Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 542314, 8 pages
http://dx.doi.org/10.1155/2013/542314

Research Article

Classification of Hospital Web Security Efficiency Using Data Envelopment Analysis and Support Vector Machine

Han-Ying Kao, Tao-Ku Chang, and Yi-Cheng Chang

Department of Computer Science and Information Engineering, National Dong Hwa University, 1, Sec. 2, Da Hsueh Rd., Shou-Feng, Hualien 97401, Taiwan

Correspondence should be addressed to Han-Ying Kao; teresak@mail.ndhu.edu.tw

Received 3 July 2013; Revised 4 September 2013; Accepted 10 September 2013

Academic Editor: Jung-Fa Tsai

This study proposes the hybrid data envelopment analysis (DEA) and support vector machine (SVM) approaches for efficiency estimation and classification in web security. In the proposed framework, the factors and efficiency scores from DEA models are integrated with SVM for learning patterns of web security performance and provide further decision support. The numerical case study of hospital web security efficiency is demonstrated to support the feasibility of this design.

1. Introduction

During the past decades, the Internet and World Wide Web (www) have been prevalent platforms for information sharing and transformation. Consequently, web security management becomes a major theme in profit as well as nonprofit organizations. Assurance of information systems security involves not only tangible costs but also intangible inputs, which makes it challenging to evaluate the performance of the investments. This section concisely introduces the basics of web security, data envelopment analysis (DEA), support vector machine (SVM), classification on web security, and the goals of this study.

1.1. Phishing and Web Security. Phishing [1–4] is a criminal activity employing both social engineering and technical subterfuge to acquire personal data such as usernames, passwords, and credit card numbers. Phishing has become a serious threat to information security and Internet privacy. An analysis of phishing attacks by the Financial Services Technology Consortium [3] produces a taxonomy consisting of six stages: planning, set up, attack, collection, fraud, and postattack.

The increasing popularity of web-based systems has resulted in phishing behaviors causing significant financial damage to both individuals and organizations. The statistics provided by the Anti-Phishing Working Group [4] show that the number of unique phishing websites detected during the fourth quarter of 2009 was 137,619 and that financial services and payment services were the most-targeted industry sectors. It is clear that financial gain is the main objective of phishing attacks. A survey by Gartner [5] reveals that between September 2007 and September 2008, more than 5 million people in the United States were affected by phishing attacks with the average loss of US$ 351 per incident, and the number of victims has increased by 39.8 percent.

Protection against attacks and unauthorized access to sensitive information is vital in the Internet. Several technical antiphishing solutions have been proposed on the server's side or client's side. Server-side defenses employ Secure Sockets Layer certificates, user-selected site images, and other security indicators to help users verify the legitimacy of web sites, while client-side defenses equip web browsers with automatic phishing-detection features or add-ons (e.g., SpoofGuard) to warn users against suspected phishing sites [6]. In addition to the technical solutions, training users on antiphishing techniques is a frequently recommended and widely used approach for countering phishing attacks. ISO and NIST security standards, which many companies are contractually obligated to follow, include security training as an important component of security compliance [7, 8]. These standards describe a three-level framework comprising

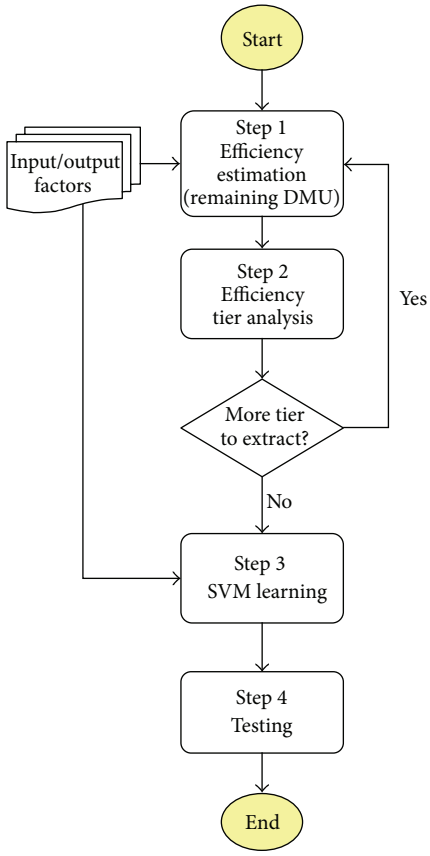

Figure 1: The research procedure.

awareness, training, and education. Users can be trained in several ways to understand how phishing attacks work [9–12].

