as testing the quality of the de-identification, gaining publicity or professional standing for performing the re-identification, etc [13]. The reasons are shown in Table 1 below.

TABLE I
THE REASONS FOR ATTEMPTING A RE-IDENTIFICATION

No	Reason
1	To test the quality of the de-identification
2	To gain publicity or professional standing for performing the re-identification
3	To embarrass or harm the organization that performed the de-identification
4	To gain direct benefit from the re-identified data
5	To cause problems such as embarrassment or harm to an individual whose sensitive information can be learned by re-identification

In ISO/TS 25237:2008(E), it explains that anonymization is another subcategory of de-identification. Because anonymization is process that removes the association between the identifying data set and the data subject, re-identification of anonymized data is not possible [12]. Therefore, this paper focuses not on anonymized data but on de-identified data.

B. De-identification guideline

In Australia's Privacy business resource 4 : De-identification of data and information, personal information is 'de-identified' if the information is no longer about an identifiable individual or an individual who is reasonably identifiable [15].

In GDPR, anonymous information is what does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. For this reason, the principles of data protection should not apply to anonymous information [8].

In UK's Anonymisation : managing data protection risk code of practice, it explains the meanings of 'not personal data' as what does not relate to and identify an individual [16].

In Republic of Korea's Personal information de-identification management guideline, de-identified information is personal information de-identified [17]. The Republic of Korea's guideline uses 'de-identification' term. In the guideline, information that is adequately de-identified is presumed to be not personal information since it can no longer identify a specific individual. In this regard, it is semantically explained as an idea of anonymization of the EU.

C. De-identification techniques

The most used method for de-identification are masking, generalization, suppression and adding random noise [18]. These methods can be described in a method of protecting statistical data. We describe the de-identification methods described above.

Masking refers to a set of direct identifier manipulation. In general, direct identifiers are removed or replaced with a random value or specific value from data set. There are redaction, randomization, and pseudonymization techniques in masking method. Redaction is a technique to remove a direct identifier from data set. Randomization is a technique to replace a direct identifier with a random value. pseudonymization is a technique to replace a direct identifier with a unique value [18].

Generalization is a set of anonymization techniques. It reduce an accuracy of data. For example, data '25 years' is generalized to '20-30 years'. In other words, generalization method is constructed from generalizing or diluting an attributes of data subjects by changing a size and scale [19]. There are hierarchy-based generalization and cluster-based generalization techniques in generalization method. Hierarchy-based generalization is based on a predefined hierarchical structure which describes that how much a reduction in an accuracy from quasi-identifier. Cluster-based generalization is based on a predefined utility policy [18].

Suppression means to delete a value of data. There are casewise deletion, quasi-identifier removal, and local cell suppression techniques in suppression method. casewise deletion is a technique to delete all records of a data set. Quasi-identifier removal is a technique to remove only quasi-identifiers of a data set. Local cell suppression, as compared to the above techniques, a more improved technique, is used to find a minimum number of quasi-identifiers required for suppression [18]. When using de-identification techniques, one important issue is data utility problem. In comparison with a casewise deletion and quasi-identifier removal techniques, local cell suppression techniques may be a better way to protect a data utility and reduce a risk of re-identification at the same time.

Adding random noise means to add noise. It may be primarily a de-identification techniques for sensitive items of personal information. This technique uses a method such as any number of addition and multiplication. Because it is added in a range of a specific mean and variance, it has special features that do not damage data utility of data set [20]. That is, it can be used as a method for solving both data utility and privacy problem.

Swapping means to replace database records with a set of predetermined variables [20]. Swapping method reduces a risk of re-identification by introducing an uncertainty of an actual data. Swapping method has an advantage, easy application, generally there is a drawback which does not hold a statistical characteristics [2].

Blank and impute means a method of filing a space portion by applying an alternative after selecting a small number of records from a micro data file, and replacing the selected filed with blank [20]. In other words, it fills blank which comes

TABLE II
THE DE-IDENTIFICATION GUIDELINES

Nation	AU	EU	UK	Republic of Korea
Guideline	Privacy business resource 4 : De-identification of data and information	General Data Protection Regulation (GDPR)	Anonymisation : managing data protection risk code of practice	Personal information de-identification management guideline
Term	de-identified information	anonymous information	not personal data	de-identified information
Definition of de-identified information	no longer about an identifiable individual or an individual who is reasonably identifiable	not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable	not relate to and identify an individual	personal information de-identified
Personal information	X	X	X	△ (presumed to be not personal information. If there is counterevidence, it is regarded as personal information) [17]

from removing original data with a calculated value by using an appropriate function(e.g., average, etc.) [2].

Blurring means a method to replace an average with a value of an item. For example, it is a typical method which replaces an average with a value of an item after classifying specific values into random groups [20].

D. Re-identification type

We confirmed the relationship between data utility and disclosure risk in Figure 1. If de-identification techniques are not sufficiently applied, data utility and disclosure risk will grow. Maybe, this is not public data but original data. On the contrary, data utility and disclosure risk may decline in public data. In other words, some de-identified data can result in a harm. When confidential information about individual such as diagnostic information is identified, disclosure happens. The types of the disclosure are identity disclosure, attribute disclosure, and inferential disclosure [21].

Identity disclosure happens when an attacker identify individual of specific data. The representative scenario which can result in identity disclosure is 're-identification by linking' [13, 21]. This is sometimes called linkage attack.

Attribute disclosure happens when confidential information about individual is identified and can be attributed to a data subject. It is similar to identity disclosure, and identity disclosure can sometimes result in attribute disclosure. However, attribute disclosure can happen without identity disclosure [13, 21].

Inferential disclosure occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of a home. As the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject [22].

In this paper, we focus on inferential disclosure. Because it can result from background knowledge. This is sometimes called background knowledge attack.

E. Re-identification research

We describe studies on re-identification of medical records. According to the research by Latanya Sweeney, she collected the Group Insurance Commission(GIC) data and the voter registration list for Cambridge Massachusetts. Then, she executed the re-identification attack using linkage attack and identified the medical records about William Weld of the total 135,000 records [23]. In another research by Latanya Sweeney, she collected the Washington state de-identified medical records and the online news data. She obtained the information for a patient(e.g., gender, age, hospital, admission month, diagnostic information, address, etc.) from the collected data and identified the 35 records of the total 81 records via a linkage attack [24].

According to the research by Khaled El Emam and Patgricia Kosseim, they collected Pharmacy data from the Children's Hospital of Eastern Ontario. Then, they executed the re-identification attack using background knowledge and identified the 1 record of the total 3,510 records [25]. The research demonstrated that if they had a sufficient background knowledge about a particular individual, it is possible to re-identify medical records. In other words, strong background knowledge for a particular individual can increase the probability of re-identification.

According to the research by Grigorios Loukides, Joshua C Denny and Bradley Malin, they confirmed that it was possible to re-identify the de-identified medical records from the linkage attack based on a diagnosis code. They found that more than 96% of the total 2,762 records were uniquely identified from a diagnosis code and that de-identified medical records may be combined with DNA information via a re-identification attack [26].

Sean Hooley and Latanya Sweeney surveyed every state and the District of Columbia to find what state released about medical records and how much identifiable information were released. They found that 33 states released hospital discharge data. Also it differed from the level that protected the hospital discharge data for each state, and they found that most of 33 states did not meet the HIPAA criteria [27]. So if de-identification techniques are not sufficiently applicable, it may be vulnerable to re-identification attacks.

F. Re-identification risk management

Khaled El Emam and Bradley Malin developed de-identification process. The process consists of 11 steps [28]. The steps of the process are like below Table 3.

Step 1 : determine that which data fields are direct identifiers in data set.

Step 2 : masking methods are applied to the direct identifiers which have been determined in Step 1.

Step 3 : there is two activities. First, we can identify adversaries and what information they may be able to access. Second, we can determine quasi-identifiers in data set.

Step 4 : determine the minimal acceptable data utility.

Step 5 : determine acceptable re-identification risk. This is called re-identification risk threshold.

Step 6 : import data from the origin database.

Step 7 : evaluate the risk of re-identification.

Step 8 : compare the actual re-identification risk with the threshold determined in Step 5.

Step 9 : if the actual re-identification risk is higher than the threshold, it is necessary to apply additional de-identification techniques to data set.

Step 10 : if the actual re-identification risk is lower than the threshold, it is necessary to perform diagnostics on the solution.

Step 11 : export de-identified data to external data set. This is final data.

TABLE III
THE DE-IDENTIFICATION PROCESS

Step	Action
Step 1	Determine direct identifiers in the data set
Step 2	Mask (transform) direct identifiers
Step 3	Perform threat modeling
Step 4	Determine minimal acceptable data utility
Step 5	Determine the re-identification risk threshold
Step 6	Import (sample) data from the source database
Step 7	Evaluate the actual re-identification risk
Step 8	Compare the actual risk with the threshold
Step 9	Set parameters and apply data transformations
Step 10	Perform diagnostics on the solution
Step 11	Export transformed data to external data set

III. METHODS

A. Factors

Since the direct identifier is specific to an individual, the direct identifier is usually removed and quasi-identifier is usually processed through Generalization. It is difficult to re-identify the de-identified data by itself. The risk of re-identification will increase when additional data which can be linked is available [13]. Data sets should be linked for re-identification. To link the data sets, they should include common quasi-identifiers. Meanwhile, the term 'quasi-identifier', first introduced by Tore Dalenius, unlike direct identifiers, can not be identified by itself. However, it allows be linked to an individual who is the subject of the data [29]. That is, quasi-identifier is a variable that allows re-identification via a connection to an individual like below Figure 2.

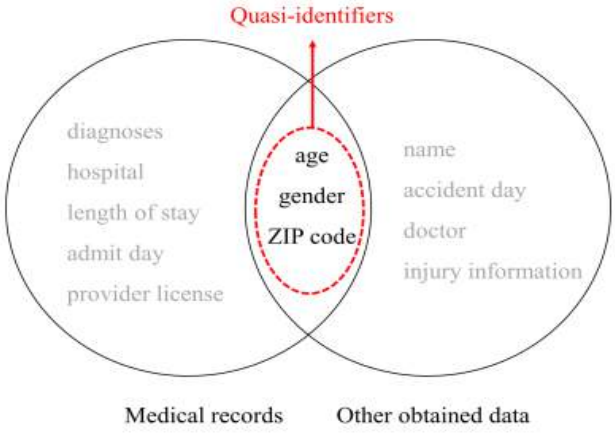


Fig. 2. Linkage attack based on Quasi-identifier

On the other hand, it is possible to infer quasi-identifiers from background knowledge. For example, if someone knew that the diagnosis of a particular patient was a prostate cancer, it can be inferred that the gender of the patient is male. Through the previous research and this study, we drew six factors which affected the probability of re-identification and could be inferred from background knowledge like below Table 4.

TABLE IV
ELEMENTS FOR INFERRING DATA FIELD

Data field	Background knowledge affecting in inferring data field	Study
Zip Code	accident information	[14]
	SNS information (profile, etc.)	[30]
	Geographic information	
Length of Stay	accident information	This study
	admit and discharge day information	
Admit day (of Week)	accident information	[14]
Discharge day (of Week)	discharge information	This study
Diagnosis*	injury information	[14]
	prescription information	
Provider license†	physical information	This study

* We replace name of the data field 'APR MDC Code' with 'Diagnosis' when defining factors.