1.2. Data Envelopment Analysis. Efficiency evaluation is a common issue in various domains and organizations, which is critical to investment analysis and resource allocation. Data envelopment analysis [13–15] is a celebrated efficiency evaluation technique and has been widely used in medical practices [16–19]. The DEA CCR ratio model developed by Charnes et al. [15] assesses the relative efficiency of decision-making units (DMUs) by maximizing the ratio of the weighted sum of outputs to that of inputs. Consider n DMUs ($j = 1, \ldots, n$) that require assessment. Each DMU consumes m inputs ($i = 1, \ldots, m$) and produces s outputs ($r = 1, \ldots, s$), denoted by $X_{1j}, X_{2j}, \ldots, X_{mj}$ and $Y_{1j}, Y_{2j}, \ldots, Y_{sj}$, respectively. The efficiency of DMU_k is computed as follows.

1.2.1. CCR Ratio Model. One has

$$\begin{aligned} \text{Max} \quad & E_k = \frac{\sum_{r=1}^{s} u_r Y_{rk}}{\sum_{i=1}^{m} v_i X_{ik}}, \\ \text{s.t.} \quad & \frac{\sum_{r=1}^{s} u_r Y_{rj}}{\sum_{i=1}^{m} v_i X_{ij}} \le 1, \quad j = 1, 2, \ldots, n, \\ & u_r, v_i \ge \varepsilon, \ r = 1, 2, \ldots, s; \ i = 1, 2, \ldots, m. \end{aligned} \tag{1}$$

Based on the CCR ratio model, the objective function E_k is maximized for every DMU_k individually. In the model, X_{ik} and Y_{rk} are the ith input and rth output of DMU_k; u_r, v_i are the weights of the outputs and inputs, respectively; ε is a small positive value which ensures that all weights are nonnegative. For computational convenience, frequently the CCR ratio model is transformed into a linear programming (LP) model by assuming [14] that

$$\sum_{i=1}^{m} v_i X_{ik} = 1. \tag{2}$$

Notably, the solution space of the CCR LP model is smaller than that of the CCR ratio model due to the constraint (2); thus, the CCR LP model finds the local optimum for the ratio model which comprises fractional terms [20].

1.3. Support Vector Machine. Support vector machine (SVM) is a popular classifier and pattern recognition method based on statistical learning [21–23]. Suppose that m training data $\{x_i, y_i\}$, $i = 1, 2, \ldots, l$, are given, where $x_i \in R^n$ are the input patterns and $y_i \in \{-1, +1\}$ are the related target values of two-class pattern classification case. Then the standard linear support vector machine is as follows:

$$\begin{aligned} & \min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i^2 \\ & \text{s.t.} \quad y_i \left[w^T x_i + b \right] \ge 1 - \xi_i, \quad \xi_i \ge 0, \ i = 1, \ldots, l, \\ & \text{that is, s.t.} \\ & w^T x_i + b \ge +1 - \xi_i, \quad \text{for } y_i = +1, \ \xi_i \ge 0, \\ & w^T x_i + b \le -1 + \xi_i, \quad \text{for } y_i = -1, \ \xi_i \ge 0, \end{aligned} \tag{3}$$

where b is the location of hyperplane relative to the origin. The regularization constant $C > 0$ is the penalty parameter of the error term $\sum_{i=1}^{l} \xi_i^2$ to determine the tradeoff between the flatness of linear functions ($w^T x_i + b$) and empirical error.