† We replace name of the data field 'Attending Provider License Number' with 'Provider license' when defining factors.

In the previous research, it inferred admit information and diagnostic information by analyzing the information on an accident [24]. This is possible because online news can contain the information such as accident date and injury contents. In addition to them, in this research, we found that it was possible to infer length of stay, discharge day and provider license from background knowledge. Length of stay can be inferred by using the information about an admission of a patient from accident information. For example, if someone knew the admit day of a particular patient and the publication date of the online news which contained the information, the patient has been hospitalized, it is possible to infer length of stay(strictly, range of length of stay). Also, if someone knew the admit day and the discharge day of a particular patient, it is possible to infer a length of stay by calculating the difference between them. It is possible to infer a discharge day of a particular patient from a publication date of an online news or from a discharge day contained in online news. It is possible to infer a provider license of the physician in charge of treatment of a particular patient. This is possible because the name of the physician in charge of treatment of the patient could be contained in online news, and the provider license could be searched in online site such as profession official database. On the other hand, it is possible to infer Zip Code of a particular patient from accident information. Also, if someone knows SNS account of a particular patient, it is possible to infer Zip Code from account information of the patient. In addition, there is a research which could infer address information from map information [30], it is possible to confirm Zip Code from the address information.

B. Data set and Subject of simulation

To calculate the probability of re-identification, we used the Hospital Inpatient Discharges 2014 (Public Use data) provided in Statewide Planning and Research Cooperative System(SPARCS) of New York state Department of Health. This data set contained de-identified data and could not include protected health information(PHI) according to HIPAA. It included total 2,365,208 records of the hospitalized patients and total 39 fields like below Table 5.

Prior to calculating the probability of re-identification, it was necessary to select the subjects of re-identification simulation. Extraction procedure of the subjects is as follows. Step 1 : Select the most frequently existed 'Facility Name' for each 'APR MDC Code' in the data set. Step 2 : Select the most frequently existed 'Attending Provider License Number' for each 'APR MDC Code' and 'Facility Name' selected Step 1 in the data set. Step 3 : From result of Step 2, total 9,160 records were generated(Table 6).

The reasons for having the extraction procedure of the subjects are as follows. Reason 1 : To include the subjects with a variety of diagnosis. Reason 2 : If the frequency of provider license for each diagnosis was low, it is difficult to analyze how much a provider license impacts on the re-identification.

TABLE V
DATA SET FIELD NAME AND EXAMPLE OF MEDICAL RECORD

No	Field name	Example
1	Health Service Area	New York City
2	accident information	Manhattan
3	Operating Certificate Number	1234567
4	Facility Id	1234
5	Facility Name	New York Hospital
6	Age Group	30 to 49
7	Zip Code - 3 digits	112
8	Gender	F
9	Race	White
10	Ethnicity	Spanish/Hispanic
11	Length of Stay	34
12	Admit Day of Week	WED
13	Type of Admission	Emergency
14	Patient Disposition	Home or Self Care
15	Discharge Year	2014
16	Discharge Day of Week	TUE
…	…	…
23	APR MDC Code	22
24	APR MDC Description	Burns
…	…	…
32	Attending Provider License Number	123456
33	Operating Provider License Number	123456
…	…	…
39	Total Costs	$4,000

TABLE VI
THE SUBJECT OF RE-IDENTIFICATION SIMULATION

No	Facility Name	Age Group	Zip Code	Gender	Length of Stay	Admit Day of Week	Discharge Day of Week	APR MDC Code	Attending Provider License Number
1	New York Hospital	30 to 49	103	F	34	WED	TUE	22	123456
…	…	…	…	…	…	…	…	…	…
9160	…	…	…	…	…	…	…	…	…

C. *Probability of re-identification*

We introduce the formula for measuring the probability of re-identification [31]-[32]. Next, we interpreted the meaning of the probability of re-identification.

$$\theta_j = \frac{1}{f_j}$$

θ refers to the probability of re-identification. f refers to the size of equivalence class. j refers to the number of equivalence class in data set. When f, the size of equivalence class, is minimum value, θ, the probability of re-identification, will be maximum value. Here it is important to find the value of j which makes f be minimized.

Next, we generated the total 64 of the combination of the six factors that affect the re-identification. In other words, j was 1, 2, …, 64. Then, the value of f could be calculated according to the value of each of j. Finally, we calculated the value of θ, the probability of re-identification, which is the inverse of f.

According to the research by Khaled El Emam and Bradley Malin, they introduced 'minimum cell size' concept in determining the threshold of re-identification like below Table 7 [28]. Cell size means the number of response corresponding to a particular condition in data set [20]. Therefore, cell size is seen to have the same concept as the equivalence class. On the other hand, k-anonymity, as one of the privacy protection models, is used to determine whether the propriety of de-identification measures is appropriate in Republic of Korea's Personal information de-identification management guideline. For example, the propriety of de-identification measures is presumed appropriate if the value of k is five for k-anonymity [17].

TABLE VII
MEANING OF IDENTIFIABLE RECORD EACH CELL SIZE

Cell size (Probability)	< 3 (> 0.33)	3 (0.33)	5 (0.2)	11 (0.09)	20 (0.05)
Meaning	Identifiable data	Highly trusted data disclosure	-	-	Highly untrusted data disclosure

In this paper, we had two assumptions for the simulation. First, it was assumed that when the size of equivalence class was 3 or less, the data was identifiable data. Second, it was assumed that when estimating the probability of re-identification, patient information about 'Facility Name', 'Age Group', and 'Gender' was known. They can be sufficiently collected from information such as online news [24], they were excluded from inferable quasi-identifier group we extracted.

IV. RESULTS

We estimated the probability of re-identification by using both prepared data set and previously extracted subject of re-identification simulation. The result of the simulation is shown in Table 8 below. The table shows the probability of re-identification according to the combinations of inferable quasi-identifiers.

Based on the results, if the number of inferable quasi-identifiers was 1, the quasi-identifier which was the most effective factor for re-identification was 'length of stay'. If the number was 2, we knew that the combination of 'length of stay and provider license' was the highest. If the number was 3, the most effective combination was 'length of stay and discharge day and provider license'. If the number was 4 or 5, we knew that the most effective combination included patient's 'zip code'. The most effective combination according to the number of the inferable quasi-identifiers is shown in Table 9 below.

This allows us to know which combination of quasi-identifiers is the most affecting re-identification of medical records. In other words, it helps us decide which quasi-identifier we must de-identify to decrease the probability of re-identification using inference attack through background knowledge.

TABLE VIII
THE PROBABILITY OF RE-IDENTIFICATION AS QUASI-IDENTIFIER COMBINATIONS

No	Combination	Probability (Number of re-identification)	No	Combination	Probability (Number of re-identification)
1	-	0% (0/9160)	33	Zip Code	0.41% (38/9160)
2	Provider license	0.28% (26/9160)	34	Zip Code & Provider license	7.89% (723/9160)
3	Diagnosis	0.01% (1/9160)	35	Zip Code & Diagnosis	2.67% (245/9160)
4	Diagnosis & Provider license	0.67% (61/9160)	36	Zip Code & Diagnosis & Provider license	9.81% (899/9160)
5	Discharge day	0% (0/9160)	37	Zip Code & Discharge day	2.22% (203/9160)
6	Discharge day & Provider license	4.44% (407/9160)	38	Zip Code & Discharge day & Provider license	23.36% (2140/9160)
7	Discharge day & Diagnosis	0.57% (52/9160)	39	Zip Code & Discharge day & Diagnosis	12.47% (1142/9160)
8	Discharge day & Diagnosis & Provider license	7.18% (658/9160)	40	Zip Code & Discharge day & Diagnosis & Provider license	25.67% (2351/9160)
9	Admit day	0% (0/9160)	41	Zip Code & Admit day	2.15% (197/9160)
10	Admit day & Provider license	4.04% (370/9160)	42	Zip Code & Admit day & Provider license	21.53% (1972/9160)
11	Admit day & Diagnosis	0.43% (39/9160)	43	Zip Code & Admit day & Diagnosis	12.05% (1104/9160)
12	Admit day & Diagnosis & Provider license	7.13% (653/9160)	44	Zip Code & Admit day & Diagnosis & Provider license	24.08% (2206/9160)
13	Admit day & Discharge day	0.04% (4/9160)	45	Zip Code & Admit day & Discharge day	9.9% (907/9160)
14	Admit day & Discharge day & Provider license	21.74% (1991/9160)	46	Zip Code & Admit day & Discharge day & Provider license	42.87% (3927/9160)
15	Admit day & Discharge day & Diagnosis	6.84% (627/9160)	47	Zip Code & Admit day & Discharge day & Diagnosis	30.72% (2814/9160)
16	Admit day & Discharge day & Diagnosis & Provider license	25.85% (2368/9160)	48	Zip Code & Admit day & Discharge day & Diagnosis & Provider license	45.94% (4208/9160)
17	Length of Stay	0.71% (65/9160)	49	Zip Code & Length of Stay	5.28% (484/9160)
18	Length of Stay & Provider license	13.1% (1200/9160)	50	Zip Code & Length of Stay & Provider license	28.48% (2609/9160)
19	Length of Stay & Diagnosis	4.67% (428/9160)	51	Zip Code & Length of Stay & Diagnosis	16.89% (1547/9160)
20	Length of Stay & Diagnosis & Provider license	16.1% (1475/9160)	52	Zip Code & Length of Stay & Diagnosis & Provider license	31.12% (2851/9160)
21	Length of Stay & Discharge day	3.46% (317/9160)	53	Zip Code & Length of Stay & Discharge day	16.98% (1555/9160)
22	Length of Stay & Discharge day & Provider license	30.72% (2814/9160)	54	Zip Code & Length of Stay & Discharge day & Provider license	49.16% (4503/9160)
23	Length of Stay & Discharge day & Diagnosis	16.46% (1508/9160)	55	Zip Code & Length of Stay & Discharge day & Diagnosis	39.08% (3580/9160)
24	Length of Stay & Discharge day & Diagnosis & Provider license	33.56% (3074/9160)	56	Zip Code & Length of Stay & Discharge day & Diagnosis & Provider license	51.31% (4700/9160)
25	Length of Stay & Admit day	3.48% (319/9160)	57	Zip Code & Length of Stay & Admit day	17.03% (1560/9160)
26	Length of Stay & Admit day & Provider license	30.53% (2797/9160)	58	Zip Code & Length of Stay & Admit day & Provider license	49.04% (4492/9160)
27	Length of Stay & Admit day & Diagnosis	16.44% (1506/9160)	59	Zip Code & Length of Stay & Admit day & Diagnosis	38.89% (3562/9160)
28	Length of Stay & Admit day & Diagnosis & Provider license	33.55% (3073/9160)	60	Zip Code & Length of Stay & Admit day & Diagnosis & Provider license	51.33% (4702/9160)
29	Length of Stay & Admit day & Discharge day	3.59% (329/9160)	61	Zip Code & Length of Stay & Admit day & Discharge day	17.64% (1616/9160)
30	Length of Stay & Admit day & Discharge day & Provider license	31.36% (2873/9160)	62	Zip Code & Length of Stay & Admit day & Discharge day & Provider license	49.84% (4565/9160)
31	Length of Stay & Admit day & Discharge day & Diagnosis	17.15% (1571/9160)	63	Zip Code & Length of Stay & Admit day & Discharge day & Diagnosis	39.93% (3658/9160)
32	Length of Stay & Admit day & Discharge day & Diagnosis & Provider license	34.24% (3136/9160)	64	Zip Code & Length of Stay & Admit day & Discharge day & Diagnosis & Provider license	52.06% (4769/9160)

TABLE IX
THE MOST EFFECTIVE COMBINATION ACCORDING TO THE NUMBER OF THE QUASI-IDENTIFIERS

Number of the inferable quasi-identifiers	Combination
1	Length of Stay
2	Length of Stay & Provider license
3	Length of Stay & Discharge day & Provider license
4	Zip Code & Length of Stay & Discharge day & Provider license
5	Zip Code & Length of Stay & Admit day & Diagnosis & Provider license

V. CONCLUSION

In this paper, before analyzing solutions to problems related to de-identification, which were data utility and re-identification, we derived optimum quasi-identifiers which have the greatest impact on re-identification of medical records. We analyzed the factors affecting re-identification and estimated the probability of re-identification based on extracted factors by using a de-identified data set. The factors were 'Zip Code', 'Length of Stay', 'Admit day', 'Discharge day', 'Diagnosis', and 'Provider license'. Especially, compared with the previous paper, we added 'Zip Code' factor affecting re-identification. As a result, we found 'Zip Code' factor had a greater impact on re-identification than the other factors when the number of inferable quasi-identifiers was more than four.