Hence, for such a generalized optimal separating hyperplane, the functional to be minimized comprises an extra term accounting the cost of overlapping errors. In fact the cost function (3) can be even more general as given below:

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i^k, \tag{4}$$

subject to the same constraints. This is a convex programming problem that is usually solved only for $k = 1$ or $k = 2$, and such soft margin SVMs are dubbed L_1 and L_2 SVMs, respectively [23].

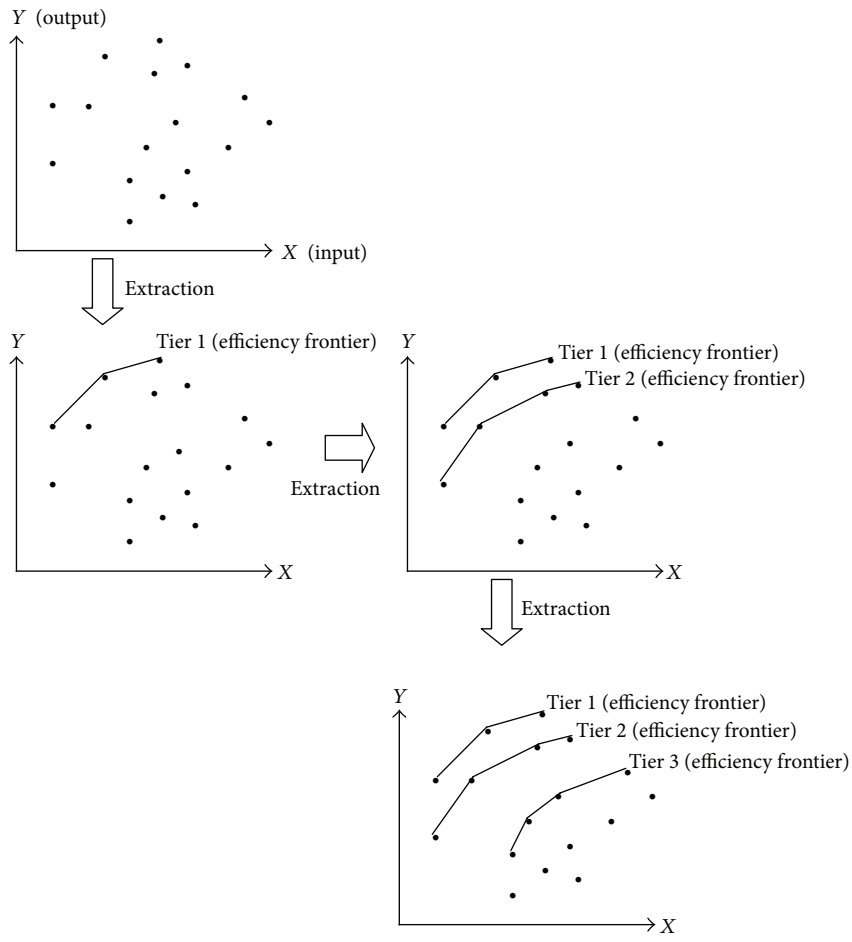

Figure 2: Tier extraction [27].

For L_1 SVMs ($k = 1$), the solution to a quadratic programming problem (3) is given by the saddle point of the primal Lagrangian shown below

$$\begin{aligned} L_p(w, b, \xi, \alpha, \beta) \quad &\text{(the primal Lagrangian)} \\ &= \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i^2 \\ &\quad - \sum_{i=1}^{l} \alpha_i \left\{ y_i \left[w^T x_i + b \right] - 1 + \xi_i \right\} - \sum_{i=1}^{l} \beta_i \xi_i, \end{aligned} \tag{5}$$

where α_i and β_i are the Lagrange multipliers.