We simulated the re-identification of medical records by using the Hospital Inpatient Discharges 2014 (Public Use data) provided in SPARCS of New York state Department of Health. From the results of the simulation, we found that the probability of re-identification was depending on the type of inferable quasi-identifier. In other words, the probability of re-identification would be either higher or lower according to the type of quasi-identifier inferred. This allows us to find the optimum quasi-identifiers for re-identification of medical records. But at the same time, this shows what we prevent from being inferred to decrease the probability of re-identification. Although it is hard to completely block the inference of information related to patient, it will be possible to decrease the probability of re-identification by means such as increasing de-identification level.

On the other hand, we describe two limitations of this paper. First, the number of inferable quasi-identifiers for re-identification of medical records may be more than six presented in this paper. Second, although this paper shows that the probability of inference is either 0 or 1, the probability of inference may actually be various. For example, the probability of inference may have various value because of variables such as the amount of collected background knowledge, the characteristics of the quasi-identifier, etc.

In order to overcome the limitations presented above, the research on extending the range of inferable quasi-identifiers and estimating the probability of inference should be done in future works.

ACKNOWLEDGMENT

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2015-0-00403) supervised by the IITP(Institute for Information & communications Technology Promotion). This paper is an extension of the Yong Ju LEE's Master's thesis and ICACT2017 conference proceeding. Thanks to the Master's thesis reviewers, who were Hun Yeong KWON, In Seok KIM and Kyung Ho LEE from School of Information Security, Korea University and the ICACT2017 reviewers.

REFERENCES

[1] H Taneja, AK Singh. "Preserving Privacy of Patients Based on Re-identification Risk," *Procedia Computer Science* 2015; 70: 448-454.
[2] V Ciriani, SDC Di Vimercati, S Foresti, P Samarati. "Microdata protection," In: *Secure data management in decentralized systems* 2007; 33: 291-321.
[3] New York State Department of Health. Hospital Inpatient Discharges (SPARCS De-Identified): 2014. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/rmwa-zns4.
[4] 5 U.S.C. §552a.
[5] 15 U.S.C. §6501.
[6] Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
[7] European Commission. "Proposal for a Regulation of the european parliament and of the council, on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012PC0011&from=EN.
[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[9] R.S.C., 1985, c. P-21 Privacy Act.
[10] Act on the Protection of Personal Information.
[11] Personal Information Protection Act.
[12] International Organization for Standardization. ISO/TS 25237:2008(E) Health Informatics — Pseudonymization.
[13] National Institute of Standards and Technology. "NISTIR 8053 De-Identification of Personal Information," [Online]. Available: http://dx.doi.org/10.6028/NIST.IR.8053.
[14] L Sweeney. "Only You, Your Doctor, and Many Others May Know," *Technology Science* 2015; 2015092903.
[15] Office of the Australian Information Commissioner. Privacy business resource 4: De-identification of data and information.
[16] Information Commissioner's Office. Anonymisation : managing data protection risk code of practice.
[17] Office for Government Policy Coordination. Personal information de-identification management guideline.
[18] K El Emam. "Methods for the de-identification of electronic health records for genomic research," *Genome medicine* 2011; 3.4: 1.
[19] ARTICLE 29 DATA PROTECTION WORKING PARTY. "Opinion 05/2014 on Anonymisation Techniques," [Online]. Available: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
[20] Park WH, HWANG JY. "Disclosure Limitation Techniques for Statistical Tables and Microdata," *Journal of The Korean Official Statistics* 2004; 9.2: 146-172.
[21] L. Xiong, J. Gardner, P. Jurczyk, J. J. Lu. "Privacy-Preserving Information Discovery on EHRs," in *Information Discovery on Electronic Health Records*. CRC Press 2009.
[22] Glossary of statistical terms, OECD. [Online]. Available : https://stats.oecd.org/glossary/detail.asp?ID=6932 .
[23] L Sweeney. "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002; 10.05: 557-570.
[24] L Sweeney. "Matching known patients to health records in Washington State data," *Available at SSRN 2289850* 2013.

[25] K El Emam, P Kosseim. "Privacy interests in prescription data, part 2: patient privacy," *IEEE Security & Privacy* 2009; 7.2: 75-78.
[26] G Loukides, JC Denny, B Malin. "The disclosure of diagnosis codes can breach research participants' privacy," *Journal of the American Medical Informatics Association* 2010; 17.3: 322-327.
[27] S Hooley, L Sweeney. "Survey of Publicly Available State Health Databases," *Available at SSRN 2277688* 2013.
[28] K El Emam, B Malin. "CONCEPTS AND METHODS FOR DE-IDENTIFYING CLINICAL TRIAL DATA," *Paper commissioned by the Committee on Strategies for Responsible Sharing of Clinical Trial Data* 2014.
[29] T Daleniusl. "Finding a needle in a haystack," *Journal of official statistics* 1986; 2.3: 329-336.
[30] Brownstein, John S., Christopher A. Cassa, and Kenneth D. Mandl. "No place to hide—reverse identification of patients from published maps," *New England Journal of Medicine* 2006; 355.16: 1741-1742.
[31] K El Emam, FK Dankar, R Vaillancourt, T Roffey, M Lysyk. "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records," *The Canadian journal of hospital pharmacy* 2009; 62.4: 307-319.
[32] K El Emam. "Guide to the de-identification of personal health information," *CRC Press* 2013.

Yong Ju LEE , was born in Republic of Korea, October 7, 1989. Yong Ju Lee earned Master's degree from School of Information Security at Korea University. His main research interests include risk management, privacy policy, de-identification and re-identification of personal information.

Kyung Ho LEE , was born in Republic of Korea, September 9, 1967. Kyung Ho Lee earned his Ph.D. degree from Korea University. He is now a Professor in School of Information Security at Korea University, and leading the Risk management Laboratory in Korea University since 2011. He was the former CISO in Naver corporation and CEO of Secubase corporation. His main research interests include information security management system(ISMS), risk management, information security consulting, privacy policy, and privacy impact assessment(PIA).

A Cooperative Spectrum Sensing Algorithm Using Leading Eigenvector Matching

Yuhui SONG

China FAW Group Corporation R&D Center, Changchun, China

syhhit@yeah.net

***Abstract*—Cognitive radio emerged as a new trend to mitigate the severe spectrum scarcity problem. As an essential problem in cognitive radio, spectrum sensing has been discussed widely recently. Blind detection techniques that sense the presence of a primary user's signal without prior knowledge of the signal characteristics, channel and noise power attract more attention than non-blind detection. The sensing algorithms based on random matrix theory which are shown to outperform energy detection especially in case of noise uncertainty. In this paper, a sensing algorithm using leading eigenvector matching (LEM) is introduced into cooperative spectrum sensing process. LEM detector uses the feature blindly learned from feature learning algorithm (FLA) as prior knowledge. The LEM algorithm involves the correlation coefficient between feature learned and leading eigenvector of sample covariance matrix as the test statistic. In this paper, we also derive the closed-form expression of the threshold in order to achieve constant false alarm rate detection. Numerical simulations show that the proposed detection algorithm performs better than the MME detector and it does not suffer from a noise power uncertainty problem while also proving to be more robust against the correlation decrease between sensing nodes.**

***Keyword*—Cognitive Radio, spectrum sensing, sample covariance matrix, leading eigenvector matching, feature learning.**

I. INTRODUCTION

COGNITIVE radio (CR) differs from conventional radio systems and is considered as an effective method to mitigate the spectrum scarcity problem. In CR, cognitive user (CU) is aware of the electromagnetic environment around it and accesses the spectrum underutilized by primary user's (PU) accordingly [1]. Spectrum sensing is an essential problem in CR which can detect the PU's signal presence and it has been widely discussed in recent decade [2]. It is simple to detect signal when the signal to noise (SNR) is high, but in practice sensing the presence of PU's signal becomes demanding because of the low SNR and shadow fading. Spectrum sensing algorithms existed can be divided into non-blind detector and blind detector according to whether it requires prior knowledge about the signal and the channel characteristics. Non-blind techniques (such as matched filter detection and cyclostationary feature detection [3]) that rely on prior knowledge give a better performance, but it is difficult to acquire prior knowledge in practice. On the other hand, blind sensing (e.g. energy detection) that do not require prior knowledge is flexible in their application.

Aforementioned detection algorithms are single-node sensing methods whose performances fall down quickly because of the multipath fading and hidden terminal problems, so cooperative spectrum sensing algorithms attract more attention. The cooperative spectrum sensing algorithms based on random matrix theory (RMT) were shown to outperform classical methods as a blind detector, especially in case of noise uncertainty which is the main disadvantage of energy detection. Most of the algorithms based on RMT utilize the differences between the distributions of eigenvalues of sample covariance matrix under H_0 and H_1 , including maximum and minimum eigenvalue (MME) [4], energy with minimum eigenvalue (EME) [5], maximum eigenvalue detection (MED) [6]. However, the algorithms based on RMT suffer from the correlation problem, i.e., its perceived performance decreases quickly when the correlation between sensing nodes decreases.

Besides eigenvalues, eigenvector is another characteristic of the covariance matrix. Feature template matching (FTM) [7] has been proposed as a single-node spectrum sensing technique which is based on the leading eigenvector and shown that it performs better than MME and the covariance absolute value (CAV) algorithms. Multiple feature matching (MFM) [8] algorithm applied the FTM algorithm in MIMO system. But the methods above did not derive the closed-form expression of the threshold and were limited in single-node spectrum sensing.

In this paper, a cooperative spectrum sensing algorithm using leading eigenvector matching (LEM) is introduced. LEM detector uses the feature blindly learned from feature learning algorithm (FLA) as prior knowledge. The correlation coefficient between feature learned and the leading eigenvector of sample covariance matrix serves as the test statistic for signal detection. The closed-form expression of the threshold is also derived in this paper. Simulation results show that the algorithm proposed is reasonable and LEM detector outperforms MME detector. It also do not suffer from a noise power uncertainty problem. Compared with MME detector, LEM detector is more robust against the decrease of correlation among the sensing nodes.

The rest of the paper is organized as follows: Sec. II

Manuscript received on April 18th, 2017. This work is a follow up of the invited journal of the accepted conference paper for the 19th International Conference on Advanced Communication Technology (ICACT-20170180).

Y. SONG is with China FAW Group Corporation R&D Center, Changchun, China ((corresponding author to provide phone: 86+15643078016; fax: 86+0431-88687522; e-mail: syhhit@yeah.net).

reviews the sensing model and basic theory of the proposed algorithm. Sec. III deals with the description of LEM detector and the threshold derivation problem. Simulation results are presented and discussed in Sec. IV. Sec. V contains the conclusions.

II. System Model and Basic Theory

A. System Model

In the CR network, there is one PU and K CUs. Denote with $x_i(n)$ the n^{th} sample received by the i^{th} PU. There are two hypotheses and H_0 indicates that the PU's signal does not exist and H_1 denotes the signal exists. The received signal samples under two hypotheses show as follows:

$$x_i(n)=\begin{cases} w_i(n) & H_0 \quad i=1,2,...,K \\ s_i(n)+w_i(n) & H_1 \quad i=1,2,...,K \end{cases} \tag{1}$$

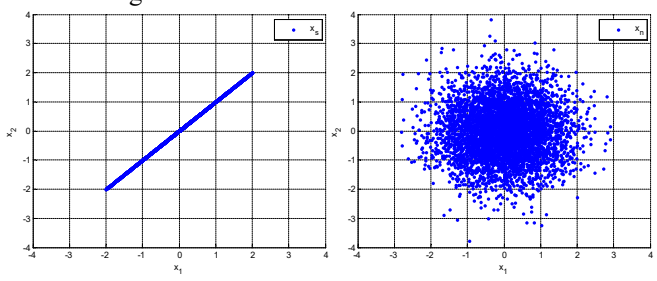
where $s_i(n)$ is the PU's signal received and $w_i(n)$ is the white Gaussian noise (WGN) with zero mean and variance σ^2. Let $X_i(n)=[x_i(n)\ x_i(n+1)\ ...\ x_i(n+N\text{-}1)]$ be a $1\times N$ vector containing N consecutive samples collected by the i^{th} CU. The j^{th} sensing segment constructed by the samples received is written as:

$$X_j=\begin{bmatrix} X_1(j) \\ X_2(j) \\ \vdots \\ X_K(j) \end{bmatrix}=\begin{bmatrix} x_1(j) & x_1(j+1) & \dots & x_1(j+N-1) \\ x_2(j) & x_2(j+1) & \dots & x_2(j+N-1) \\ \vdots & \vdots & \ddots & \vdots \\ x_K(j) & x_K(j+1) & \dots & x_K(j+N-1) \end{bmatrix} \tag{2}$$

The sample covariance matrix is $R_j=\frac{1}{N}X_jX_j^T$ and its leading eigenvector is φ_j.

B. Basic Theory

x_s indicates a 2×1 amplitude modulation (AM) signal vector and x_n indicates a 2×1 WGN vector. Denote with x_{s+n} a 2×1 AM signal with WGN vector, which means $x_{s+n}=x_s+x_n$. The elements of the vector are x_1 and x_2 respectively. Fig. 1 shows that the distribution of WGN is random, while x_s has the same characteristic angle with x_{s+n}, which means that their leading eigenvector is similar. The algorithm proposed uses the characteristic to distinguish signal from noise. The similar situation in three-dimension is shown in Fig. 2.



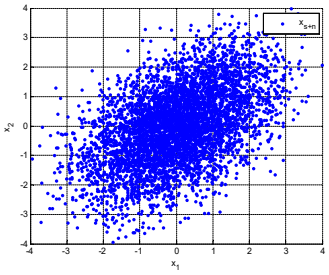


Fig. 1. The distribution under two-dimension

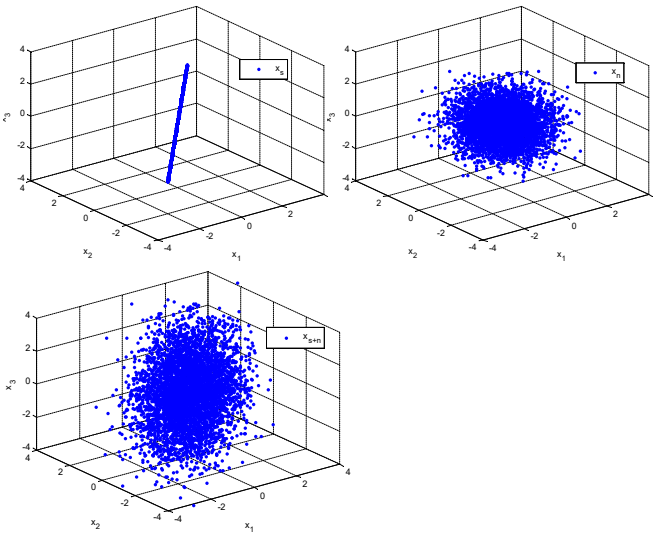


Fig. 2. The distribution under three-dimension

Mathematical theorem about the sensitivity of eigenvectors also explain the characteristic above, that is, the sensitivity of the eigenvector depends on the separation between the corresponding eigenvalue and other eigenvalues.

Under H_0, the sample covariance matrix approximates the diagonal matrix σ^2E. The leading eigenvector is sensitive and random because the maximum eigenvalue equals to other eigenvalues, which means the similarity between two leading eigenvector of two sample covariance matrices is low. Instead, the leading eigenvector remains stable under H_1 because the maximum eigenvalue is much larger than other eigenvalues and the similarity is high.

According to the definition of the eigenvalues and eigenvectors, the relation among sample covariance matrix R, its maximum eigenvalue λ and leading eigenvector I can be written as the equation:

$$RI=\lambda I \tag{3}$$

While noise exists, it can be expressed as:

$$(R+\sigma^2E)I=(\lambda+\sigma^2)I \tag{4}$$

It is shown that the leading eigenvector I of matrix R is also the leading eigenvector of $R+\sigma^2E$. Therefore, I remains stable regardless of the change of noise variance, which is more robust against noise.

III. Detection Algorithm

A. Leading Eigenvector Matching Algorithm

Leading eigenvector is also called signal feature in pattern recognition and it has the greatest mutual information with

original signal. Compared with the randomness of leading eigenvector of WGN, the leading eigenvector of WGN is more stable. If PU's signal exists, highly correlated leading eigenvector can be detected in consecutive sensing segments X_j and X_{j+1}. Due to the robustness of signal feature, it can be learned by blind FLA.

The feature φ_s can be learned blindly from J sensing segments by following steps:

1) Extract feature φ_j and φ_{j+1} from X_j and X_{j+1};
2) Compute correlation coefficient via cosine similarity formula:

$$\rho_{j,j+1}=\left|\frac{\left\langle\varphi_j,\varphi_{j+1}\right\rangle}{\left|\varphi_j\right|\left|\varphi_{j+1}\right|}\right|=\left|\varphi_j^T\varphi_{j+1}\right| \tag{5}$$

3) J-1 correlation coefficients can be calculated from J sensing segments. If $\rho_{m,m+1}=\max_{j=1,2,\ldots,J-1}\{\rho_{j,j+1}\}$, signal feature φ_s is learned as φ_{m+1}.

With the prior knowledge φ_s, we have LEM detector:

1) Compute the received signal sample covariance matrix $X_{current}$ and corresponding leading eigenvector $\varphi_{current}$;
2) Compute correlation coefficient $\rho_{s,current}$ between φ_s and $\varphi_{current}$;
3) H_1 is true if $\rho_{s,current}>\varepsilon$, where ε is the threshold determined by desired P_f.

Compared with MME detector, both of the algorithms need to solve eigenvector and eigenvalue problem and their time complexity is almost same. But the LEM detector need to learn the feature φ_s by FLA which requires extra computation and time. In practical applications, the feature can be learned ahead and stored in local fusion center memory.

B. Threshold

It is necessary to obtain the expression of the false-alarm probability. On the one hand, the false-alarm probability can be used to illustrate the performance of the detection algorithm. On the other hand, the threshold can be obtained by given target false-alarm probability. In the proposed algorithm, the probability of false alarm is defined as:

$$P_f=p(\left|\varphi_s^T\varphi_{current}\right|>\varepsilon\left|H_0\right.) \tag{6}$$

Under H_0, we have the following result:

$$R_s,R_{current}\sim\frac{1}{S}wishart(S,\sigma^2I) \tag{7}$$

where $wishart(\bullet)$ is Wishart distribution. Let φ_s and $\varphi_{current}$ be the leading eigenvectors of R_s and $R_{current}$, and the normalized covariance matrices are defined as:

$$\begin{aligned}&C_s=\frac{1}{\sigma^2}R_s,C_{current}=\frac{1}{\sigma^2}R_{current}\\&C_s,C_{current}\sim\frac{1}{S}wishart(S,I)\end{aligned} \tag{8}$$

Let A and B be the matrices containing the normalized eigenvectors of C_s and $C_{current}$ respectively. We have $A=[a_1,a_2\cdots a_K]$, $B=[b_1,b_2\cdots b_K]$ where the eigenvectors are arranged in descending order. And the matrices A^T and B^T converge in distribution to Haar [9]. It is known that A, B and B^T are unitary matrices and we have following result:

$$f(A^TB)=f(A^T(B)^T)=f(A^T) \tag{9}$$

where $f(\bullet)$ is the Probability Density Function. Because A^TB and A^T converge in the same distribution, the elements of the matrices also converge in the same distribution and we have following result:

$$f(\left|a_1^Tb_1\right|)==f(\left|a_{11}\right|) \tag{10}$$

where a_{11} is the element at the first row and the first column in matrix A.

According to the property of unitary matrix, it can be known that a_{11}^2 converges in distribution to Beta with parameters $\alpha=\frac{1}{2}$ and $\beta=\frac{K-1}{2}$. The probability density function of $T=\left|a_1^Tb_1\right|=\left|\varphi_s^T\varphi_{current}\right|$ can be written as:

$$f(T)=\frac{\Gamma(\frac{K}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{K-1}{2})}\frac{(1-T^2)^{(K-3)/2}}{T} \tag{11}$$

And we have the expression of false-alarm probability as follows:

$$\begin{aligned}P_f&=p(\left|\varphi_s^T\varphi_{current}\right|>\varepsilon\left|H_0\right.)\\&=\frac{\Gamma(\frac{K}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{K-1}{2})}\int_\varepsilon^1\frac{(1-T^2)^{(K-3)/2}}{T}dT\end{aligned} \tag{12}$$

Hence, we derive the decision threshold as a function of the false-alarm probability:

$$\varepsilon=\sqrt{F_{Beta}^{-1}(1-P_f,\frac{1}{2},\frac{K-1}{2})} \tag{13}$$

where $F_{Beta}^{-1}(\cdot)$ is the inverse cumulative distribution function of Beta distribution. It is shown that the threshold is a function of the target P_f and the dimension of the covariance matrix K. So the algorithm proposed in this paper is a blind sensing algorithm that do not require any prior knowledge about the signal and the channel characteristics.