Due to the KKT conditions, a dual Lagrangian function has to be maximized as follows:

$$\begin{aligned} L_d(\alpha) \quad &\text{(the dual Lagrangian)} \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j x_i^T x_j, \\ \sum_{i=1}^{l} \alpha_i y_i &= 0, \quad 0 \le \alpha_i \le C, \; i = 1, \ldots, l. \end{aligned} \tag{6}$$

In learning a nonlinear classifier, we can define a kernel and the dual Lagrangian to be maximized as follows:

$$\begin{aligned} L_p(\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right), \\ \sum_{i=1}^{l} \alpha_i y_i &= 0, \quad 0 \le \alpha_i \le C, \; i = 1, \ldots, l, \end{aligned} \tag{7}$$

where $K(x_i, x_j) = \phi^T(x_i)\phi(x_j)$ is the kernel function which maps the training vector x_i into a higher dimensional space. Popularly used kernel types include linear, polynomial, Gaussian, radial basis, and sigmoid [23].

1.4. Estimation and Classification on Web Security. Bose and Leung [2] investigate antiphishing preparedness of banks in Hong Kong by analyzing the websites of the registered Hong Kong banks. They compute the score for each bank by averaging the performance of the bank's website in three aspects, accessibility, usability, and information content. Later on Chen et al. [24] assess the severity of phishing attacks in terms of their risk levels and the potential loss in market value suffered by the targeted firms. They analyze 1030 phishing

Table 1: The factors for the DEA model.

Input factors
Clarity of purpose
X_1: security information is clearly structured on homepage;
X_2: adequate security measures have been adopted from official homepage or login page;
X_3: security policy documents are provided on homepage;
X_4: news or events on the Internet security are reported;
X_5: fault detection and responses adequately function.
Communication
X_6: content is frequently updated;
X_7: the transmission platform of interhospital medical records is built.
Security framework
X_8: availability of help desk services;
X_9: how sound are the security methods or protocols used;
X_{10}: first aid directions for computer security risks.
Output factors—the benefits of web security
Y_1: user satisfaction;
Y_2: progress of ISO 27001 accreditation.

alerts released on a public database and financial data related to the targeted firms using a hybrid method that predicts the severity of the attack. Nishanth et al. [25] employ a two-stage soft computing approach for data imputation to assess the severity of phishing attacks, which involves K-means algorithm and multilayer perception (MLP), probabilistic neural network (PNN), and decision trees (DT). Similar machine-learning techniques are employed by Lakshmi and Vijaya [26] for modelling the prediction task. The supervised learning algorithms, namely, multilayer perception, decision tree induction, and naïve Bayes classification, are used for exploring the results.

This study intends to integrate DEA and SVM for web efficiency estimation and classification for several key reasons. First, as the medical informatics and security gain growing attention, a practical evaluation scheme is needed. We develop the DEA to assess the relative efficiency of the hospitals as the pioneer study in the related field. Second, in addition to evaluating the current websites at one snapshot, some websites may be reviewed as potential data set. An efficient and reliable classifier is essential to discriminate future data. Among wide machine learning methods, SVM is relatively robust and convincing, so we integrate DEA and SVM to build the efficiency classification platform. Third, compared with related studies, this work emphasizes web security preparedness instead of potential web attacks detection. That is, we assess web security from a proactive view not limited to technical aspect. The rest of this paper is organized as follows. Section 2 addresses the problem and methods. Section 3 presents the numerical case study of web security analysis in medical institutions. Finally, the concluding remarks are given in Section 4.

2. The Method

This section develops the hybrid DEA and SVM approaches for efficiency classification. Consider n DMUs ($j = 1, \ldots, n$) that require assessment. Each DMU consumes m inputs ($i = 1, \ldots, m$) and produces s outputs ($r = 1, \ldots, s$), denoted by $X_{1j}, X_{2j}, \ldots, X_{mj}$ and $Y_{1j}, Y_{2j}, \ldots, Y_{sj}$, respectively. In the proposed framework, the factors and efficiency scores from DEA models are integrated with SVM in learning patterns of DMUs' performance and provide further decision support. The procedure of the hybrid methods is demonstrated in Figure 1.