IV. Simulation

In this section, we present the simulation results to evaluate the performance of the proposed algorithm. The PU's signal is the AM signal whose carrier frequency is 702 KHz and the sample ratio is 4MHz. The numbers of the CUs and samples are 32 and 1000 respectively and the SNR is -20dB. The simulation results are obtained by 10,000 Monte Carlo trials.

A. Distribution of the test statistic

Fig. 3 shows the frequency distribution of the test statistic under two hypotheses, i.e., H_0 and H_1. It is shown that most of the test statistic under H_0 is less than 0.4 while the majority of the test statistic under H_1 is greater than 0.4. It is obvious that the test statistic under H_0 and H_1 can be separated well by a given threshold, e.g., 0.45, in this situation.

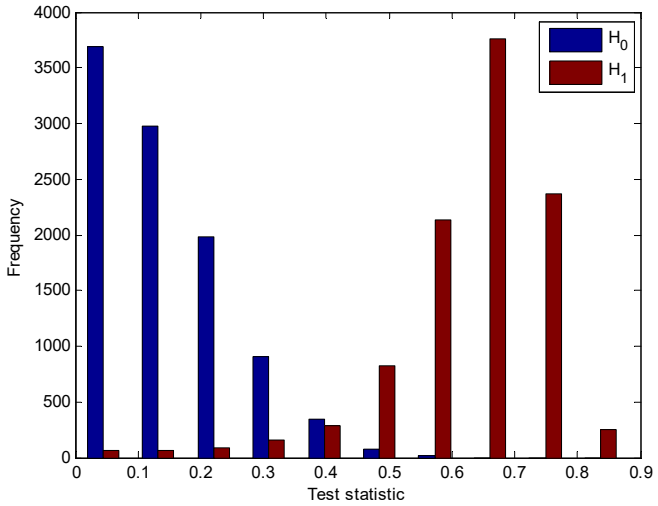


Fig. 3. Frequency distribution of the test statistic under H_0 and H_1

Fig. 4 presents the estimated and empirical cumulative distribution function (CDF) of the test statistic under H_0 respectively with different K. The accuracy of the estimated CDF determines the accuracy of the threshold to achieve target false-alarm probability. It is shown that the estimated CDF matches well with the empirical CDF with different K.

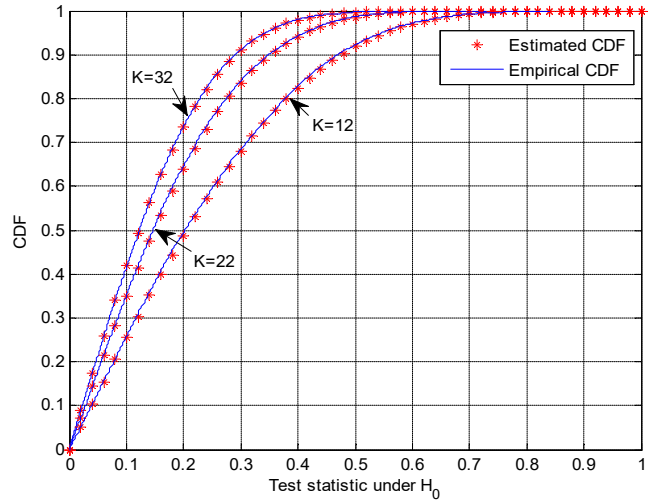


Fig. 4. Estimated and empirical CDF of the test statistic under H_0

B. Comparison of ROC for MME and LEM

MME algorithm is a classical cooperative spectrum sensing method based on the random matrix theory. Denote $\lambda_{\max}$ and $\lambda_{\min}$ with maximum and minimum eigenvalues of sample covariance matrix respectively. The ratio between maximum and minimum eigenvalues $\lambda_{\max}/\lambda_{\min}$ is used to be the test statistic. We will compare MME detector with LEM detector from several aspects.

Receiver operating characteristic (ROC) curve is an essential graphical plot that illustrates the performance of a binary classifier system. Fig. 5 shows the ROC's comparison between MME detector and LEM detector and the latter has a better performance obviously compared with the former. Generally speaking, the spectrum sensing algorithm need to achieve constant false alarm rate (CFAR) detection and according to 802.22 working group, the target false-alarm probability in the CR is required to be less than 10%. As shown in the figure, when the false-alarm probability is 1%, the detection probability for MME detector and LEM detector is 64% and 92% respectively. Because it is unnecessary to estimate the noise power to obtain the threshold, both the MME detector and LEM detector do not suffer from a noise uncertainty problem which is the main disadvantage of the energy detector.

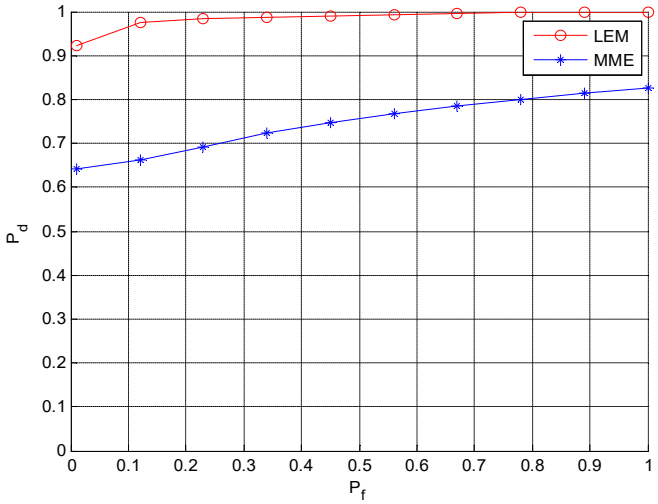


Fig. 5. ROC for MME and LEM

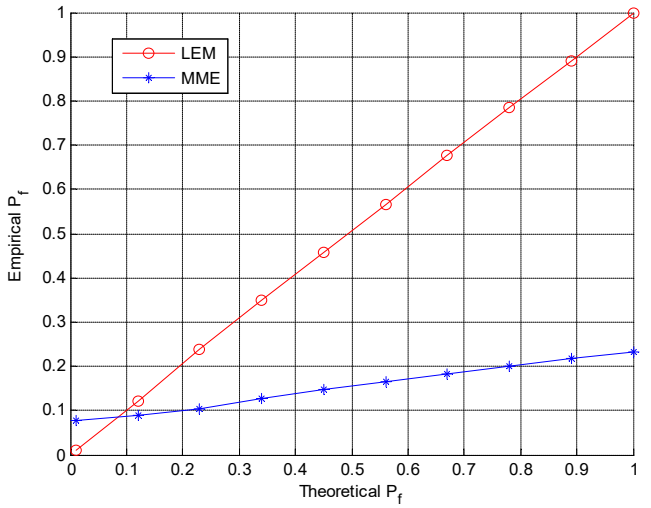


Fig. 6. Empirical P_f vs. theoretical P_f for MME and LEM

The abscissa coordinate of Fig. 6 is the theoretical false-alarm probability and the vertical coordinate is the empirical false-alarm probability. It shows that the curve of LEM algorithm is roughly diagonal, that is, the theoretical false-alarm probability is approximately equal to the theoretical false-alarm probability, which proves the correctness of the decision threshold derivation. As for MME algorithm, the empirical false-alarm probability deviates from the theoretical false-alarm probability significantly. MME

detector is based on the asymptotic random matrix theory that requires the dimension of the matrix is infinite and it is impossible in practice, which leads to the deviation of the threshold.

C. Comparison of performance under different SNR

Fig. 7 represents that the detection probability for MME detector decreases sharply with the decrease of SNR while the detection probability for LEM detector decreases slowly. When the SNR is -20dB, the detection probability for MME detector and LEM detector is 65% and 95% respectively.

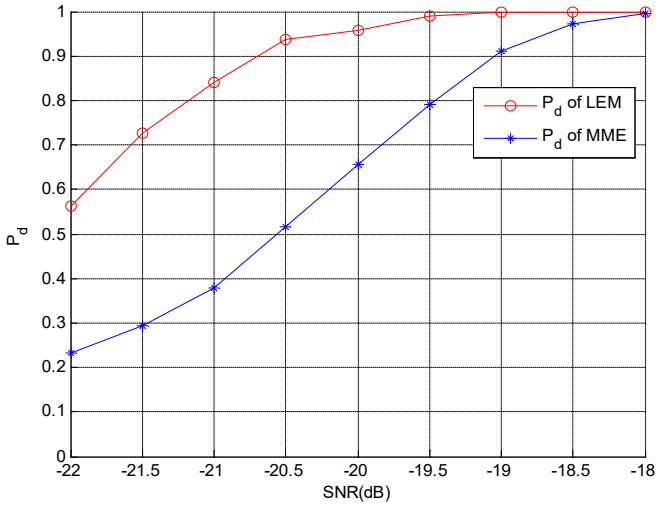


Fig. 7. P_d vs. SNR at $P_f = 5\%$ for MME and LEM

In Fig. 8, we investigate the probability of false alarm versus SNR. The false-alarm probability for LEM detector reaches 5% and fluctuates slightly around it. The false-alarm probability for MME detector is slightly higher than that for LEM algorithm because the threshold derivation of LEM algorithm is more accurate.

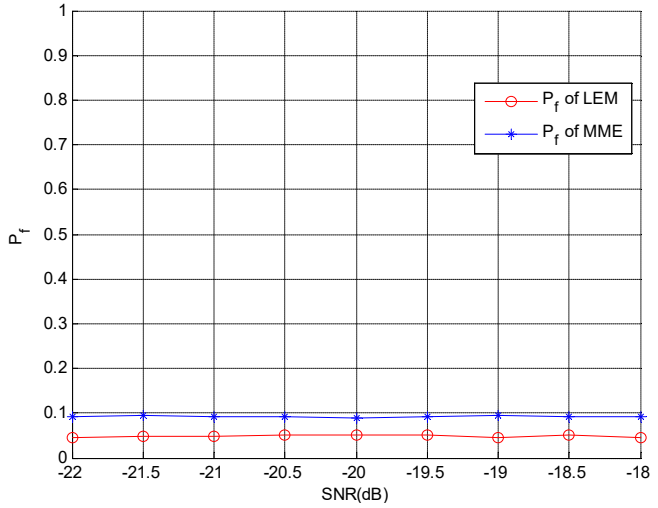


Fig. 8. P_f vs. SNR at $P_f = 5\%$ for MME and LEM

D. The impact of the correlation between sensing nodes

Reference [3] points out that MME detector requires the signal of sensing nodes are highly correlated, otherwise the detection probability falls down quickly. Due to its simplicity and flexibility, exponential model is widely adopted to describe correlation. ρ is the correlation coefficient between two sensing nodes that is related to the angular spread, wavelength and the distance between two nodes. Fig. 9 represents the detection probability for MME detector and LEM detector under different correlation coefficient. It is shown that the detection probability of MME approximates to zero as $\rho = 0.95$ while the probability of detection for LEM decreases a little in the same situation, which means the LEM algorithm is more robust against the decrease of correlation among the sensing nodes.