Step 1 (efficiency evaluation). Based on (1) and (2), this study first evaluates the efficiency of the training data of the DMUs. The efficiency E_k of DMU_k is defined as in (1).

Step 2 (efficiency tier analysis). This step iteratively discriminates the fully productive group ($E_k = 1$) and the subproductive group ($E_k < 1$). Using tier analysis [27], the DMUs are divided according to their efficiency scores. Then, the fully productive group is moved to the current tier, while the remaining DMUs are kept for further tier extraction. The algorithm is described as follows.

S1: set $i = 1$. Set SC = {all DMUs}. Set t Max = 1.

S2: compute the efficiency scores of DMUs in SC.

S3: store the efficient DMUs with score 1 in Tier(i). Set SC = SC \ Tier(i).

S4: determine if the extraction process continues. If yes, set $i = i + 1$ and go to S2;

else set t Max = i and go to S5.

S5: for $i = 1$ to t Max

output DMUs in Tier(i).

end.

The procedure of this step is demonstrated in Figure 2. Each DMU will belong to one tier thereafter.

Step 3 (SVM learning). Here the classification schema is learned as

$$L_p(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} t_i t_j \alpha_i \alpha_j K\left(x_i \mid y_i, x_j \mid y_j\right), \quad \sum_{i=1}^{l} \alpha_i t_i = 0, \quad 0 \le \alpha_i \le C,\ i = 1, \ldots, l, \tag{8}$$

where t_i stands for the tier that DMU_i belongs to and $x_i \mid y_i$ is the vector combining the input and output factors of DMU_i. Since there can be more than two tiers, so this is a multiclass classification problem.

Step 4 (testing). The set of testing data will be used to validate the classification model.

In the next section, the case of hospital web security efficiency will be thoroughly studied for demonstrating the procedure developed above.

Table 2: The efficiencies from DEA.

Hospital ID	Efficiency*	Tier
1	1.000	1
2	0.778	2
3	0.903	2
4	0.667	3
5	0.889	2
6	0.920	2
7	1.000	1
8	0.808	2
9	1.000	1
10	0.122	4
11	0.970	2
12	0.669	3
13	0.264	4
14	0.332	4
15	1.000	1
16	0.303	4
17	1.000	1
18	0.678	3
19	1.000	1
20	0.295	4
21	1.000	1
22	0.155	4
23	0.750	2
24	1.000	1
25	1.000	1
26	0.295	4
27	1.000	1
28	0.264	3
29	1.000	1
30	0.295	4
31	0.176	4
32	1.000	1
33	0.335	4
34	1.000	1
35	1.000	1
36	0.264	3
37	0.122	4
38	0.848	2
39	0.848	2
40	0.264	4
41	1.000	1
42	0.145	4
43	0.122	4
44	0.284	3
45	0.801	3
46	0.264	3
47	0.264	4
48	0.801	3
49	1.000	1
50	0.125	4
51	0.961	2
52	1.000	1
53	0.145	4
54	1.000	1
55	1.000	1
56	1.000	1
57	0.176	4
58	0.122	4
59	0.388	3
60	0.388	3
61	0.388	3
62	0.388	3
63	0.388	3
64	0.388	3
65	0.388	3
66	0.388	3
67	1.000	1
68	1.000	1
69	0.800	2
70	0.388	3
71	0.778	2
72	0.388	3
73	0.332	4
74	0.388	3
75	0.388	4
76	0.388	3
77	1.000	1
78	0.388	3
79	0.295	4
80	0.332	3
81	0.388	3
82	0.388	3
83	0.332	3
84	0.388	3
85	0.332	3
86	0.388	3
87	0.332	4
88	0.388	3
89	1.000	1
90	0.388	3
91	1.000	1

*The efficiency from the first round evaluation without tier analysis.

3. Case Study

This study investigates 91 medical institutes in Taiwan, among which 8 (8.79%) are medical centers, 45 (49.45%) are metropolitan hospitals, and 38 (41.76%) are local community hospitals. To assess the hospitals' efficiencies in web security, 10 input factors in 3 categories (*clarity of purpose*, *communication*, and *security framework*) and 2 output factors (*user satisfaction* and *progress of ISO 27001 accreditation*) are defined in Table 1. Two professional users with web

Table 3: The distribution of efficiencies.

Efficiency	No. of DMU	Percentage (%)
1	25	27.47
0.900–0.999	4	4.40
0.800–0.899	7	7.69
0.700–0.799	3	3.30
0.600–0.699	3	3.30
0.300–0.399	20	21.98
0–0.2999	29	31.87

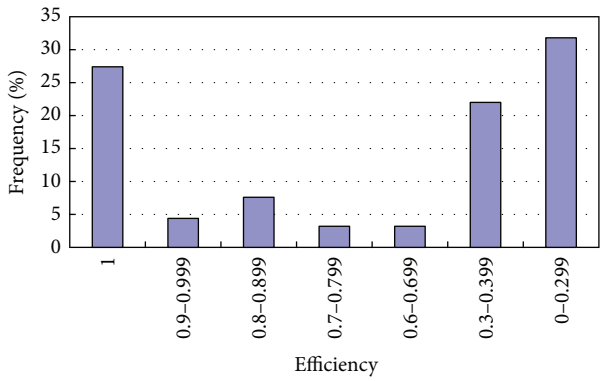

Figure 3: The distribution of efficiencies.

security expertise independently assess the web sites of these hospitals. All items are scored between 1 and 9 by the predefined measures, where 9 means total consistency between the statement and the practice while 1 stands for the opposite. After the reviewers observe the sample web sites, they give the scores according to the level of conformability to the factors in Table 1. The scores from each reviewer will be averaged as the input/output values for the DEA models. Most of the factors are nearly objective except *user satisfaction* (Y_1). Notably, the input variables are surrogate variables to construct the investments to the web security.

Step 1 (efficiency evaluation). By (1) and (2), the efficiencies of the information security investment in hospitals are computed. The distributions of the results are summarized in Tables 2 and 3 and Figure 3.

Step 2 (efficiency tier analysis). By each iteration of the tier analysis algorithm in Section 2, DEA determines one productive group of hospitals ($E_k = 1$) and the other group of subproductive hospitals ($E_k < 1$). Then the fully efficient group is extracted and the other proceeds to the next step. Each DMU will belong to an efficiency tier T thereafter. The members of the tiers are distributed as in Table 4 and Figure 4.

Steps 3 and 4 (*SVM learning and testing*). In this step, we use the utility LIBSVM [28] to build the SVM classification models. Four types of kernel functions are learned, including linear, polynomial, radial basis, and sigmoid. The accuracy in testing by different number of tiers and kernel function types is compared. The report is shown in Table 5 and Figure 5.

Table 4: The distribution of efficiency tiers.

Tier	No. of DMU	Percentage (%)
1	25	27.47
2	12	13.19
3	31	34.07
4	23	25.27

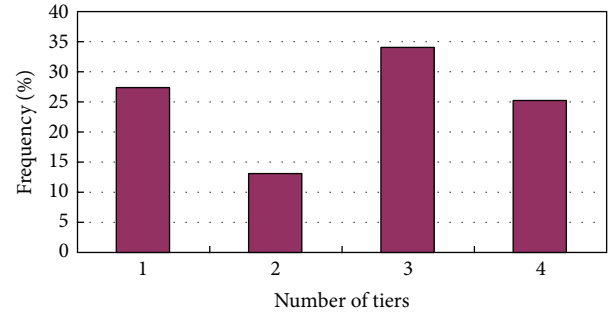

Figure 4: The distribution of tiers.