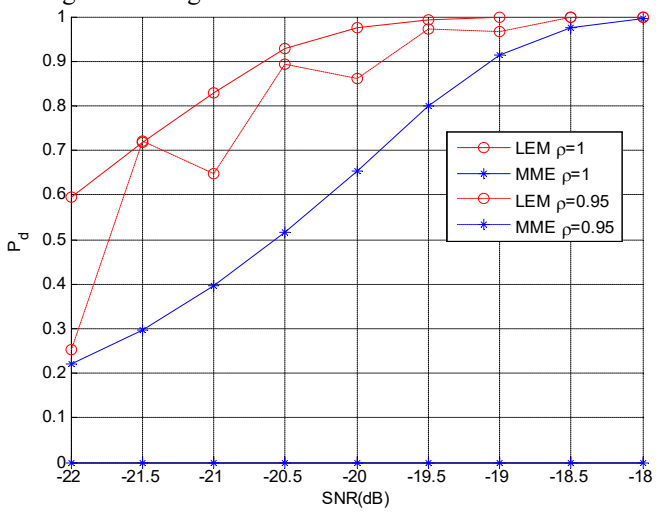


Fig. 9. P_d vs. SNR at $P_f = 5\%$ with different correlation

E. The impact of the parameter K and N

The detection probability for LEM detector with various numbers of CUs K and samples N is shown in Fig. 7 and Fig. 8 respectively. The simulation results show that the number of cooperative CUs K and the sample size N play the similar role in detection performance. It is obvious that the detection probability varies with the number of sensing nodes K proportionally. In addition, the probability of detection increases as the number of samples N grows.

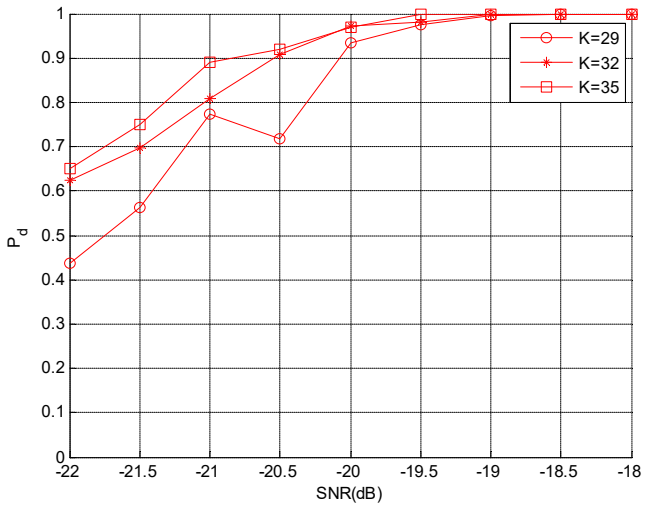


Fig. 10. P_d vs. SNR at $P_f = 5\%$ for LEM with different K

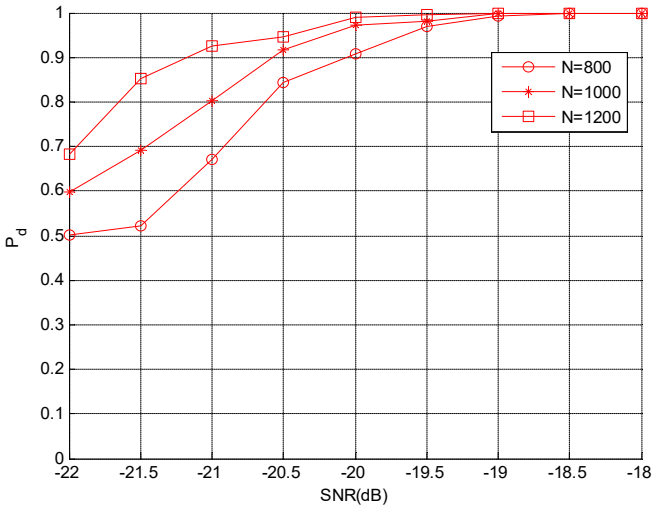


Fig. 11. P_d vs. SNR at P_f = 5% for LEM with different N

V. Conclusions

In this paper, a cooperative spectrum sensing algorithm using leading eigenvector matching is introduced. While PU's signal does not exist, the leading eigenvector is random. But when the signal is present, the leading eigenvector is stable. Due to its robustness, the feature can be learned blindly by FLA and LEM detector uses the feature as prior knowledge. The correlation coefficient between feature learned and the leading eigenvector of sample covariance matrix serves as the test statistic for signal detection. The closed-form expression of the threshold is also derived in this paper. Simulation results show that the algorithm proposed is reasonable and LEM detector outperforms MME detector. It also do not suffer from a noise power uncertainty problem. Compared with MME detector, LEM detector is more robust against the decrease of correlation among the sensing nodes. However there are some inherent flaws in this approach. A feature can only be learned in the presence of the desired PU signal, it cannot be learned in the presence of noise or in the presence of any other signal.

References

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," in *IEEE Personal Communications*, vol. 6, no. 4, pp. 13-18, Aug 1999.

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," in *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201-220, Feb. 2005.

[3] D. Cabric, S. M. Mishra and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.*, 2004, pp. 772-776 Vol.1.

[4] F. Penna, R. Garello and M. A. Spirito, "Cooperative spectrum sensing based on the limiting eigenvalue ratio distribution in wishart matrices," in *IEEE Communications Letters*, vol. 13, no. 7, pp. 507-509, July 2009.

[5] Y. Zeng and Y. C. Liang, "Eigenvalue-based spectrum sensing algorithms for cognitive radio," in *IEEE Transactions on Communications*, vol. 57, no. 6, pp. 1784-1793, June 2009.

[6] Y. C. Liang, G. Pan and Y. Zeng, "On the Performance of Spectrum Sensing Algorithms Using Multiple Antennas," *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, Miami, FL, 2010, pp. 1-5.

[7] P. Zhang, R. Qiu and N. Guo, "Demonstration of Spectrum Sensing with Blindly Learned Features," in *IEEE Communications Letters*, vol. 15, no. 5, pp. 548-550, May 2011.

[8] F. A. Bhatti, G. B. Rowe and K. W. Sowerby, "Spectrum sensing using feature vectors," *2012 IEEE International Conference on Communication Systems (ICCS)*, Singapore, 2012, pp. 448-452.

[9] Li Yingxue, Lei Jing, Zhong Shiyuan, Huangchunming, Huangchao. Covariance Blind Detection Method Based on Eigenvector in Cognitive Radio Networks [J]. Telecommunications Science, 2015, 31(11): 2015306.

Yuhui SONG was born in Changchun city, China, on March 15, 1993. He received the B.S. degree in electronic and information engineering from Harbin Engineering University, Harbin, China, in 2015. He received the M.S. degree in electronic and communication engineering from Harbin Institute of Technology, Harbin, China, in 2017. He is currently an engineer in China FAW Group Corporation R&D Center, Changchun, China. His research interests include spectrum sensing in cognitive radio.

Evolving Neural Network Intrusion Detection System for MCPS

Nishat Mowla*, Inshil Doh**, KiJoon Chae*

*Department of Computer Science and Engineering

**Department of Cyber Security

Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 120750, Korea

nishat.i.mowla@gmail.com, isdoh1@ewha.ac.kr, kjchae@ewha.ac.kr

***Abstract*— Medical Cyber Physical Systems (MCPS) are some of the most promising next generation technologies so far. Like many other systems connected to a wider network such as internet, MCPS are also vulnerable to various forms of network attacks. For detecting such diverse forms of attack, we need smart and efficient mechanisms. Human intelligence is good enough to track such attacks but when it is a huge number of traffic it is no more a feasible process to detect them manually as it is time consuming and computationally intensive. Machine learning techniques embracing artificial intelligence are emerging as powerful tools to detect abnormalities in the network data. Supervised Neural Networks are some of the most efficient techniques to perform such classification. In this paper, we propose an evolving neural network technique that evolves based on classification, elimination and prioritization while focusing on time, space and accuracy to efficiently classify the four major types of network attack traffic found in an effectively pruned KDD dataset. We also show a leap of performance with hyper-parameter optimization which highly enhances the benefit of our proposed mechanism. Finally, the new performance gain is compared with a boosted Decision Tree. We believe our proposed mechanism can be adopted to new forms of attack categories and sub-categories.**

***Keyword*— MCPS, Machine Learning, Neural Networks, Intrusion Detection System**

I. INTRODUCTION

THIS is an era of various body worn devices that can record multiple physiological signals, such as ECG and heart rate or even more sophisticated devices that measure physiological markers such as body temperature, skin resistance, gait, posture, and EMG. Medical Cyber Physical Systems are the much-promised technologies which aim at providing remote healthcare to patients using the sensor information collected from such body worn devices [14]. With great prospect come great responsibilities. The data collected from these devices can be stored in a public or private cloud to be later analysed by the hospital authorities. Therefore, assuring the accessibility of the personal health information during the transmission from the sensory networks to the cloud and from the cloud to doctors' mobile devices will necessitate the design of an intelligent malicious traffic detection system which would prevent normal traffic from getting the proper connection [15].

Machine learning classification techniques are popular when it comes to the issue of classifying normal from abnormal. Among them recently deep learning techniques such as Neural Network are shown to act as powerful tools in order to classify various forms of network attack exploits.

In this paper, we propose an evolving neural network based intrusion detection system for detecting the four key major forms of network attack types by evolving the multi-class data to a 2-class problem following classification based data pruning and class prioritization.

We discuss some of the related works in sections II. In section III we discuss our proposed mechanism. Section IV shows our performance evaluation results followed by a discussion of our proposed mechanism in section V. Finally, section VI concludes our paper.

II. RELATED WORKS

A. Intrusion Detection

Intrusion Detection Expert System was first proposed by Dorothy E. Denning [1]. It had a rule-based expert system to detect known types of intrusions with a statistical anomaly detection component based on profiles of users, host systems and the target systems. Later, a new version called Next-Generation Intrusion Detection Expert System was developed [2].

The idea of using anomaly detection came into mainstream with DARPA Intrusion Detection Evaluation in information security released in 1998 and 1999 in conjunction with the MIT [3]. However, it was shown that the DARPA datasets are not appropriate to simulate real network systems [4] initiating the need for development of new datasets for developing IDS.

Manuscript received on Mar. 21, 2017. This work is sponsored by Basic Science Research Program through the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP), and a follow-up of the invited journal to the accepted & presented paper of the 19th International Conference on Advanced Communication Technology (ICACT2017), and Grant ID is 2016R1A2B4015899. Kijoon Chae is the corresponding author.

Chae. Kijoon. Author is with Ewha Womans University, Seoul, 120750 Korea (corresponding author to provide phone: +82-10-3726-6157; e-mail: kjchae@ ewha.ac.kr).

Mowla Nishat. Author, is with Ewha Womans University, Seoul, 120750 Korea. (e-mail: nishat.i.mowla@gmail.com).

Doh Inshil. Author is with Ewha Womans University, Seoul, 120750 Korea. (e-mail: isdoh1@ewha.ac.kr).

B. Machine Learning Techniques for IDS

Various forms of existing machine learning techniques are used for developing IDS. [5] and [6] discusses a survey of these techniques. Among them one of the most promising techniques called the neural network consists of a collection of actions to transform a set of inputs to a set of searched outputs through a set of simple processing units, or nodes and connections between them. There are schemes for both supervised and unsupervised learning techniques such as multi-layer perceptron [7] and self-organizing maps [8] respectively. Neural networks are ideal when we consider all the various forms of network attack traffic that we can experience based on the misuse detection model and the anomaly detection model [9]. Neural networks have also been ideally combined with clustering techniques to achieve promising performance [18]. Different existing dataset are used to evaluate the performance of IDS using neural networks in many research works [10].

C. Modern IDS

Modern IDS have difficulty in dealing with high speed network traffic while attackers can utilize that to hide their exploits by IDS overloading with irrelevant information while executing an attack [11]. A memory efficient multiple character-approaching architecture suited for ASIC implementations was proposed in [12]. The focus mainly went into memory management which could reduce the accuracy. Therefore, to manage higher traffic throughput and increasing link speed hardware accelerators were used to create various forms of NIDS. [13] depicts that while working with a huge number of data, a two-class problem is always more accurate than multi-class problem. In our approach, we try to combine lessons from all the related works and develop a more accurate mechanism that also considers space and time efficiency as will be discussed in the next section.

III. Proposed Mechanism

There is always a trade-off between time, space and accuracy while designing an efficient Intrusion Detection System. When we increase the number of classes to be distinguished, the accuracy of the machine learning model decreases while the IDS becomes slow. Again, when we decrease the number of attributes, the accuracy of the system goes down while the IDS work faster. Reducing the number of attributes is not always a good idea since various forms of attacks can only be classified when we have an abundant number of attributes to distinguish them from others. Considering all these aspects, we try to create a mechanism which is not only accurate but also considers space and time efficiency. Therefore, to overcome the various penalties to the techniques due to huge data handling, we utilize the benefit of a two-class problem since it is time efficient and enhances accuracy as we will also discuss later. As we go down the process, we also try to make it space efficient by a logical elimination process. This step allows us to make our system more space efficient. However, we also make sure all the classes come in the process of classification step by step while we try to maintain a 2-class problem. The approach is to start with the two basic classes namely normal and attack classes. Once the data are identified not to belong to the normal class we eliminate the normal class instances and re-construct a two-class problem from the later class by taking one class of attack as our prioritized attack and the other class as other attack type class. If the data are not identified to be our prioritized attack type, then the instances for this attack class are removed and a new two class problem is constructed by following the same prioritization procedure as shown in Fig. 1

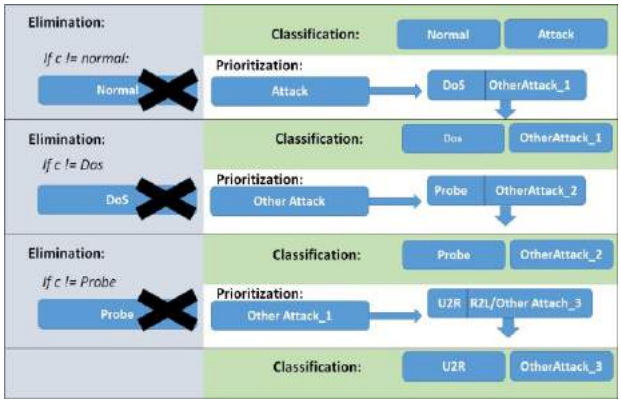


Fig. 1. Evolving ANN based 2-class IDS mechanism.

Fig. 1 shows our evolved ANN mechanism working with these the three major steps of classification, elimination and prioritization. We prioritize the attacks as DoS, Probe, U2R and R2L respectively. Our mechanism in the form of an algorithm for the four major types of network attack traffic in shown in Fig. 2.

Algorithm 1 Evolving ANN Intrusion Detection System

```
class =sum of 2 or more classes;
normal = 1;
attack =0;
run classification test;
i = 1;
if c == 0 && class ==TRUE then
    eliminate normal instances;
    dos = 1;
    other_attack_i = 0;
    if c == 0 && class ==TRUE then
        eliminate DoS instances;
        probe =1;
        other_attack_i+1 =0;
        if c == 0 && class==TRUE then
            eliminate probe instances;
            U2R =1;
            R2L/other_attack_i+2 =0;
            if c == 0 && class ==TRUE then
                traffic is R2L;
            else
                traffic is U2R;
        else
            traffic is probe;
    else
        traffic is DoS;
else
    traffic is normal;
```

Fig. 2. Evolving ANN Intrusion Detection System Algorithm for current four major network attack types.

Neural Networks classify with feature inputs by training a network formed with weights to derive higher level features that can be classified by a non-linear activation function. As shown in the Fig 3, x_i are the feature vectors input to the ANN system. In our case, we used 41 features provided by the KDD dataset [17]. u_j and u_k are the hidden layers which are also

called the intermediary output layers. u_l is the final output layer which helps us to identify the classes. w_{ij}, w_{jk} and w_{kl} are the weight from x_i to u_j, u_j to u_k and u_k to u_l respectively. Finally, a sigmoidal function is used at the outer layer to classify the input to an output class. Fig 3 shows the basic workflow of our ANN (Artificial Neural Network) based 2-class IDS mechanism in general.

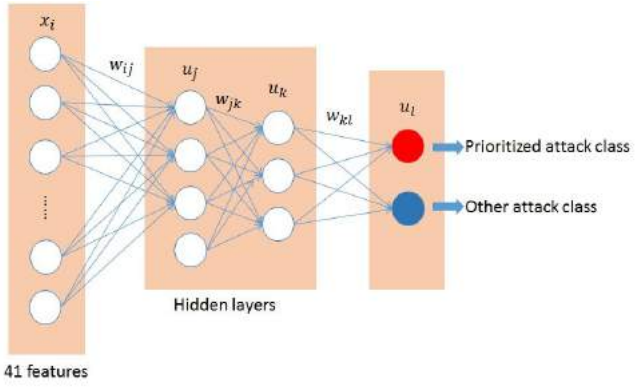


Fig. 3. ANN based 2-class IDS mechanism.

Currently there are 4 major types of network attack traffic namely DoS, Probe, R2L and U2R. Among them DoS refers to all the network traffic flooding attack types. Relevant features include source bytes, packet rates etc. Probe attacks are attacks conducted by sending meaningless packets in order to gain knowledge about the network. They are often detected by features such as duration of connection or source bytes. R2L refers to remote access attacks where the attacker tries to gain access to a remote system. Relevant features include duration of connections, service requested or failed log-in attempts. U2R is the type of attack in which the attacker tries to log-in to a normal account and then gain root administrator access. They are often identified by features such as number of files created or number of shell prompts invoked [16]. On these 4 classes of attack traffic and 1 class of normal traffic, we apply our evolving mechanism which, as discussed before, can be summarized to follow three major steps discussed below.

Classification

Our classification follows the ANN model of Multi-Layer Perceptron (MLP) working on a 2-class problem. Initially a normal class and an attack class are taken as the 2-classes.

Elimination

After a successful classification, the class with the lowest possibility is eliminated to effectively prune the analysed network traffic and a new 2-class problem is constructed from the later class.

Prioritization

In this step, a class with higher priority to be analysed is taken as the first form of class while making the other class as other attack class.

Our proposed mechanism is further optimized with hyper-parameter optimization with learning rate different datasets behave differently in different learning rates. Finally, we compare our performance gain with a highly-optimized Decision Tree algorithm.

IV. Performance Evaluation

We used all 41 features of KDD99 dataset [17] and evaluated the training time and detection accuracy for different attack types. We use a total of 1200 data instances from all the four different kinds of network attack traffic along with normal network traffic. We also took samples from all the subclasses of the 4 major types of network attack traffic. Table 1 shows all the sub-types of the 4-major network attack traffic that we used in our simulation [16]. In our first experiment, we show how the training time decreases as we decrease the number of classes from n to 2. Fig. 4 illustrates that as we decrease the number of classes, the number of training time also decreases significantly.

TABLE I
Network Attack Traffic

Attack Class	Attack Types
DoS	Back, Land, Neptune, Pod, Smurf, Teardrop
Probe	Satan, Ipsweep, Nmap, Portsweep
R2L	Guess_Password, Ftp_write, Imap, Phf, Warezmaster
U2R	Loadmodule

The four network intrusion traffic classes which are further sub-classified by KDD99 to accumulate samples from all the existing attack categories belonging to these four main classes of network attacks.

Fig. 4. Change in training time as the number of classes are decreased.

The above figure also depicts that if we have a huge number of classes reducing the sub-types to higher level types can effectively make the system faster. Next, we show how the time reduces as we use an evolved neural network.

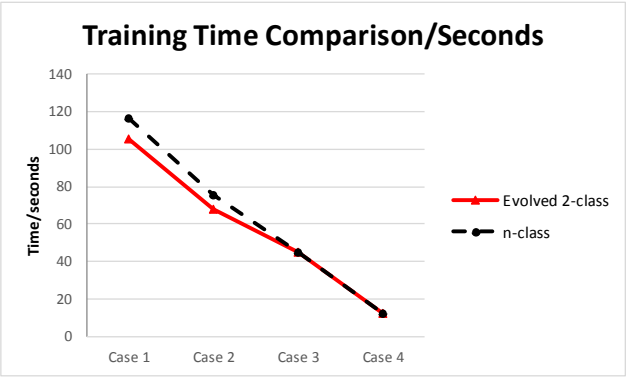


Fig. 5. Training time comparison between n-class and evolved 2-class.

Fig. 5 shows the comparison between the linear class reduction in neural network and an evolved neural network

class reduction performance. Here case 1 is all the 5 classes included (Normal, DoS, Probe, R2L, U2R), case 2 is all the 4 classes included (DoS, Probe, R2L, U2R), case 3 is all the 3 classes included (Probe, R2L, U2R) and class 4 is the 2 classes included (R2L and U2R). In case of our evolved 2-class mechanism case 1 means Normal and Attack class, case 2 means DoS and other attack class, case 3 means Probe and other attack class, case 4 means R2L and U2R class. As can be seen from the following figure our evolved 2-class mechanism is more time efficient than the normal n-class mechanisms and the difference of time efficiency tends to be higher as we increase the initial total number of classes in case 1.

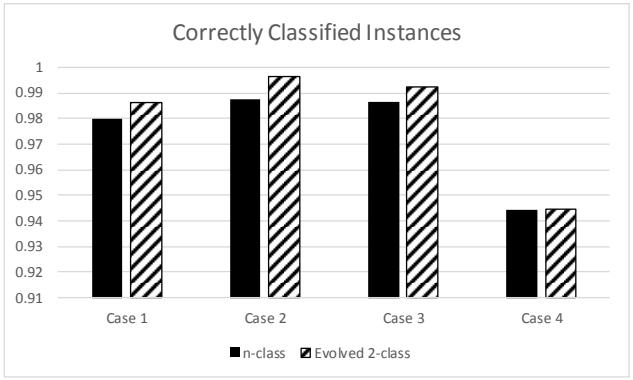


Fig. 6. Correctly classified instances.

Fig. 6 shows the correct classification results of our evolved neural network 2-class model versus the normal n-class neural network classification. As can be seen from the above figure our evolved 2-class model has higher correct classification in all the 4 cases of network attack traffic analysis. Since we have four major network attack traffic categories our model has only 4 cases. We believe our proposed mechanism can scale to other types of attacks with a higher number of classifications.

In the next experiment, we vary the learning rate from 0.1 to 0.0001 and observe the performance gain with a varied learning rate for our evolved neural network. Fig. 7, Fig 8, Fig 9 and Fig 10 shows the performance with varied learning rate for normal vs attack, DoS vs other attack, Probe vs other attack, and R2L vs U2R.