The distribution of efficiencies in Figure 3 manifests the unbalanced pattern where most hospitals lie in the two extremes of efficiency scales. However, by tier analysis, the distribution of tiers is nearly even except the second tier with the lowest number of hospitals.

From the results, the kernel functions with satisfactory prediction accuracy are linear (90.11% in average), radial basis (89.37% in average), and polynomial (87.18% in average), while the sigmoid function results in the lowest accuracy (average of 55.31%). Obviously, the linear, radial basis, and polynomial functions are more appropriate kernel types for web security efficiency classification in this case. The pattern of tier distribution is possibly the reason why those three kernel types outperform the sigmoid function in SVM classification.

From the perspective of data tier refinement, 2 tiers get the highest accuracy (average of 88.74%) while 4 tiers have the lowest rate (average of 72.53%). The results show that fewer data tiers obtain better accuracy in classification, which is consistent with the rule of thumb.

4. Conclusions

This study proposes the data envelopment analysis and support vector machine approaches for efficiency estimation and classification. For the feasibility of data collection, we use the surrogate variables to construct the tangible and intangible input factors. In defining the output factors, we define an objective variable of *ISO 27001 accreditation progress* and a subjective one of *user satisfaction*, which evaluate the efficiency from not only technical perspective but also users' perception. In the proposed framework, the factors and efficiency scores from DEA models are integrated with SVM for learning patterns of expected web security performance. From the case study, linear and radial basis kernel functions

Table 5: The accuracy from SVM (%).

Kernel type	2 tiers	3 tiers	4 tiers	Average
Linear	94.51	90.11	85.71	90.11
Polynomial	91.21	85.71	84.62	87.18
Radial basis	96.70	85.71	85.71	89.37
Sigmoid	72.53	59.34	34.07	55.31
Average	88.74	80.22	72.53	**80.49**

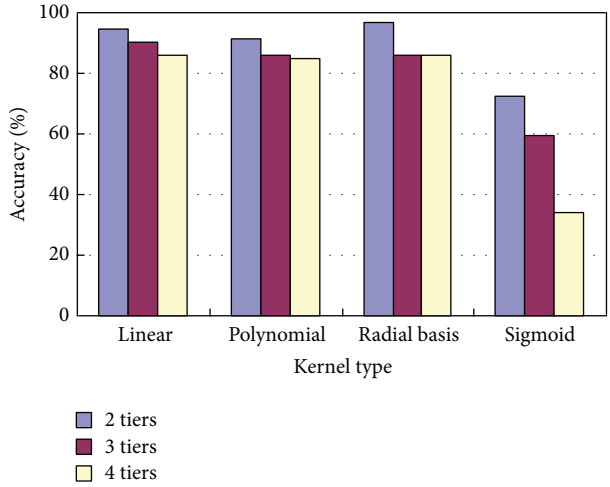

Figure 5: The results from SVM classifiers.

have superior performance in classification. Also classification with fewer data tiers obtains better accuracy in testing, which is consistent with the rule of thumb. This design integrates performance estimation and pattern learning to provide decision support in medical information security.

Acknowledgments

The authors are indebted to the anonymous reviewers for their careful reading and suggestions to enhance the quality of this paper. This work is supported by the National Science Council, Taiwan (Grant no. NSC 102-2410-H-259-039-, NSC 101-2221-E-259-030).

References

[1] M. Jakobsson and S. Myers, *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*, Wiley-Interscience, Hoboken, NJ, USA, 2007.

[2] I. Bose and A. C. M. Leung, "Assessing anti-phishing preparedness: a study of online banks in Hong Kong," *Decision Support Systems*, vol. 45, no. 4, pp. 897–912, 2008.

[3] R. Wetzel, "Tackling phishing," *Business Communications Review*, vol. 35, no. 2, p. 46, 2005.

[4] Anti-Phishing Working Group, Phishing Activity Trends Report, 2009, http://www.antiphishing.org/reports/apwg_report_Q4_2009.pdf.