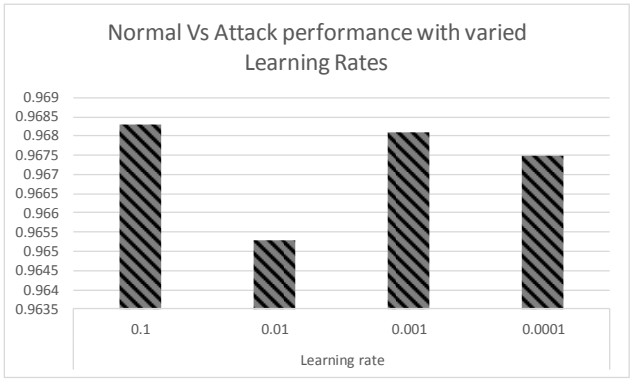


Fig. 7. Performance gain of normal vs attack with varied learning rates.

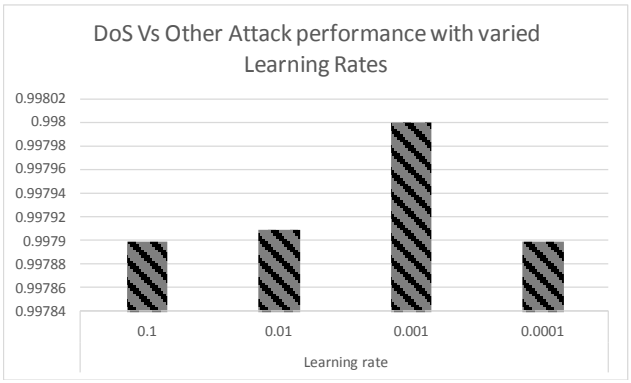


Fig. 8. Performance gain of DoS vs other attack with varied learning

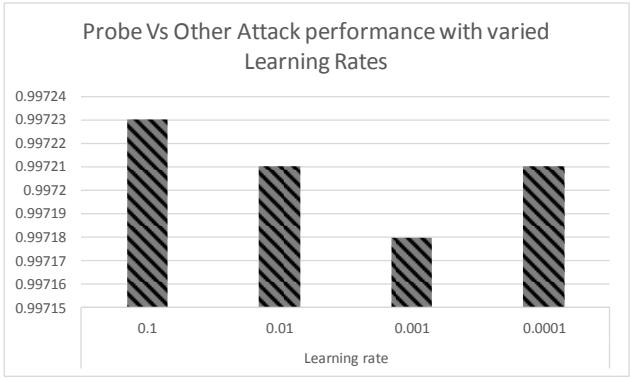


Fig. 9. Performance gain of Probe Vs other attack with varied learning rates.

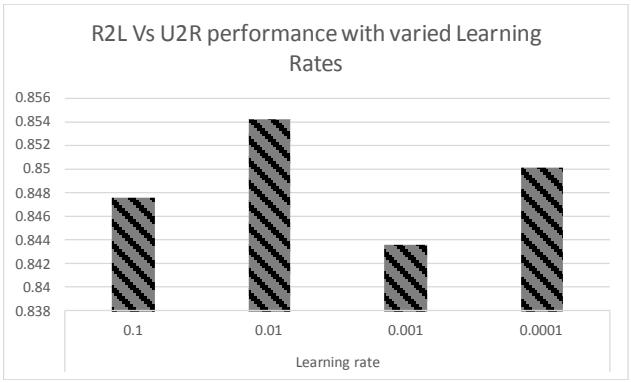


Fig. 10. Performance gain of R2L vs U2R with varied learning rates.

As can be seen from the above figure, the evolved two-class normal vs attack, DoS vs other attack, Probe vs other attack and R2L vs U2R achieves the highest performance with a learning rate of 0.1, 0.001, 0.1, and 0.01 respectively. Therefore, we use these optimized hyper-parameter values and compare the optimized evolving neural networks with a boosted Decision Tree. Fig. 11 shows the performance of the above four cases named case 1, case 2, case 3, and case 4.

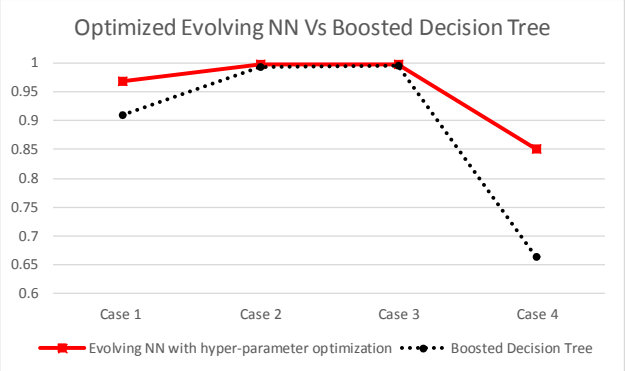


Fig. 11. Performance of hyper-parameter optimized Evolving Neural Network compared with boosted Decision Tree.

The above figure depicts that our evolved neural network with optimized hyper-parameters can outperform the state of the art Boosted Decision Tree. The performance is promising and it depicts that simpler neural networks can be optimized with techniques such as pairwise learning and hyper-parameter optimization to achieve similar and higher performance than more computationally intensive efficient machine learning algorithms.

V. Discussion

The performance gain of this paper is credited to the fact that as we decrease the number of classes in concern we make the classification borderline simpler. Thus, the classifier's complexity is reduced which can be evolved every time to create a two-class problem and solved pairwise to find the specific class in concern. The reduction in complexity is also contributing to the time efficiency of our mechanism. Besides the elimination process to create a new two-class problem allows us to make the problem space smaller and thus saving space.

Finally, the combination of evolved pairwise learning with hyperparameter optimization creates an ultimate leap of performance while reducing the complexity and making the problem space smaller but robust. The idea, thus, achieves a unique combination of high performance, speed with efficient space consumption.

VI. Conclusion

In this paper, we have proposed an Intrusion Detection System inspired by evolving neural network classification technique in order to detect the key 4 different types of attack traffic that can occur in a Medical Cyber Physical System network. We have shown that our proposed mechanism enhances the performance of the traditional supervised multilayer perceptron neural network. With certain hyper-parameter optimization, our mechanism can also achieve promising performance. With optimized hyper-parameters, our mechanism can outperform state-of-the-art algorithms such as boosted Decision Tree. Our mechanism, however, doesn't have a standardized mechanism for attack prioritization yet. Therefore, in future work we hope to identify and develop techniques to prioritize the attack traffic class based on attack prediction mechanisms. Also, we will consider other attack classes and evolving mechanism with clustered neural network.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2016R1A2B4015899). Kijoon Chae is the corresponding author.

References

[1] D. E. Denning, "An Intrusion-Detection Model," *IEEE Symposium on Security and Privacy*, 1986, pp. 118–131.
[2] D. Anderson, T. Frivold, and A. Valdes, "Next generation Intrusion Detection Expert System (NIDES): A summary," *SRI Int.*, no. May 1995, p. 47, 1995.
[3] M. Lincoln Laboratory, "DARPA Intrusion Detection Data Sets." [Online]. Available: https://www.ll.mit.edu/ideval/data/. [Accessed: 07- Apr-2016].
[4] J. McHugh, "Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," *ACM Trans. Inf. Syst. Security*, vol. 3, no. 4, pp. 262–294, 2000.
[5] J. Singh and M. J. Nene, "A Survey on Machine Learning Techniques for Intrusion Detection Systems," *Int. J. Adv. Res. Computer Communication Eng.*, vol. 2, no. 11, pp. 4349–4355, 2013.
[6] S. K. Wagh, "Survey on Intrusion Detection System using Machine Learning Techniques," *Int. J. Computer Appl.*, vol. 78, no. 16, pp. 30–37, 2013.
[7] C. Qiu, J. Shan, B. Polytechnic, and B. Shandong, "Research on Intrusion Detection Algorithm Based on BP Neural Network," *Int. J. Security and its Applications*, vol. 9, no. 4, pp. 247–258, 2015.
[8] L. Vokorokos, A. Baláž, and M. Chovanec, "Intrusion detection system using self-organizing map," *Informatica*, vol. 6, no. 1, pp. 1–6, 2006.
[9] J.-P. Planquart, "Application of Neural Networks to Intrusion Detection," 2001.
[10] S. K. Sahu, S. Sarangi, and S. K. Jena, "A detail analysis on intrusion detection datasets," Souvenir 2014 IEEE Int. Adv. Computer Conf. IACC 2014, pp. 1348– 1353, 2014.
[11] V. Paxson, "Bro: a system for detecting network intruders in real-time," *Computer Networks*, vol. 31, no. 23–24, pp. 2435–2463, 1999.
[12] H. Lu, K. Zheng, B. Liu, X. Zhang, and Y. Liu, "A memory-efficient parallel string matching architecture for high-speed intrusion detection," *IEEE J. Sel. Areas Communication*, vol. 24, no. 10, pp. 1793–1803, 2006.
[13] L. Dhanabal and S. P. Shantharajah, "A Study on NSLKDD Dataset for Intrusion Detection System Based on Classification Algorithms," *Int. J. Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 446–452, 2015.
[14] M. Alam, S. Abedin, M. Ameen and C. Hong, "Web of Objects Based Ambient Assisted Living Framework for Emergency Psychiatric State Prediction," *Sensors*, Vol. 16, No. 9, September 2016.
[15] O. Kocabas, T. Soyata, and M. K. Aktas, "Emerging Security Mechanisms for Medical Cyber Physical Systems", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vo. 13, No. 3, June 2016.
[16] S. Potluri, C. Diedrich, "Accelerated deep neural networks for enhanced Intrusion Detection System", 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1-8, September 2016.
[17] KDD 99 dataset, http://tunedit.org/repo/KDD_Cup/KDDCup99.arff
[18] W. Gang, H. Jinxing, M. Jian, H. Lihua. "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering." *Expert Systems with Applications*, Vol. 37, No. 9, pp. 6225-6232, 2010.

Nishat Mowla received the B.S degree in Computer Science from Asian University for Women, Chittagong, Bangladesh in 2013, an M.S. degree in Computer Science and Engineering from Ewha Womans University, Seoul, Korea in 2016. She is currently a PhD student at Ewha Womans University, Seoul, Korea. Her research interests include next generation network security, IoT

network security and network traffic analysis.

Inshil Doh received the B.S. and M.S. degrees in Computer Science at Ewha Womans University, Korea, in 1993 and 1995, respectively, and received the Ph.D. degree in Computer Science and Engineering from Ewha Womans University in 2007. She is currently an assistant professor of Computer Science and Engineering at Ewha Womans University, Seoul, Korea. Her research interests include wireless network, sensor network security, and M2M network security.

Prof. Chae received the B.S. degree in mathematics from Yonsei University in 1982, an M.S. degree in computer science from Syracuse University in 1984, and a Ph.D. degree in electrical and computer engineering from North Carolina State University in 1990. He is currently a professor in Department of Computer Science and Engineering at Ewha Womans University, Seoul, Korea. His research interests include sensor network, smart grid, CDN, SDN and IoT, network protocol design and performance evaluation.

Volume 6 Issue 4, Jul. 2017, ISSN: 2288-0003

ICACT-TACT JOURNAL

Global IT Research Institute

1713 Obelisk, 216 Seohyunno, Bundang-gu, Sungnam Kyunggi-do, Republic of Korea 13591

Business Licence Number : 220-82-07506, Contact: secretariat@icact.org Tel: +82-70-4146-4991