[5] Gartner, "Gartner says number of phishing attacks on U.S. consumers increased 40 percent in 2008," 2009, http://www.gartner.com/it/page.jsp?id=936913.

[6] N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, "Client-side defense against web-based identity theft," in *Proceedings of the Annual Network and Distributed System Security Symposium (NDSS '04)*, 2004.

[7] ISO, "ISO/IEC 27001:2005—information technology—security techniques—information security management systems—requirements," Tech. Rep., International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), 2005.

[8] NIST, "NIST special publication 800-12: an introduction to computer security—the NIST handbook," Tech. Rep., National Institute of Standards and Technology, 2004.

[9] eBay, "Spoof email tutorial," 2010, http://pages.ebay.com/education/spooftutorial/.

[10] S. Srikwan and M. Jakobsson, "Using cartoons to 8 teach internet security," Tech. Rep., DIMACS, 2007.

[11] SonicWALL Phishing and Spam IQ Quiz, 2010, http://survey.mailfrontier.com/survey/quiztest.html.

[12] S. A. Robila and J. W. Ragucci, "Don't be a phish: steps in user education," in *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE '06)*, pp. 237–241, New York, NY, USA, June 2006.

[13] R. D. Banker, A. Charnes, and W. W. Cooper, "Some models for estimating technical and scale inefficiencies in data envelopment analysis," *Management Science*, vol. 30, no. 9, pp. 1078–1092, 1984.

[14] A. Charnes and W. W. Cooper, "Programming with linear fractional functionals," *Naval Research Logistics Quarterly*, vol. 9, pp. 181–186, 1962.

[15] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, 1978.

[16] J. Garcia-Lacalle and E. Martin, "Rural versus urban hospital performance in a "competitive" public health service," *Social Science and Medicine*, vol. 71, no. 6, pp. 1131–1140, 2010.

[17] S. J. Chang, H. C. Hsiao, L. H. Huang, and H. Chang, "Taiwan quality indicator project and hospital productivity growth," *Omega*, vol. 39, no. 1, pp. 14–22, 2011.

[18] M. Caballer-Tarazona, I. Moya-Clemente, D. Vivas-Consuelo, and I. Barrachina-Martínez, "A model to measure the efficiency of hospital performance," *Mathematical and Computer Modelling*, vol. 52, no. 7-8, pp. 1095–1102, 2010.

[19] Y. Chen, J. Du, H. D. Sherman, and J. Zhu, "DEA model with shared resources and efficiency decomposition," *European Journal of Operational Research*, vol. 207, no. 1, pp. 339–349, 2010.

[20] C. H. Huang and H. Y. Kao, "On solving the DEA CCR ratio model," *International Journal of Innovative Computing, Information and Control*, vol. 4, no. 11, pp. 2765–2773, 2008.

[21] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[22] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[23] L. Wang, Ed., *Support Vector Machines: Theory and Applications*, Springer, Berlin, Germany, 2005.

[24] X. Chen, I. Bose, A. C. M. Leung, and C. Guo, "Assessing the severity of phishing attacks: a hybrid data mining approach," *Decision Support Systems*, vol. 50, no. 4, pp. 662–672, 2011.

[25] K. J. Nishanth, V. Ravi, N. Ankaiah, and I. Bose, "Soft computing based imputation and hybrid data and text mining: the case of

predicting the severity of phishing alerts," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10583–10589, 2012.

[26] V. S. Lakshmi and M. S. Vijaya, "Efficient prediction of phishing websites using supervised learning algorithms," *Procedia Engineering*, vol. 30, pp. 798–805, 2012.

[27] H. K. Hong, S. H. Ha, C. K. Shin, S. C. Park, and S. H. Kim, "Evaluating the efficiency of system integration projects using data envelopment analysis (DEA) and machine learning," *Expert Systems with Applications*, vol. 16, no. 3, pp. 283–296, 1999.

[28] C. C. Chang and C. J. Lin, "LIBSVM—a library for support vector machines," 2003, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.